USING PROTEIN INTERACTOME NETWORKS TO UNDERSTAND HUMAN DISEASE AND EVOLUTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

By

Jishnu Das

August 2016

USING PROTEIN INTERACTOME NETWORKS TO UNDERSTAND HUMAN DISEASE AND EVOLUTION

Jishnu Das

Cornell University 2016

Proteins are macromolecules that perform many key biological functions in living organisms. However, despite their functional complexity, they do not act in isolation but are part of a large underlying cellular network – the protein interactome. In my thesis, I focus on ways in which the structure, topology and properties of this interactome network can be leveraged to better understand mechanisms of human disease and evolution at a molecular level. This work can be broken down into 3 broad categories – (1) in chapters 2-4, I focus on combining protein structure with interactome networks to better understand disease mechanisms via loss or gain of specific functions; (2) in chapters 5 and 6, I use gene expression data in conjunction with networks to study interaction dynamics and how this is useful in predicting cancer outcome; (3) in chapters 7 and 8, I examine global principles of protein network evolution from yeasts to human. At the beginning of each chapter, I have outlined both my role and the role of my colleagues in conducting the research described in the chapter.

BIOGRAPHICAL SKETCH

Jishnu Das was born and raised in Kolkata, India. From an early age, he become interested in analyzing complex problems, perhaps in part due to the Indian academic system favoring and selecting for the best problem solvers. He graduated from St. Xavier's Collegiate School Kolkata and was admitted into the Indian Institute of Technology, Kanpur after having cleared an entrance examination that only allows in the top 1% of applications based solely on their scores. However, despite having a penchant for solving hard problems, Jishnu always felt that the true challenge lay in translating these skills to real-life problems.

In college, Jishnu's chosen major was bioengineering. The reason behind choosing biology was that problems in the life sciences are most challenging due to their complexity, and being an engineer would let him deal with numbers. Numbers and computing have immensely fascinated him from his high-school days. Here, computational biology emerged as an extremely interesting career option as it had the complexity of biological problems and yet it had numbers. In 2010, he joined the field of Computational Biology at Cornell University with a broad interest in systems biology and functional genomics. He found an excellent fit in the lab of Dr. Haiyuan Yu, where he pursued graduate study in exploring mechanisms of human disease and evolution through biological networks.

Dedicated to Bapia, Mummum, Dadan and Ranju

ACKNOWLEDGEMENTS

I am who I am because of my family. My parents have gone through a tremendous amount of effort hardship just to make sure that I got the very best at every stage in life and enjoyed opportunities that they never had. My dad, a pediatrician, inspired me to be thoughtful, creative and inquisitive. But above all, it was he who introduced me to the scientific process; I would not have had the core skills to perform research without him. I gained a lot of my quantitative skills from my mother, an economics professor. She taught me the value of diligence and perseverance at a young age. She has made an incredible amount of personal sacrifice to hold our family together and without my mom, I would never have made it thus far. My paternal grandfather was a friend, philosopher and guide throughout my childhood. With him, I explored the literary universe, unhinged and unrestricted like Jonathan Livingston Seagull. My extended family – other grandparents, uncles, aunts, cousins and in-laws have also been very loving and extremely supportive of all my career choices.

In terms of my mentors at Cornell, I am very lucky to have found an inspirational and creative advisor like Dr. Haiyuan Yu. Being Haiyuan's first graduate student, I truly learnt the ropes from him and was incredibly lucky to have almost uninhibited access to him whenever necessary. Haiyuan has a penchant for bringing out the best in all his trainees. He taught me not just how to do science but also how to setup a research lab from scratch, an experience that will hopefully be very useful to me later in life. Dr. Andrew Clark, a member of my committee was instrumental in recruiting me to Cornell. Despite being very busy, Andy has been exceptionally supportive and accessible. Andy

has taught me several aspects of population genetics and molecular evolution. I have been especially impressed by how he can juggle so many things and yet give the sharpest and most pointed suggestions all the time. Dr Jim Booth, my third committee member, has been the incredibly helpful statistician to whom we turned to for advice whenever we were stuck with a complex statistical problem. Other professors at Cornell who have had a major influence in shaping a couple of my projects and the trajectory of my graduate career are Dr. Marcus Smolka and Dr Florentina Bunea.

I was also lucky to be part of a great lab environment involving people with a wide range of skills, both computational and experimental. On the computational side, when I started off, I learnt a lot from working with Xiujuan Wang, a motivated postdoc in the lab who achieved a lot in her relatively short stay. I also really enjoyed working with graduate students Yu Guo and Michael Meyer over the years, both talented graduate students. I have worked more with Michael and he has become both a great colleague on whom I can rely on for everything and a perfect sounding board for a lot of my ideas. On the experimental side, I have greatly enjoyed working with graduate students Tommy Vo and Robert Fragoza. I have been paired up with Tommy and Robert on different projects and have learnt a lot from them, especially their incredible work ethic and attention to detail. A lot of the work described in this thesis would not have been possible without them as most of our projects have both a computational and an experimental side. Two postdocs on the experimental side, Xiaomu Wei and Jin Liang have also been very helpful, often performing critical experiments that can make or break a paper.

Finally, it would have been impossible for me to go through graduate school without the support of my loving wife, Niranjana Natarajan. She has been a friend, a

partner, an intellectual equal and above all a soulmate with whom I could share everything. She too is wrapping up a PhD from Johns Hopkins. While this has meant traveling well over 100,000 miles during the course of the last 6 years just to be together on weekends, the bond that we share has made all geographical distances seem small.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH iii
ACKNOWLEDGEMENTS v
TABLE OF CONTENTS vii

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 HINT: HIGH-QUALITY PROTEIN INTERACTOMES AND THEIR APPLICATIONS IN UNDERSTANDING HUMAN DISEASE

2.1 ABSTRACT 5
2.2 INTRODUCTION 6
2.3 RESULTS 8
2.4 DISCUSSION 19
2.5 MATERIALS AND METHODS 19
2.6 FIGURE AND TABLE LEGENDS 22
2.7 REFERENCES 24

CHAPTER 3 ELUCIDATING COMMON STRUCTURAL FEATURES OF HUMAN PATHOGENIC VARIATIONS USING LARGE-SCALE ATOMIC-RESOLUTION PROTEIN NETWORKS

3.1 ABSTRACT 41
3.2 INTRODUCTION 42
3.3 RESULTS 44
3.4 DISCUSSION 52
3.5 MATERIALS AND METHODS 53
3.6 FIGURE AND TABLE LEGENDS 59
3.7 REFERENCES 61

CHAPTER 4 A MASSIVELY PARALLEL PIPELINE TO CLONE DNA VARIANTS AND EXAMINE MOLECULAR PHENOYTPES OF HUMAN DISEASE MUTATIONS

4.1 ABSTRACT 72
4.2 INTRODUCTION 73
4.3 RESULTS 75
4.4 DISCUSSION 86
4.5 MATERIALS AND METHODS 90
4.6 FIGURE AND TABLE LEGENDS 105

4.7 REFERENCES 108

CHAPTER 5 GENOME-SCALE ANALYSIS OF INTERACTION DYNAMICS REVEALS ORGANIZATION OF BIOLOGICAL NETWORKS

5.1 ABSTRACT 129
5.2 INTRODUCTION 130
5.3 RESULTS 131
5.4 DISCUSSION 136
5.5 MATERIALS AND METHODS 137
5.6 FIGURE AND TABLE LEGENDS 138
5.7 REFERENCES 139

CHAPTER 6 ENCAPP: ELASTIC-NET-BASED PROGNOSIS PREDICTION AND BIOMARKER DISCOVERY FOR HUMAN CANCERS

6.1 ABSTRACT 147
6.2 INTRODUCTION 148
6.3 RESULTS 150
6.4 DISCUSSION 161
6.5 MATERIALS AND METHODS 163
6.6 FIGURE AND TABLE LEGENDS 169
6.7 REFERENCES 173

CHAPTER 7 CROSS-SPECIES PROTEIN INTERACTOME MAPPING REVEALS SPECIES-SPECIFIC WIRING OF STRESS-RESPONSE PATHWAYS

7.1 ABSTRACT 189
7.2 INTRODUCTION 190
7.3 RESULTS 192
7.4 DISCUSSION 202
7.5 MATERIALS AND METHODS 204
7.6 FIGURE AND TABLE LEGENDS 221
7.7 REFERENCES 224

CHAPTER 8 A PROTEOME-WIDE FISSION YEAST INTERACTOME REVEALS NETWORK EVOLUTION PRINCIPLES FROM YEASTS TO HUMAN

8.1	ABSTRACT	248
8.2	INTRODUCTION	249
8.3	RESULTS	250
8.4	DISCUSSION	267
8.5	MATERIALS AND METHODS	269
8.6	FIGURE AND TABLE LEGENDS	284
8.7	REFERENCES	290

CHAPTER 1

A NETWORK PERSPECTIVE

John Donne wrote – “No man is an island”, golden words in today’s interconnected and interdependent world. But it is not just humans who are part of intrinsic networks; the molecules that make us human are also part of complex biological networks. In the following chapter, I outline the relevance of the network perspective in biology.

1.1 INTRODUCTION

The building blocks of life, nucleic acids and proteins, are complex bio-molecules that can carry out numerous biological functions. However, despite their functional complexity, the vast majority of these molecules do not act in isolation. Rather, they are part of well-coordinated biological networks that are robust, fault-tolerant and highly efficient. In my thesis, I focus on one such biological network – the protein-protein interaction network. In this network, the nodes are proteins and edges between these nodes represent biophysical interactions between proteins. My dissertation focuses on ways in which we can use the protein network to elucidate molecular mechanisms of human disease and evolution.

In the first part of my thesis – chapters 2, 3 and 4, I discuss how interaction networks can be combined with protein structures to generate three-dimensional structurally resolved networks. I then study disease mutations in the concept of these 3D networks to obtain novel insights into mechanisms of disease progression.

In the second part of my thesis – chapters 5 and 6, I focus on combining gene expression data with protein networks to understand the dynamics of interactions, as well as improve cancer outcome prediction.

In the third and final part of my thesis – chapters 7 and 8, I study the evolution of protein networks from yeasts to human. I develop statistical frameworks that can be used to

compare protein networks from different organisms and enunciate the molecular basis of network evolution.

The overall theme of this thesis is to show how network structure, organization and dynamics can be leveraged to gain key insights into mechanisms of complex biological processes. In future, similar methods may be applied to other biological networks such as transcriptional or metabolic networks, to uncover hitherto unknown mechanisms.

CHAPTER 2

HINT: High-quality protein interactomes and their applications in understanding human disease

In the following chapter, we describe the development of HINT, a high-quality repository of protein-protein interactions in different organisms. I am the sole first author of the paper resulting from this chapter (Das and Yu, BMC Systems Biology 2012). There is another paper on which I am a co-first author (Meyer*, Das* et al Bioinformatics 2013 *=Equal contribution) that builds on HINT and describes the generation of structurally-resolved interactomes. I do not devote a separate chapter to it as a lot of the relevant conceptual details are covered in Chapter 3.

2.1 ABSTRACT

A global map of protein-protein interactions in cellular systems provides key insights into the workings of an organism. A repository of well-validated high-quality protein-protein interactions can be used in both large- and small-scale studies to generate and validate a wide range of functional hypotheses. We develop HINT (http://hint.yulab.org) - a database of high-quality protein-protein interactomes for human, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. These were collected from several databases and filtered both systematically and manually to remove low-quality/erroneous interactions. The resulting datasets are classified by type (binary physical interactions vs. co-complex associations) and data source (high-throughput systematic setups vs. literature-curated small-scale experiments). We find strong sociological sampling biases in literature-curated datasets of small-scale interactions. An interactome without such sampling biases was used to understand network properties of human disease-genes - hubs are unlikely to cause disease, but if they do, they usually cause multiple disorders. HINT is of significant interest to researchers in all fields of biology as it addresses the ubiquitous need of having a repository of high-quality protein-protein interactions. These datasets can be utilized to generate specific hypotheses about specific proteins and/or pathways, as well as analyzing global properties of cellular networks. HINT will be regularly updated and all versions will be tracked.

2.2 INTRODUCTION

Numerous recent efforts in systems biology have tried to characterize the set of all possible pairwise physical interactions or the binary protein "interactome" of an organism (Bader et al., 2008; Cusick et al., 2005; Vidal, 2005). Most proteins perform their functions through interactions (Pawson and Nash, 2000). Thus, these large-scale maps are critical in elucidating the biological roles of functional products of genes that are identified by large-scale genome and cDNA sequencing projects. Because most of these efforts are discovery-oriented and try to explore previously unknown functionalities, it is of utmost importance to ensure that the resultant maps are of high quality. Erroneous results at this stage could propagate into both ill-conceived hypotheses and futile downstream experiments. Moreover, it has been shown that high-quality interaction networks can provide key insights into fundamental topological and biological properties of cellular systems (Batada et al., 2006, 2007; Bertin et al., 2007; Han et al., 2004). Although there are numerous databases (Hu et al., 2009; Kerrien et al., 2012; Keshava Prasad et al., 2009; Licata et al., 2012; Mewes et al., 2011; Salwinski et al., 2004; Stark et al., 2011; Turner et al., 2010) that try to systematically curate the entire repository of interactions for different organisms, there has been very little effort in filtering out unreliable ones. This has led to low overlaps between independent publications and resultant confusion as to which interactions are correct (Cusick et al., 2009; Venkatesan et al., 2009; Yu et al., 2008).

There are two major types of protein-protein interaction data – binary physical interactions and co-complex associations. While some databases distinguish between these two orthogonal datasets, others fail to do so. Binary interactions represent a direct biophysical interaction between two proteins. On the other hand, co-complex associations provide information about co-membership in a complex. A lot of these associations may actually represent indirect interactions

(Cusick et al., 2009; Yu et al., 2008). The biological information conveyed by these two kinds of interactions is different and for many applications it is necessary to have a clear distinction between these two.

There are two major methods to obtain a global map of binary interactions – literature-curation (LC) and high-throughput experiments (HT) (Cusick et al., 2009). LC refers to systematically collecting interaction data from thousands of small-scale studies directed at validating a single or a few specific hypotheses. On the other hand, HT experiments produce large-scale interaction maps. Because most LC data are generated by hypothesis-driven experiments, it is much easier to infer biological function from those studies as compared to HT experiments. On the other hand, although the search space of some HT experiments might be focused on certain functional groups, most HT experiments are not designed to detect the presence or absence of specific interactions. Any experiment can have two kinds of bias – "assay bias" and "sampling bias". The first arises because no assay is perfect and all experiments – HT or small-scale have their own characteristic biases (von Mering et al., 2002). However, small-scale studies also have a sampling bias, i.e., they are typically focused on one or a few proteins of interest and hence selectively sample interactions from only a part of the search space. HT experiments are free of this sampling bias, i.e., the search space is scanned without *a priori* expectations (Venkatesan et al., 2009; Yu et al., 2008). Thus, for many global topological analyses, it is often necessary to use only the HT datasets.

Here, we describe a publicly available protein-protein interaction database, HINT (*High-quality INT*eractomes) that directly addresses the above three issues and provides high-quality binary and co-complex interactions for human, *S. cerevisiae* and, *S. pombe.* The binary interactomes have also been divided into LC and HT subsets. Using these datasets, we show that there are

significant sociological sampling biases in LC datasets, i.e., well-studied proteins tend to have more interactions in LC datasets for both human and *S. cerevisiae*. Finally, using only the high-quality HT interactions for human, we find that disease genes (i.e., genes that have a causal connection with one or more diseases) with more interactions tend to cause more diseases. Even though this result is unexpected in light of previous findings that interaction hubs are less likely to cause disease (Feldman et al., 2008; Goh et al., 2007), it will help understand mechanisms of various disease processes and develop corresponding treatments.

2.3 RESULTS

Data source for protein-protein interactions

The set of all protein-protein interactions for the three organisms was downloaded from the public databases – BioGrid (Stark et al., 2011), DIP (Salwinski et al., 2004), HPRD (Keshava Prasad et al., 2009), IntAct (Kerrien et al., 2012), iRefWeb (Turner et al., 2010), MINT (Licata et al., 2012), MIPS (Mewes et al., 2011) and VisAnt (Hu et al., 2009). Not all three organisms were present in all the databases. Though some of the databases mentioned above store both genetic and physical interactions, only physical interactions were used in building the interactomes. Certain tools (Szklarczyk et al., 2011; Turner et al., 2010) also provide scoring schemes for protein-protein interactions. However, we do not include these for HINT as they integrate both computational predictions and experimentally determined interactions. Our goal is to provide a repository of only experimentally well-validated high-quality protein-protein interactions.

Building the database

Figure 2.1 summarizes how HINT was built. For each organism and each source database, a filter was applied to generate non-redundant lists of appropriate interactions for the two categories – binary and co-complex. The filter classifies interactions into the correct groups and removes ones that are inadequately supported by experimental evidence. The binary interactions were further classified as HT and LC based on the nature of the experiments that produced them. If the experiment in support of the interactions discovers greater than a cutoff number of interactions, it is classified as HT. To determine the cutoff, we calculated the distribution of number of interactions reported by each unique publication. The cutoff (>=100 interactions) corresponds to the top 0.5 percentile of studies, when all publications are ranked in decreasing order of interactions reported per study. For co-complex associations, there exists no such clear distinction between HT and LC because the average number of interactions detected in a single experiment is significantly higher.

The next step was to remove low-quality interactions. For ones supported by HT publications, a non-redundant list of papers was compiled and each publication was manually examined to verify that the actual experiments used by the authors agree with the evidence codes cited by the curators. All papers for which there was an error in this matching process were removed. Moreover, papers that do not validate the interactions obtained were also not included in HINT. Although some HT affinity purification followed by mass spectrometry (AP/MS) experiments producing co-complex associations report confidence scores, most binary HT experiments do not. For co-complex interactions, we require all interactions to be reported by two papers or more to ensure quality. For HT binary experiments, some report datasets of different levels of confidence – usually, core vs. non-core. We always include the highest-quality dataset (i.e., the core set only). Moreover, we ensure that every single interaction included is high-quality (please

see Quality control section). Within this high-quality dataset, the users of HINT are free to choose their own confidence cutoff based on any combination of the number of supporting publications and evidence code. For LC interactions, it is not possible to replicate this process, as the number of papers is too high. It has been shown that a large fraction of the LC interactions supported by a single publication cannot be verified (Cusick et al., 2009; Turinsky et al., 2010). Curation is an extremely painstaking process and we acknowledge that there may be some high-quality interactions supported by only one publication. However, it is impossible to distinguish them from the larger fraction that has been demonstrated to be of lower quality/erroneous (Venkatesan et al., 2009; Yu et al., 2008). Our goal here is to present to the community only a high-quality dataset that is free of potential biases due to differential curation of the same source publication. Only those LC interactions that are supported by two or more publications are preserved in our database. Table 1 provides a summary of the source databases used (version and download date). Table 2 reports the number of high-quality interactions in each of these databases in each category.

For the binary network, we generated two sub-interactomes – the high-quality LC (HQ-LC) and the high-quality HT (HQ-HT) sub-interactomes. Interactions that are supported by both forms of evidence belong to both.

Table 3 provides summary statistics for the different interactomes and their sub-classes. The numbers refer to unique entries and any interaction validated in multiple orientations (e.g., bait and prey in binary interaction detection experiments) or by different research groups is counted as a single entity. We find that the average degree for *S. pombe* is much lower than that of human or *S. cerevisiae* for both binary and co-complex data. This shows that the *S. pombe* interactome is still mostly unexplored. There is also a sharp increase in the average degree from binary to co-

complex for *S. cerevisiae*. This is expected given that models to generate topologies of co-complex networks tend to include several or all possible combinations (Bader and Hogue, 2002). However, the same does not hold true for human. This probably indicates that the human co-complex interactome is underexplored as compared to the *S. cerevisiae* one.

Figure 2.2 depicts the binary and co-complex interactomes for human and *S. cerevisiae*. The degree distribution of each of the networks is also illustrated and these plots correspond well with the theoretical expectation of the networks being scale-free (Barabasi and Bonabeau, 2003). It is not possible to produce these plots for *S. pombe* as the interactome for this organism is severely underexplored.

Quality control

There has been a great deal of effort in the literature at discovering new protein-protein interactions in different species to gain an understanding of the entire interactome of that organism. However, due to experimental errors and inaccurate curation, databases often contain interactions that are low quality/erroneous (Cusick et al., 2009). Since accuracy is of paramount importance in generating new hypotheses using these interaction data, it is essential to have an easily accessible repository of high-quality binary protein-protein interactions. HINT is a repository created by combining information from commonly used databases. To ensure quality control, we adopt the following protocol. Since the number of HT publications is relatively low as compared to the vast number of small-scale studies, we manually inspect each of the HT studies. We ensure that high-quality HT experiments included in HINT have been verified by orthogonal traditional assays (e.g., co-immunoprecipitation). Some experiments that do not perform any validation of their screen are considered low-quality and therefore removed. More

recently, we developed a statistical framework to comprehensively evaluate the quality of HT datasets verified by orthogonal assays in both human and *S. cerevisiae* (Venkatesan et al., 2009; Yu et al., 2008). Using this framework, we can quantitatively and experimentally measure the quality of individual interactions, as well as the whole dataset. The quality of interactions reported by a HT experiment can be measured by two independent statistical parameters – the number of interactions validated, i.e., the "validation rate" and the number of interactions that could be re-tested in the validation carried out, i.e., the "retest rate". The first parameter is a measure of the confidence associated with the validation carried out (i.e., more confidence can be associated with the results when a larger fraction of the reported interactions are validated), while the second one directly assays the reproducibility of the HT experiment. We carried out a comprehensive re-curation for all HT experiments included in HINT. Only those HT experiments that satisfy have a validation rate of >50% and a recuration rate of >75% are included in HINT. This ensures that only the highest quality HT experiments are included in HINT.

On the other hand, since it is impossible to manually check all small-scale studies, we require two independent publications to report the same interaction for it to be included in our dataset. This is because while some interactions from dedicated small-scale studies are high-quality and have been repeated multiple times in the literature, a significant fraction of interactions from small-scale experiments are not easily reproducible. In fact, many of the interactions that cannot be reproduced are supported by only one publication, were not produced by dedicated experiments and were often not even mentioned in the paper (Cusick et al., 2009). More importantly, it has been experimentally shown that such interactions are indeed of low quality (Venkatesan et al., 2009; Yu et al., 2008). Thus, our repository of high-quality interactions

contains only manually validated HT experiments and interactions from small-scale studies that have been reported at least twice in the literature.

To further validate the filtering approach used we adopted the following method. For each organism and interaction type, percentage overlaps between all pairs of databases that contain data relevant to that category were calculated before and after filtering. Since all these databases are curating the same information, we would expect the overlaps between any two of them to be high. However, that is not the case and we find low overlaps between pairs of databases. This supports our hypothesis that some of the information contained in these datasets is low-quality/incorrect. However, if our filtering scheme successfully removes these low-quality/incorrect interactions, the pairwise overlap between databases should increase considerably after filtering. We find that this is indeed the case. For each organism and interaction type, there is a significant enrichment in the average pairwise overlap between databases after filtering (Figure 2.3; *P*-values calculated using a cumulative binomial test). Specifically, let the maximum number of interactions for a certain organism and interaction type that can be common to a particular database pair before and after filtering be denoted by Mb_i and Ma_i respectively, where i is an index to denote the database pair. Let the percentage overlaps before and after filtering for that pair be denoted by Pb_i and Pa_i respectively. The average percentage overlap for that organism and interaction type before (*AvPB*) and after filtering (*AvPA*) are calculated as:

$$AvPB = \frac{\sum_i Mb_i \, x \, Pb_i}{\sum_i Mb_i}$$

$$AvPA = \frac{\sum_i Ma_i \, x \, Pa_i}{\sum_i Ma_i}$$

Querying the database

The database has two major parts – a query interface and a batch download for the entire interactomes of the organisms. The pooled interactions can be queried in the following manner.

The organism of interest is selected from a drop-down menu followed by entering the query proteins separated by semi-colons. Up to a maximum of 10 proteins can be entered per query. The database supports Entrez gene IDs (Maglott et al., 2007) and gene names for proteins in human and ORF names and gene names for proteins in *S. cerevisiae* (Cherry et al., 1997) and *S. pombe* (Matsuyama et al., 2006). The user also has the option of specifying the cutoff number of publications for each of the query proteins. One can also specify a particular evidence type for searching interactions. For each interacting protein, the gene name is listed in the first column followed by the list of Pubmed IDs of the papers supporting this interaction in column 2. The last column lists the PSI-MI evidence code (Hermjakob et al., 2004) that describes the kind of evidence supporting the interaction. The gene names are linked to the NCBI Entrez Gene database (Maglott et al., 2007) for human and *S. cerevisiae* and the GeneDB database (Hertz-Fowler et al., 2004) for *S. pombe*. The PubMed IDs link to the NCBI website for the relevant abstracts.

For batch download, separate links are provided for binary and co-complex interactomes for each organism. The binary interactome is also divided into the LC and HT networks. One notes here that the LC and HT networks are not completely mutually exclusive. There are certain protein-

protein interactions that have been discovered both by HT experiments and by LC. There are included in both interactomes.

Using HINT, it will now be possible to analyze, visualize, and generate reliable hypotheses about a part of or the complete interactome of the three different organisms – human, *S. cerevisiae* and *S. pombe.* Future efforts may be directed at similarly collecting and filtering data for other organisms and also updating the current dataset based on new findings**.**

Binary vs co-complex

HINT clearly distinguishes between binary and co-complex interactions. The binary network represents direct interactions between two proteins. On the other hand, the co-complex network merely indicates membership of a group and does not necessarily imply pairwise interactions. In most cases, the exactly topology of the complex is unknown. Two primary methods – the spoke model and the matrix model are used to represent these complexes. However, both models are approximations and merely suggest possible topologies (Bader and Hogue, 2002). Since different reports base their choice of model on study-specific conditions, all co-complex associations were included as curated in the source databases. No re-curation was performed. Moreover, compared to co-complex interactome models, binary maps have a greater fraction of transient signaling connections and inter-complex connections (Das et al., 2012; Yu et al., 2008). Since these two datasets represent fundamentally different biological entities, their overlap is low and it is important to differentiate between them in certain studies. For example, recent studies have examined how mutations may either lead to complete loss of gene products or edge-specific changes in the interactome (Dreze et al., 2009; Zhong et al., 2009). We show in a recent study that the pathogenesis of human disease can be better understood by looking at the position of

mutations on interaction interfaces (Wang et al., 2012). These approaches are applicable to direct binary interactions, as it is more difficult to infer interface pairs from co-complex associations. The latter can be resolved using information on three-dimensional structures of protein complexes if these are available. Thus, based on the context, it may be more appropriate to use one interactome over the other. Moreover, there are significant differences in the topological properties of these two networks. We calculated the clustering coefficient (Watts and Strogatz, 1998) and the edge betweenness (Girvan and Newman, 2002) for the different interaction networks in HINT. Clustering coefficient measures the density of clustering in an interaction network (Watts and Strogatz, 1998). We find that co-complex networks have a significantly higher clustering coefficient ($P < 10^{-8}$ in both cases as calculated by a two-sample Kolmogorov-Smirnov test) than binary networks. This shows that co-complex associations tend to be much more dense in terms of topological structure. Edge betweenness is used to detect community structure in networks. A higher betweenness value for an edge indicates that it connects different modules and disrupting this edge will fragment the network into disjoint components (Girvan and Newman, 2002). We find that binary networks for both human and *S. cerevisiae* have a significantly higher betweenness ($P < 10^{-8}$ in both cases as calculated by a two-sample Kolmogorov-Smirnov test) than co-complex networks for the two organisms. This suggests that co-complex associations form tightly regulated modules and binary interactions are often used to form links between these modules. We did not use the *S. pombe* networks for our global topological calculations as these interactomes are highly underexplored at this stage and the small number of interactions available make the networks unsuitable for global analyses.

HT protein-protein interactions in understanding human disease

People have realized in the last decade that a human disease is rarely the consequence of an isolated abnormality in a particular gene but is generally the outcome of complex perturbations of the underlying cellular network (Barabasi et al., 2011). This has led to systematic studies of interactome networks and numerous insights have been obtained from such studies. The structure of these networks is governed by key biological principles and changes in their global properties may be linked to human disease (Vidal et al., 2011). Further advances in such studies are expected to uncover the biological significance of disease-associated mutations discovered by genome-wide association studies (Manolio, 2010) and help in identifying biomarkers and novel drug targets (Barabasi et al., 2011).

Previous studies have shown that protein hubs tend to be essential genes (Jeong et al., 2000; Yu et al., 2004). Therefore, one interesting question is whether a lot of the hubs are disease genes. Using the HT interactome, we examined the distribution of disease genes across number of protein-protein interactions. We found that disease-genes tend not to be hubs (Figure 2.4A; error bars correspond to standard error of the mean assuming a binomial distribution). This result is consistent with earlier studies that find that disease genes are usually non-essential and occupy peripheral positions in the human interactome (Feldman et al., 2008; Goh et al., 2007). The finding is logical in light of an evolutionary argument – for essential genes, mutations would be more likely to affect fitness to the extent of causing embryonic lethality (Feldman et al., 2008; Goh et al., 2007).

However, we were unable to reproduce the same results using the LC interactome (Figure 2.4A; error bars correspond to standard error of the mean assuming a binomial distribution). There is a significant increase ($P < 10^{-8}$ as calculated by a one-way ANOVA) of percentage of disease genes with degree for proteins that have at least one interaction. This led us to believe that the

difference could be due to study biases in the LC data. To systematically analyze if this is true, we plotted the average number of publications against the number of interactions of proteins separately for the HT and LC interactomes. Intuitively, there should be no strong correlation between these two entities as the number of publications associated with a protein should have no connection with its degree. The average number of publications does not vary significantly with degree for the HT dataset but increases dramatically for the LC interactome (see Figures 2.4B and 2.4C). This illustrates the strong study bias in the LC data – proteins with a greater number of interactions tend to be revisited more often by small-scale studies. Our results are consistent with earlier findings that the degree of proteins in the LC interactome is strongly correlated with the number of publications associated with them (Pfeiffer and Hoffmann, 2007; Yu et al., 2008). This makes the LC interactome unsuitable for global topological analyses. The low overlap between the HT and LC interactomes also confirms that these are in fact two separate networks that need to be appropriately used based on the context.

To further investigate whether protein interactomes can help us understand disease mechanisms and uncover previously unknown disease genes, we used the HT human interactome to analyze what fraction of disease genes are disease-hubs, i.e., genes causing multiple diseases. We examined the distribution of disease-hubs as a function of their degrees (Figure 2.4D; error bars correspond to standard error of the mean assuming a binomial distribution). We observed that proteins with a higher number of interactions are significantly more likely to be disease hubs ($P < 10^{-8}$ as calculated by a one-way ANOVA). Though this may seem contradictory to earlier findings in Figure 2.4A, these two are in fact independent results. It is true that if a disease gene has more interactions, there is a higher probability of its fitness being affected. However, in Figure 2.4D, we focused only on disease genes. By virtue of the fact that these are observed in

the population as disease genes, their mutations are less likely to cause embryonic lethality. Therefore the evolutionary constrains in Figure 2.4A do not apply here. It is logical to expect that a disease protein with multiple interactions will have a greater propensity for causing multiple diseases. This is because a protein with more interactions is involved in more biological functions (Yu et al., 2004). This result also means that protein-protein interactions are important in the pathogenesis of many human diseases. Further studies on alteration of interactions by disease mutations may reveal insights into molecular mechanisms of various diseases and provide information about potential drug targets.

2.4 DISCUSSION

HINT is a comprehensive repository of high-quality binary and co-complex physical interactions in human, *S. cerevisiae*, and *S. pombe.* It establishes and implements systematic techniques for separating interactions based on both type (i.e., binary and co-complex) and data-source (i.e., LC and HT). Making these distinctions is critical for many applications. Using only the HT dataset, we demonstrated that human disease genes with a greater number of interactions tend to cause more diseases. Future directions involve implementation of the same techniques for other organisms of biological interest.

2.5 MATERIALS AND METHODS

Evidence codes and ID-mapping

As one of the primary goals of the database is to clearly distinguish binary interactions from co-complex associations, two separate and mutually exclusive lists of evidence codes were created – one for each category. An evidence code is a unique number assigned by the PSI-MI initiative to a particular form of experimental information in support of an interaction (Hermjakob et al., 2004). The lists used for both categories can be found in Tables 2.1, 2.2, and 2.3. Using these lists, all the interactions were classified into binary and/or co-complex. Interactions supported by evidence codes that are in neither of the two lists are excluded. Different databases use different gene identifiers and as this may lead to error, all gene identifiers in each database were converted to Entrez gene IDs for human and ORF names for *S. cerevisiae* and *S. pombe*. For each of the organisms, gene names are also provided in the bulk download files. Mapping files we obtained from Uniprot (2007) and the NCBI gene databases.

As described earlier, for an interaction to qualify as high-quality, it has to have at least one manually verified HT evidence code or at least two LC evidence codes supporting it.

Protein-protein interactions and human disease genes

To look at the distribution of human disease genes across number of protein-protein interactions, the following protocol was used. The total number of human proteins is taken to be 20,000. For the LC and the HT interactomes, we separately calculated the number of proteins that take part in 1, 2, 3, and >=4 interactions respectively. The difference of 20,000 and the sum of proteins in all these categories represents the number of proteins that have no known interactions in that particular network. Thus we have the number of proteins with 0, 1, 2, 3, and >=4 interactions for both interactomes. The mapping between human genes and diseases is obtained from OMIM

(Amberger et al., 2009) and HGMD (Stenson et al., 2009).Then the following formula was used to calculate the percentage of disease genes in each category (PG_i):

$$PG_i = \frac{N_i \, x100}{T_i}$$

where N_i is the number of disease genes in bin i and T_i is the total number of genes in bin i.

Here each bin corresponds to the number of interactions – 0, 1, 2, 3, and >=4 respectively. These values have been shown in Figure 2.4A. The error bars represent standard error of the mean assuming a binomial distribution (each gene is either involved or not involved in disease).

To calculate the sub-percentage of disease hubs in each category (PH_j), the following formula was used:

$$PH_j = \frac{N_j \, x100}{T_j}$$

where N_j is the number of disease hubs in bin j and T_j is the total number of disease genes in bin j

Here each bin corresponds to the number of interactions – 0, 1, and >=2 respectively and a disease hub is any disease gene implicated in three or more diseases. These values have been shown in Figure 2.4D. The error bars represent standard error of the mean assuming a binomial distribution (each protein is either a disease hub or it is not).

2.6 FIGURE AND TABLE LEGENDS

Figure 2.1

Flow diagram depicting the series of steps used to build HINT.

Figure 2.2

Binary and co-complex interactomes and degree distribution plots for human and *S. cerevisiae*

Figure 2.3

Average overlap percentage between all pairs of databases for binary and co-complex interactions in human, *S. cerevisiae* (*S.c.*), and *S. pombe* (*S.p.*) before and after filtering.

Figure 2.4

A. Percentage of disease genes within proteins that have 0, 1, 2, 3, and >=4 interactions respectively.

B. Plot of average number of publications associated with a protein versus the cumulative degree of the protein in the HT and LC interaction networks in human.

C. Plot of average number of publications associated with a protein versus the cumulative degree of the protein in the HT and LC interaction networks in *S. cerevisiae*.

D. Percentage of disease hubs within disease genes that have 0,1, and >=2 interactions respectively.

Table 2.1

List of PSI-MI evidence codes used to classify binary interactions and co-complex associations.

Table 2.2

Mapping used to convert MIPS evidence codes to PSI-MI evidence codes.

Table 2.3

Mapping used to convert VisAnt evidence codes to PSI-MI evidence codes.

2.7 REFERENCES

(2007). The Universal Protein Resource (UniProt). Nucleic Acids Res *35*, D193-197.

Amberger, J., Bocchini, C.A., Scott, A.F., and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res *37*, D793-796.

Bader, G.D., and Hogue, C.W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol *20*, 991-997.

Bader, S., Kuhner, S., and Gavin, A.C. (2008). Interaction networks for systems biology. FEBS Lett *582*, 1220-1224.

Barabasi, A.L., and Bonabeau, E. (2003). Scale-free networks. Sci Am *288*, 60-69.

Barabasi, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat Rev Genet *12*, 56-68.

Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hurst, L.D., and Tyers, M. (2006). Stratus not altocumulus: a new view of the yeast protein interaction network. PLoS Biol *4*, e317.

Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hurst, L.D., and Tyers, M. (2007). Still stratus not altocumulus: further evidence against the date/party hub distinction. PLoS Biol *5*, e154.

Bertin, N., Simonis, N., Dupuy, D., Cusick, M.E., Han, J.D., Fraser, H.B., Roth, F.P., and Vidal, M. (2007). Confirmation of organized modularity in the yeast interactome. PLoS Biol *5*, e153.

Cherry, J.M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R.K.*, et al.* (1997). Genetic and physical maps of Saccharomyces cerevisiae. Nature *387*, 67-73.

Cusick, M.E., Klitgord, N., Vidal, M., and Hill, D.E. (2005). Interactome: gateway into systems biology. Hum Mol Genet *14 Spec No. 2*, R171-181.

Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.R., Simonis, N., Rual, J.F., Borick, H., Braun, P., Dreze, M.*, et al.* (2009). Literature-curated protein interaction datasets. Nat Methods *6*, 39-46.

Das, J., Mohammed, J., and Yu, H. (2012). Genome scale analysis of interaction dynamics reveals organization of biological networks. Bioinformatics.

Dreze, M., Charloteaux, B., Milstein, S., Vidalain, P.O., Yildirim, M.A., Zhong, Q., Svrzikapa, N., Romero, V., Laloux, G., Brasseur, R.*, et al.* (2009). 'Edgetic' perturbation of a C. elegans BCL2 ortholog. Nat Methods *6*, 843-849.

Feldman, I., Rzhetsky, A., and Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A *105*, 4323-4328.

Girvan, M., and Newman, M.E. (2002). Community structure in social and biological networks. Proc Natl Acad Sci U S A *99*, 7821-7826.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proc Natl Acad Sci U S A *104*, 8685-8690.

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

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C.*, et al.* (2004). The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol *22*, 177-183.

Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K.*, et al.* (2004). GeneDB: a resource for prokaryotic and eukaryotic organisms. Nucleic Acids Res *32*, D339-343.

Hu, Z., Hung, J.H., Wang, Y., Chang, Y.C., Huang, C.L., Huyck, M., and DeLisi, C. (2009). VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res *37*, W115-121.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L. (2000). The large-scale organization of metabolic networks. Nature *407*, 651-654.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U.*, et al.* (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res *40*, D841-846.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A.*, et al.* (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res *37*, D767-772.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E.*, et al.* (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res *40*, D857-861.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res *35*, D26-31.

Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. N Engl J Med *363*, 166-176.

Matsuyama, A., Arai, R., Yashiroda, Y., Shirai, A., Kamata, A., Sekido, S., Kobayashi, Y., Hashimoto, A., Hamamoto, M., Hiraoka, Y.*, et al.* (2006). ORFeome cloning and global analysis

of protein localization in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol *24*, 841-847.

Mewes, H.W., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K.F., Stumpflen, V.*, et al.* (2011). MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res *39*, D220-224.

Pawson, T., and Nash, P. (2000). Protein-protein interactions define specificity in signal transduction. Genes Dev *14*, 1027-1047.

Pfeiffer, T., and Hoffmann, R. (2007). Temporal patterns of genes in scientific publications. Proc Natl Acad Sci U S A *104*, 12052-12056.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res *32*, D449-451.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X.*, et al.* (2011). The BioGRID Interaction Database: 2011 update. Nucleic Acids Res *39*, D698-704.

Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S., and Cooper, D.N. (2009). The Human Gene Mutation Database: 2008 update. Genome Med *1*, 13.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P.*, et al.* (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res *39*, D561-568.

Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M., and Wodak, S.J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. Database (Oxford) *2010*, baq026.

Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M., and Wodak, S.J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford) *2010*, baq023.

Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Vidal, M. (2005). Interactome modeling. FEBS Lett *579*, 1834-1838.

Vidal, M., Cusick, M.E., and Barabasi, A.L. (2011). Interactome networks and human disease. Cell *144*, 986-998.

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

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotech *30*, 159-164.

Watts, D.J., and Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. Nature *393*, 440-442.

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

Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., and Gerstein, M. (2004). Genomic analysis of essentiality within protein networks. Trends Genet *20*, 227-231.

Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D.*, et al.* (2009). Edgetic perturbation models of human inherited disorders. Mol Syst Biol *5*, 321.

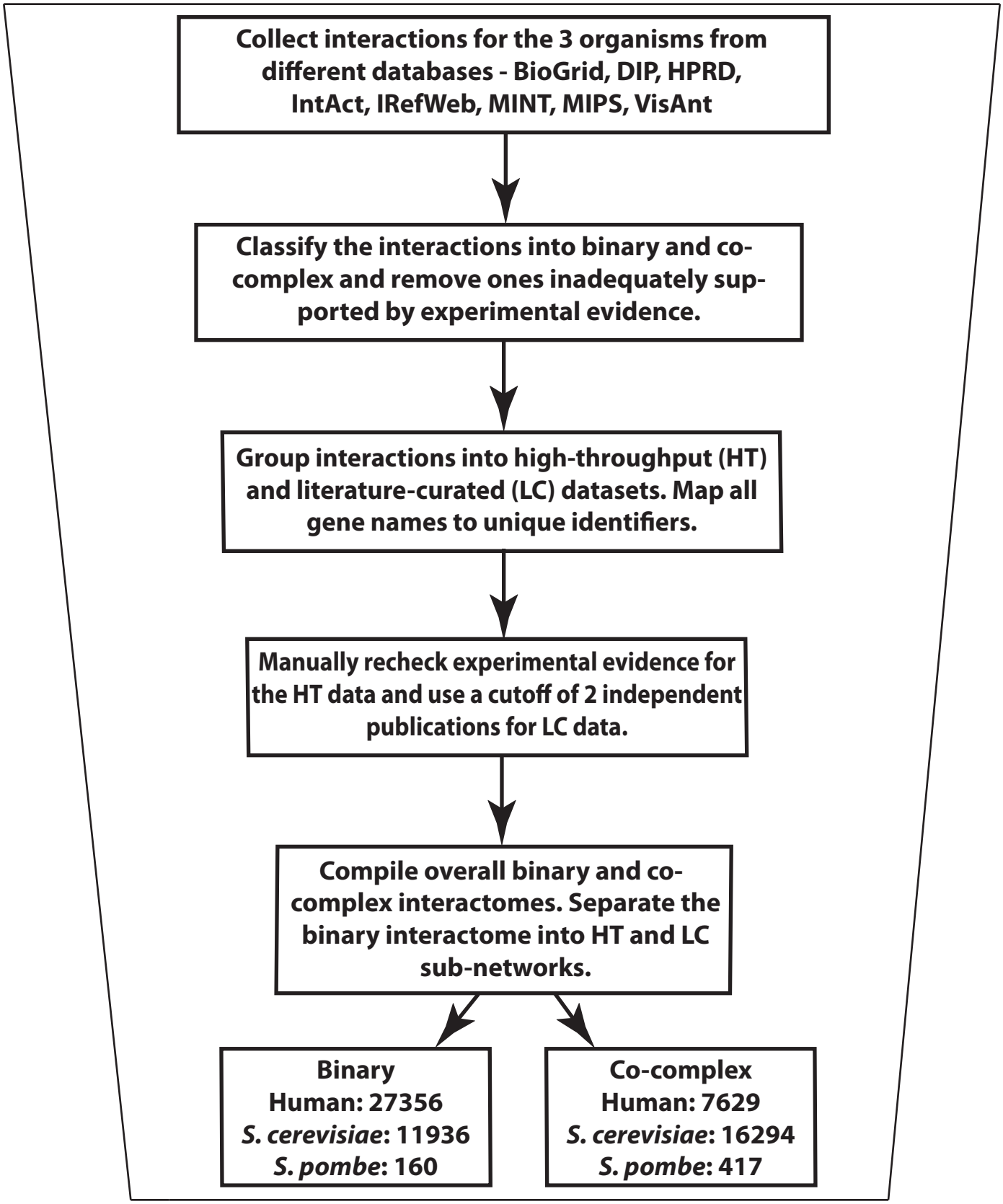
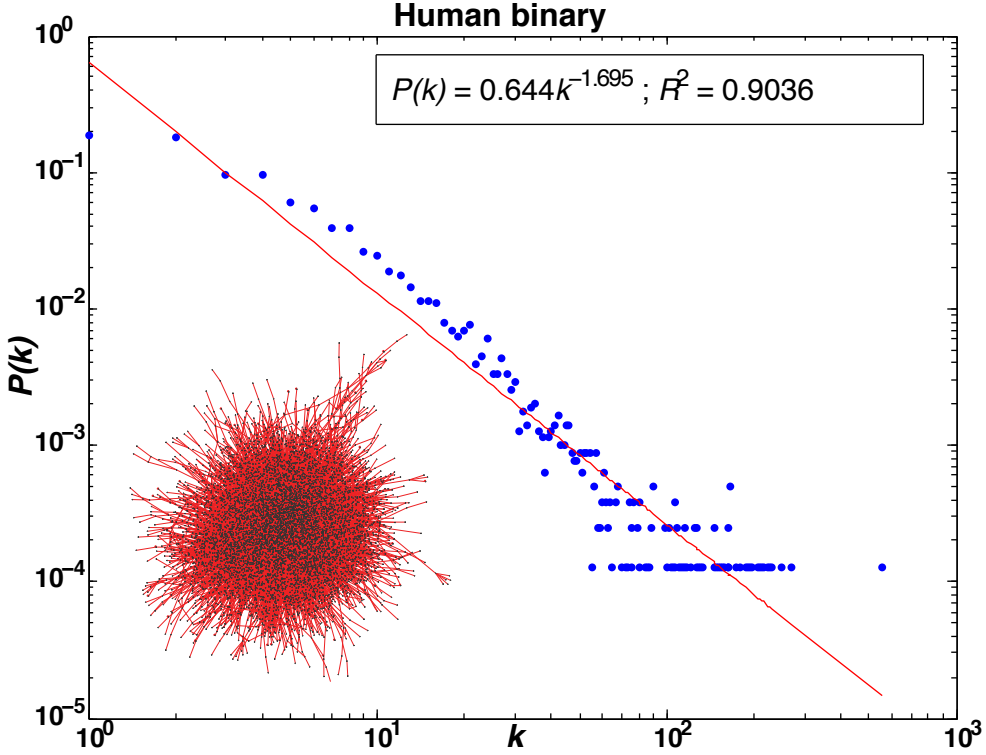
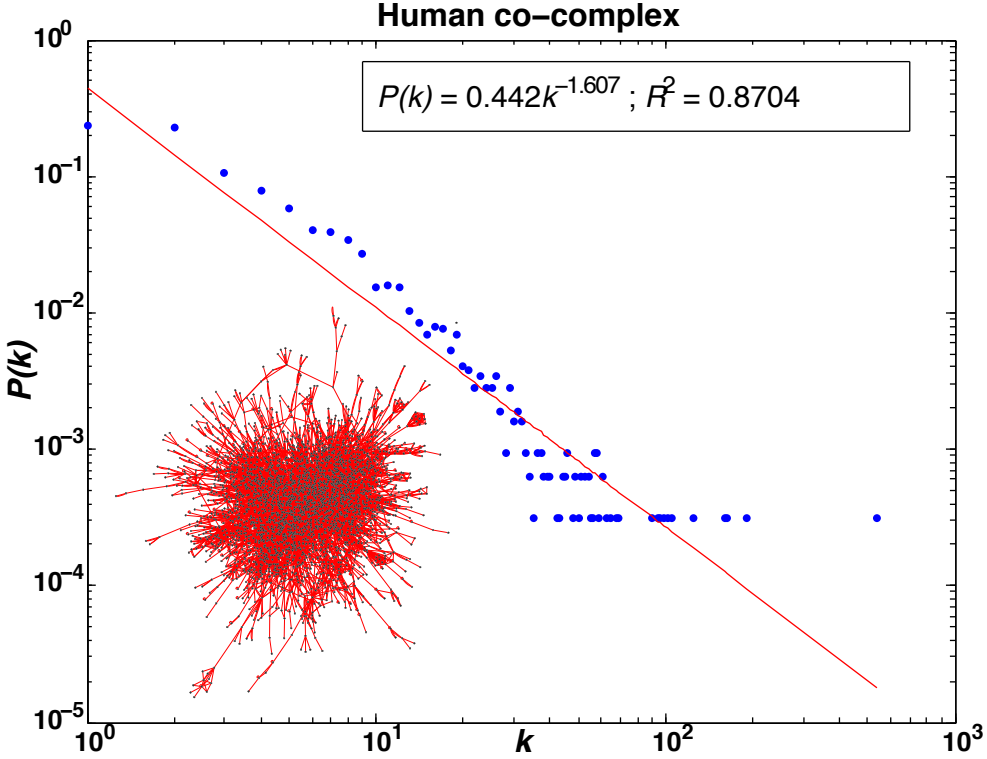
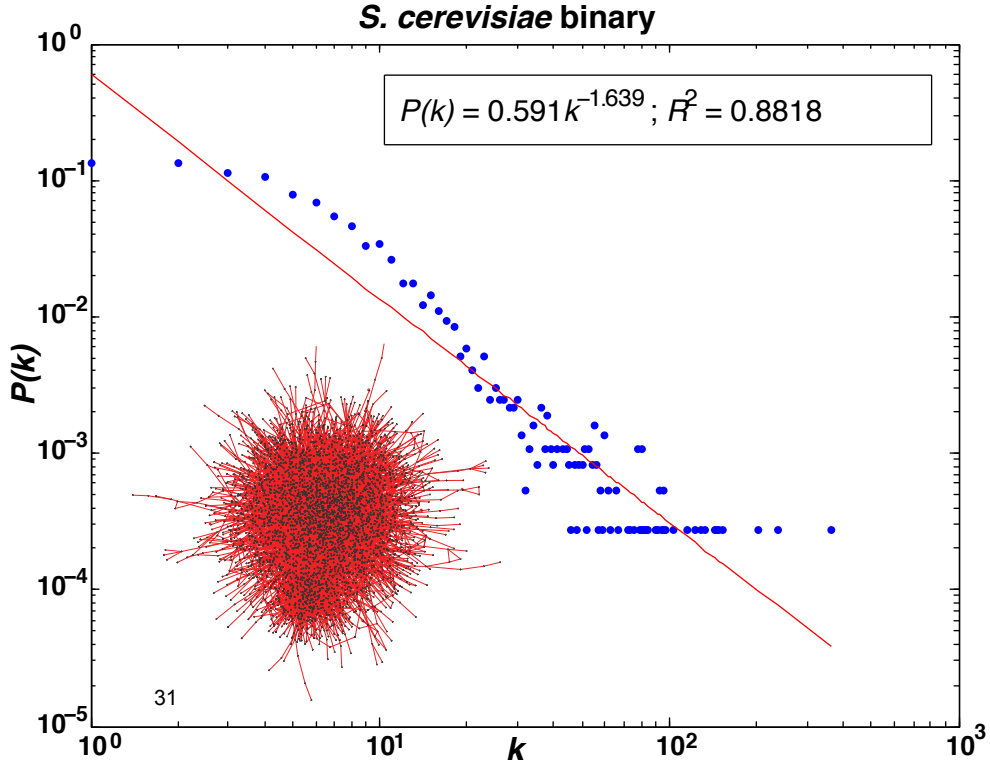

Collect interactions for the 3 organisms from different databases - BioGrid, DIP, HPRD, IntAct, IRefWeb, MINT, MIPS, VisAnt
Classify the interactions into binary and co-complex and remove ones inadequately supported by experimental evidence.
Group interactions into high-throughput (HT) and literature-curated (LC) datasets. Map all gene names to unique identifiers.
Manually recheck experimental evidence for the HT data and use a cutoff of 2 independent publications for LC data.
Compile overall binary and co-complex interactomes. Separate the binary interactome into HT and LC sub-networks.
Binary
Human: 27356
S. cerevisiae: 11936
S. pombe: 160
Co-complex
Human: 7629
S. cerevisiae: 16294
S. pombe: 417

Figure 2.1


Human binary
$P(k) = 0.644k^{-1.695}$; $R^2 = 0.9036$
P(k)
k


Human co–complex
$P(k) = 0.442k^{-1.607}$; $R^2 = 0.8704$
P(k)
k


S. cerevisiae binary
$P(k) = 0.591k^{-1.639}$; $R^2 = 0.8818$
P(k)
k
31

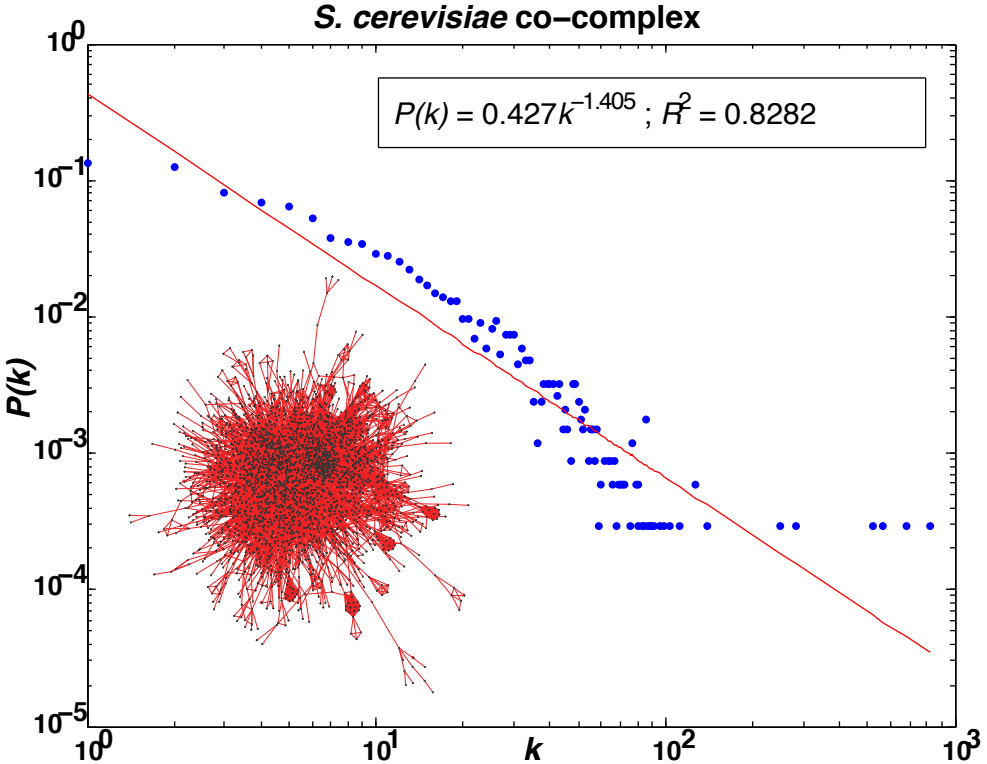

S. cerevisiae co–complex
$P(k) = 0.427k^{-1.405}$; $R^2 = 0.8282$
P(k)
k

Figure 2.2

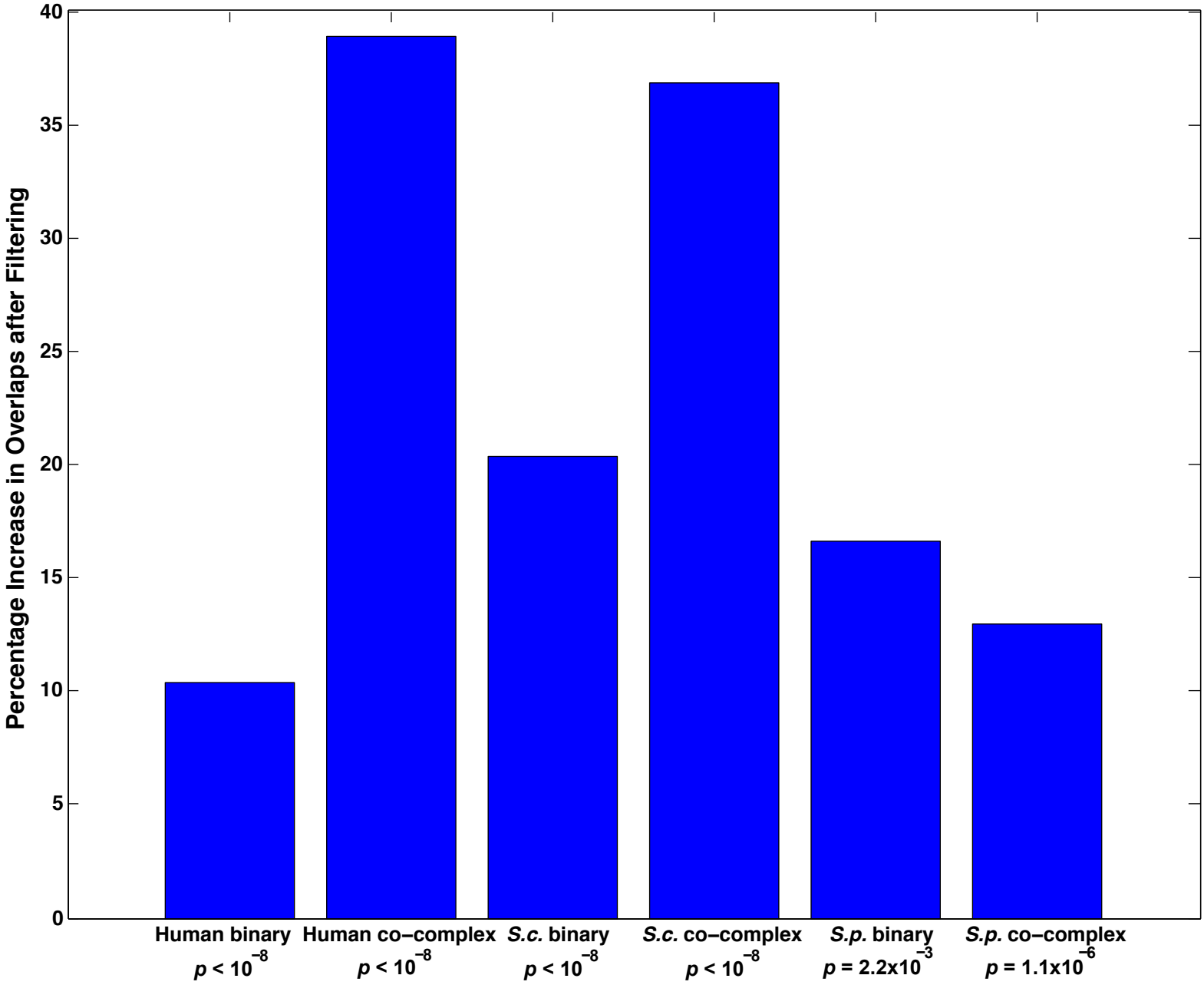
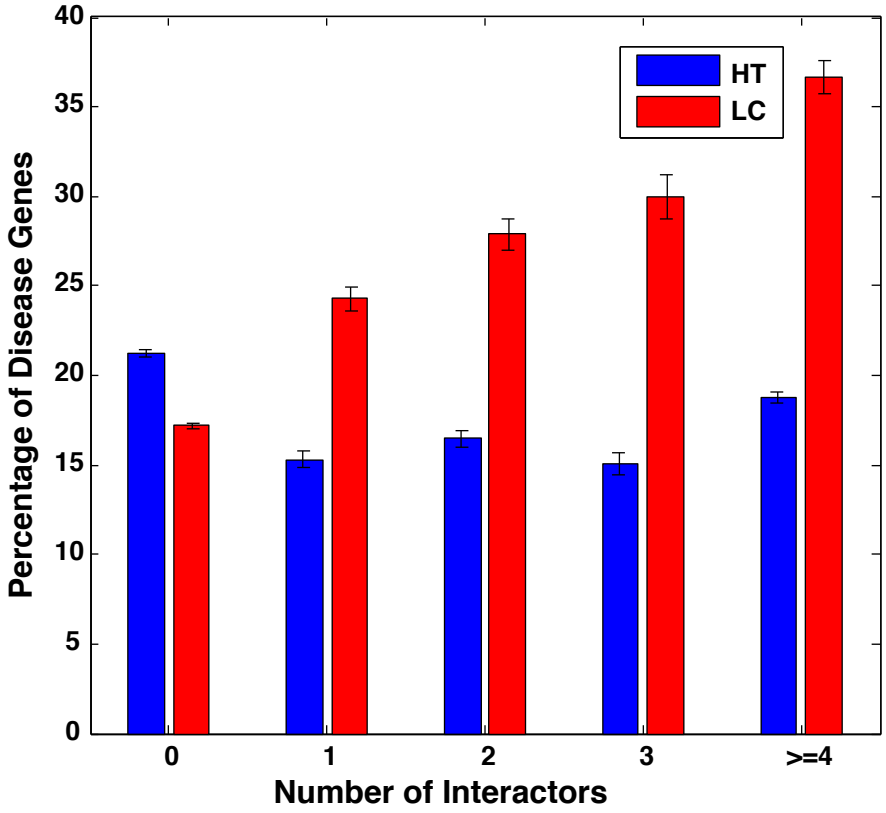
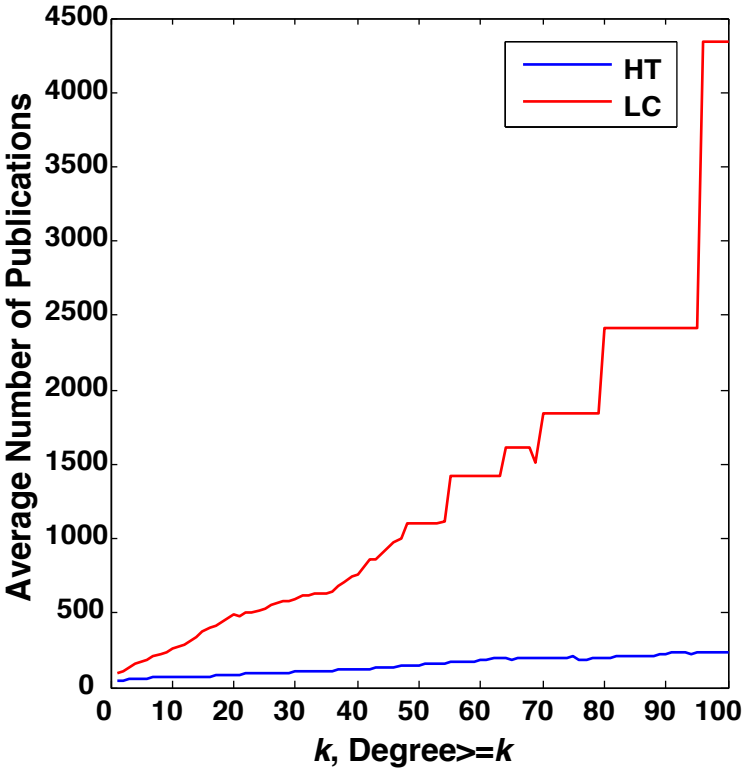
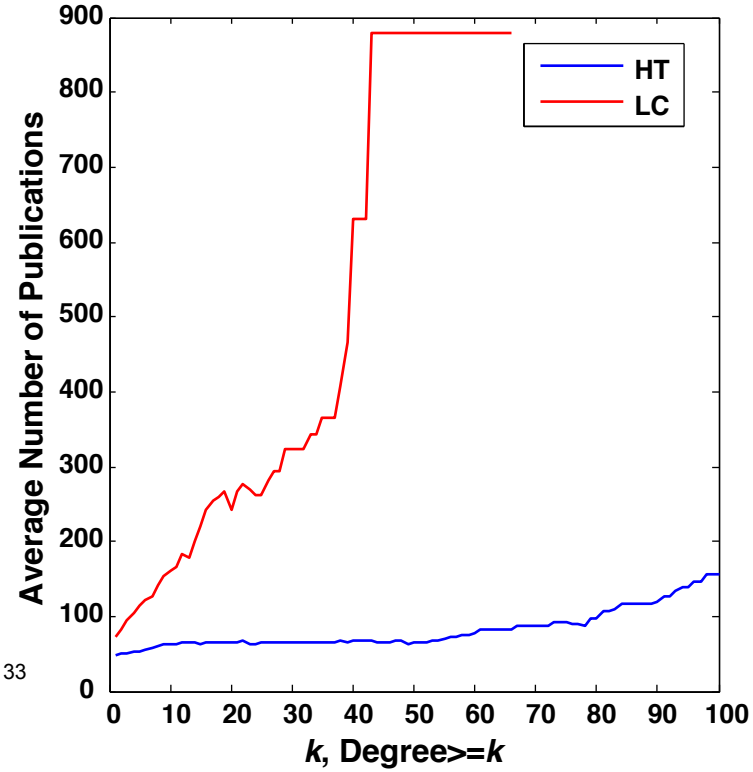

Percentage Increase in Overlaps after Filtering
40
35
30
25
20
15
10
5
0
Human binary
p < 10^−8
Human co−complex
p < 10^−8
S.c. binary
p < 10^−8
S.c. co−complex
p < 10^−8
S.p. binary
p = 2.2x10^−3
S.p. co−complex
p = 1.1x10^−6

Figure 2.3


HT
LC
Percentage of Disease Genes
Number of Interactors
0
1
2
3
>=4


HT
LC
Average Number of Publications
k, Degree>=k


HT
LC
Average Number of Publications
k, Degree>=k

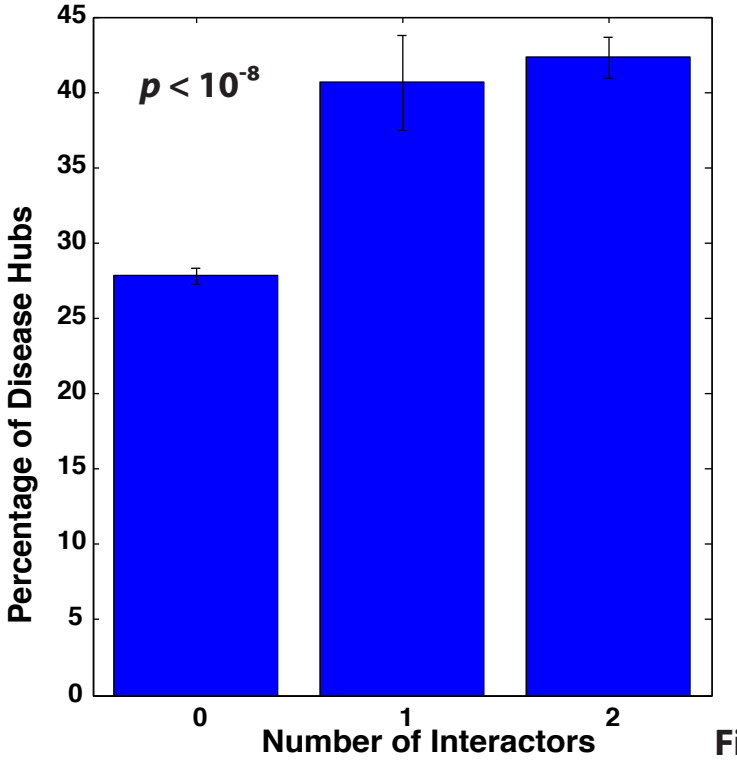

p < 10-8
Percentage of Disease Hubs
Number of Interactors

Figure 2.4

TABLE 2.1

Binary Interactions	
Experimental Evidence	**PSI-MI Evidence Code**
array technology	0008
beta galactosidase complementation	0010
beta lactamase complementation	0011
bioluminescence resonance energy transfer	0012
adenylate cyclase complementation	0014
circular dichroism	0016
classical fluorescence spectroscopy	0017
two hybrid	0018
coimmunoprecipitation	0019
transmission electron microscopy	0020
cosedimentation	0027
cosedimentation in solution	0028
cosedimentation through density gradient	0029
cross-linking study	0030
protein cross-linking with a bifunctional reagent	0031
dynamic light scattering	0038
electron microscopy	0040
electron paramagnetic resonance	0042
far western blotting	0047
filamentous phage display	0048
filter binding	0049
fluorescence correlation spectroscopy	0052
fluorescence polarization spectroscopy	0053
fluorescence-activated cell sorting	0054
fluorescent resonance energy transfer	0055
isothermal titration calorimetry	0065
lambda phage display	0066
light scattering	0067
molecular sieving	0071
nuclear magnetic resonance	0077
peptide array	0081
phage display	0084
protein array	0089
protein complementation assay	0090
chromatography technology	0091
reverse ras recruitment system	0097
scintillation proyimity assay	0099
static light scattering	0104
surface plasmon resonance	0107
t7 phage display	0108
dihydrofolate reductase reconstruction	0111
ubiquitin reconstruction	0112
x-ray crystallography	0114
yeast display	0115
ion eychange chromatography	0226
reverse phase chromatography	0227

green fluorescence protein complementation assay	0229
mammalian protein protein interaction trap	0231
transcriptional complementation assay	0232
blue native page	0276
ley-a dimerization assay	0369
toy-r dimerization assay	0370
two hybrid array	0397
two hybrid pooling approach	0398
two hybrid fragment pooling approach	0399
comigration in non denaturing gel electrophoresis	0404
competition binding	0405
deacetylase assay	0406
electron tomography	0410
enzyme linked immunosorbent assay	0411
enzymatic study	0415
fluorescence microscopy	0416
kinase homogeneous time resolved fluorescence	0420
in-gel kinase assay	0423
protein kinase assay	0424
phosphatase assay	0434
protease assay	0435
saturation binding	0440
homogeneous time resolved fluorescence	0510
methyltransferase assay	0515
methyltransferase radiometric assay	0516
enzymatic footprinting	0605
lambda repressor two hybrid	0655
antibody array	0678
reverse two hybrid	0726
gal4 vp16 complementation	0728
luminescence based mammalian interactome mapping	0729
comigration in gel electrophoresis	0807
comigration in sds page	0808
bimolecular fluorescence complementation	0809
y-ray fiber diffraction	0825
y ray scattering	0826
phosphotransfer assay	0841
immunodepleted coimmunoprecipitation	0858
intermolecular force	0859
demethylase assay	0870
atomic force microscopy	0872
acetylation assay	0889
surface plasmon resonance array	0921
polymerization	0953

Co-complex Associations	
Experimental Evidence	**PSI-MI Evidence Code**
affinity chromatography technology	0004
anti bait coimmunoprecipitation	0006
anti tag coimmunoprecipitation	0007
mass spectrometry studies of complexes	0069
pull down	0096
affinity technology	0400
tandem affinity purification	0676

Table 2.2

Binary Interactions		
MIPS Evidence Code	**MIPS Description**	**Corresponding PSI-MI Code**
902.01.01.02.01.01	co-immunoprecipitation	0019
902.01.01.02.01.03	centrifugation	0027
902.01.01.02.01.05.01	cross linking, chemical	0031
902.01.01.02.01.05.02	cross linking, UV	0430
902.01.01.02.01.06	in vitro reconstitution	0492
902.01.01.02.01.07	two hybrid	0018
902.01.01.02.01.08	overlay	0047
902.01.01.02.01.09.01	FRET	0055
902.01.01.02.01.09.02	scintillation proximity assay	0099
902.01.01.02.01.10	surface plasmon resonance	0107
902.01.01.02.01.11	phage display	0084
902.01.01.02.01.13.01	electron microscopy	0040
902.01.01.02.01.13.02	NMR	0077

Co-complex Associations		
MIPS Evidence Code	**MIPS Description**	**Corresponding PSI-MI Code**
902.01.01.02.01	physical	0013
902.01.01.02.01.01.02	epitope tag co-ip	0007
902.01.01.02.01.02	affinity chromatography	0004
902.01.01.02.01.02.01	affinity chromatography, native	0004
902.01.01.02.01.02.02	affinity tag chromatography	0004

TABLE 2.3

Binary Interactions		
VisAnt Evidence Code	**VisAnt Description**	**Corresponding PSI-MI Code**
M0010	Co-immunoprecipitation	0019
M0011	Co-sedimentation	0027
M0012	Competition binding	0405
M0013	Copurification	0025
M0014	Cross-linking studies	0030
M0015	Electron microscopy	0040
M0018	Molecular sieving	0071
M0021	Western blot	0113
M0024	Immunoprecipitation	0019
M0026	In vitro binding	0492
M0029	Monoclonal antibody	0671
M0032	Sizing Column	0071
M0033	Cosedimentation	0027
M0034	Two-hybrid	0018
M0035	X-ray	0114
M0049	Surface plasmon resonance	0921
M0050	Phage display	0084
M0051	ELISA	0411
M0052	Fluorescence technology	0051
M0053	Filter binding	0928
M0060	Far western	0047
M0061	Resonance energy transfer	0055
M0062	Electron microscopy	0040
M0066	Enzymatic study	0415
M0068	Protein array	0089
M0069	Protein complementation assay	0090
M0070	NMR	0077
M0071	X-ray crystallography	0114
M0079	Co-fractionation	0027
M0085	Chromatography	0091
M0092	Peptide array	0081
M0095	Protein kinase assay	0424
M0096	Blue native PAGE	0276
M0097	Comigration in gel electophoresis	0404
M0100	Ubiquitin reconstruction	0112
M0101	Phosphatase assay	0434
M0103	Isothermal titration calorimetry	0065

Co-complex Associations		
VisAnt Evidence Code	**VisAnt Description**	**Corresponding PSI-MI Code**
M0006	Affinity column	0400
M0028	Mass spectrometry of complex	0069
M0044	Affinity precipitation	0400
M0045	Affinity technology	0400
M0065	Anti-tag co-IP	0007
M0067	Pull down	0096

M0074	Reconstituted complex	0069
M0088	Tandem affinity purification	0676
M0089	Anti-bait Co-IP	0006
M5001	Tandem affinity mass spectrometry	0032

CHAPTER 3

Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks

In the following chapter, we examine human disease mutations in the context of structurally-resolved protein networks. I am the sole first author of the manuscript resulting from this chapter (Das et al Human Mutation 2014). I am also a co-first author on another related manuscript (Wang*, Wei*, Thijssen*, Das* et al Nature Biotechnology 2012 *=Equal contribution); I have not devoted a separate chapter to that paper in my thesis as the core concepts common to both papers are covered in this chapter.

3.1 ABSTRACT

With the rapid growth of structural genomics, numerous protein crystal structures have become available. However, the parallel increase in knowledge of the functional principles underlying biological processes, and more specifically the underlying molecular mechanisms of disease, has been less dramatic. This notwithstanding, the study of complex cellular networks has made possible the inference of protein functions on a large scale. Here, we combine the scale of network systems biology with the resolution of traditional structural biology to generate a large-scale atomic-resolution interactome-network comprising 3,398 interactions between 2,890 proteins with a well-defined interaction interface and interface residues for each interaction. Within the framework of this atomic-resolution network, we have explored the structural principles underlying variations causing human inherited disease. We find that in-frame pathogenic variations are enriched at both the interface and in the interacting domain, suggesting that variations not only at interface "hot-spots", but in the entire interacting domain can result in alterations of interactions. Further, the sites of pathogenic variations are closely related to the biophysical strength of the interactions they perturb. Finally, we show that biochemical alterations consequent to these variations are considerably more disruptive than evolutionary changes, with the most significant alterations at the protein interaction interface.

3.2 INTRODUCTION

The functions of a protein are inherently bound up with its three-dimensional structure – both regular secondary structures and disordered elements play a role in modulating function(Lahiry et al., 2010). Protein structures are often so intricate that even comparatively minor structural alterations can cause dramatic changes in function. Since such disruptions often lead to disease(Celli et al., 1999; Haberle et al., 2011), a significant amount of effort has been invested in attempting to determine the principles underlying complex structure-function relationships in human proteins. To date, however, most of this effort has been directed towards understanding how individual folds, domains or structural motifs carry out specific cellular functions(Andreeva et al., 2008; Pearl et al., 2005). Furthermore, most proteins carry out their functions by interacting with other proteins, all of which are part of a complex cellular network termed the "interactome"(Vidal, 2005; Vidal et al., 2011).

Recently, studies have become focused on how protein networks can be used to infer function and how changes in these networks can lead to human disease(Barabasi et al., 2011; Vidal et al., 2011). However, these efforts have had only limited success because protein networks are still incomplete(Vidal et al., 2011) and studies to date have treated proteins as mere graph-theoretical points in a mathematical network rather than as biological entities with their own structural details and chemical properties(de Souza, 2012; Wang et al., 2012). The importance of structural considerations has been well-recognized in predicting protein-protein interactions(Tuncbag et al., 2011; Zhang et al., 2012) and functional residues for each interaction(Marks et al., 2012). However, although structure has been widely employed to understand the evolutionary impact of single nucleotide polymorphisms (SNPs) (Bao and Cui,

2005; David et al., 2012; Sunyaev et al., 2001), the number of studies which have examined pathogenic variations in a structural context has been limited(Studer et al., 2013). To address this deficiency, we previously used a domain-level interaction network to show that in-frame pathogenic variations tend to be enriched within interacting domains(Wang et al., 2012). However, interacting domains comprise not only interface residues that are directly involved in the physical interaction between the two proteins but also other non-contact residues. In our earlier study, we did not differentiate between these two categories of amino acid residues. Since it is generally considered that interface residues mediate protein-protein interactions(Jones and Thornton, 1996), it is of paramount importance to examine the differential distribution of pathogenic variations between interface and non-interface residues within interacting domains. Moreover, only at the resolution of individual amino acid residues is it possible to ascertain structural (i.e., biophysical and biochemical) principles governing pathogenic processes.

To this end, we present here a large-scale atomic-resolution human interactome network by systematically identifying the interaction interfaces and corresponding residues mediating all interactions with available co-crystal structures in the Protein Data Bank (PDB)(Berman et al., 2000). Using this atomic-resolution interactome network, we analyze the distribution of pathogenic variations in different regions of human proteins focusing on interface and non-contact residues within interacting domains. We also explore how the locational specificity of these variations is directly associated with the strength of the interactions they disrupt. Finally, we examine biochemical properties of human pathogenic variations and compare them to their evolutionary counterparts.

3.3 RESULTS

Atomic-resolution structural analysis of pathogenic variations and their molecular mechanisms

Pathogenic variations belong to two broad categories – in-frame variations (both missense variations and in-frame microinsertions and microdeletions) and truncating variations (both nonsense point variations and frameshift insertions or deletions)(Zhong et al., 2009). We previously found that in-frame pathogenic variations are non-randomly distributed in proteins – indeed, they tend to be enriched within interacting domains. On the other hand, truncating variations do not show any particular trend with regard to their distribution in different parts of the protein(Wang et al., 2012).

It has been commonly accepted that interface residues mediate interactions between proteins(Hu et al., 2000; Jones and Thornton, 1996). Moreover, it is generally believed that "only a small portion of interface residues, the so-called hot spot residues, contribute the most to the binding energy of the protein complex"(Assi et al., 2010). These hot-spots are often the targets of drug molecules(Wells and McClendon, 2007). Owing to the limits of resolution of our previous study(Wang et al., 2012), we were able to perform the investigation only at the domain level, not at the level of individual residues. Employing the newly derived atomic-resolution interactome network, we set out to systematically examine whether pathogenic variations tend to specifically alter interface residues, as our previous results suggested might be the case. This network is higher resolution that other structurally resolved networks(Khurana et al., 2013; Wang et al., 2012) as it reports not just interacting domains for 3,398 interactions, but individual amino acids residues mediating each interaction. We calculated the enrichment of in-frame variations at the

interaction interface, the remainder of the interacting domain, and the rest of the protein. We found that in-frame variations are enriched significantly both at the interface and in the remainder of the interacting domain (odds ratio = 1.67, $P < 10^{-3}$ for interface residues; odds ratio = 1.75, $P < 10^{-3}$ for the remainder of the interacting domain; Figure 1A). To confirm that the observed trends are robust, we performed the same calculations with only the fraction of the protein in the actual co-crystal, because in many cases the crystallized structure does not contain full-length proteins. 62.6% of all the pathogenic variations used for our calculations in Fig. 1a are present within co-crystal structures. Using only these variations, our results remain unchanged – in-frame variations are enriched at both the interface and in the remainder of the interacting domain even if we consider only residues depicted within the co-crystal structures. To assess the significance of a decrease in solvent accessibility, we used randomly chosen cutoffs – decreases of 0.5 $Å^2$, 2 $Å^2$ and 5 $Å^2$ in solvent accessible surface area to define 3 alternate sets of interface residues. Using these 3 sets of residues, we repeated our calculations in Fig. 1a. We find that our results remain unchanged with all 3 alternate sets of residues. This shows that our results are robust to the choice of cutoff for decrease in solvent accessible area to define interface residues. In fact, the sets of interface residues are very similar for any cutoff between 0.5 $Å^2$ to 5 $Å^2$.

Our result shows that it is not simply the interface residues, but rather the interacting domain in its entirety that plays an important role in pathology for many disease genes. As a negative control, we calculated the distribution of 94,084 missense non-synonymous single nucleotide polymorphisms (SNPs) from ESP6500 in 2,829 genes and found that these were distributed randomly across the protein (Figure 1B). Most genes contain relatively few pathogenic variations and SNPs. Moreover, there is no significant difference ($P = 0.33$) in the

distribution of pathogenic variations and SNPs across various genes, confirming that the differences observed in the distribution of disease-associated variants and SNPs are not due to gene-specific distribution biases. To further confirm that SNPs are indeed randomly distributed across proteins, we repeated our calculations with only those genes that contain at least one disease-associated variant (i.e., those genes used for the calculations in Figure 1A) and found that SNPs in these genes are also randomly distributed across the length of the protein. Moreover, even if we consider SNPs present only within co-crystal structures, we find that they are still randomly distributed across proteins.

We also note that in-frame variations outside the interface were enriched in buried residues (Figure 1C). The importance of buried residues in maintaining the overall stability of the protein is well established(Gromiha et al., 1999). It has been suggested that in-frame and truncating variations have distinct disruption modes – the former is likely to disrupt specific interactions whereas the latter usually leads to degradation of the entire protein leading to a loss of all interactors(Zhong et al., 2009). Our results suggest that even for in-frame variations, the possible molecular mechanisms by which variations at or near the interface (and distant from it) affect protein-protein interactions are likely to be distinct: those at the interface are more likely to alter specific interactions, thereby causing the mutated protein either to lose or acquire specific functions; by contrast, in-frame variations in other non-interacting regions are more likely to disrupt the core of the protein and lead to incorrect folding and/or degradation of the protein, resulting in the loss of all interactions for the mutated protein **(**Figure 1D**)**.

To further understand the effects of variations in the interacting domain outside the interface, we examined the effects of two disease-associated variants on the *PTS-PTS* interaction (Figure 2A). Using site-directed mutagenesis PCR, we introduced the two variants – R25Q and

R9C on *PTS*. Although the R25Q variant is located on the PTPS domain that mediates the *PTS-PTS* interaction, it is not at an interface residue. Using yeast two-hybrid, we confirmed that wild-type *PTS* interacts with itself (Figure 2B). However the R25Q variant disrupts this interaction (Figure 2B). On the other hand, the R9C variant lies outside the interface mediating the *PTS-PTS* interaction. Using yeast two-hybrid, we confirmed that this variant (R9C) does not affect the interaction (Figure 2B). This shows that variations in the interacting domain outside the interface can disrupt protein interactions, whereas the same interactions can remain unaffected by variants outside the corresponding interacting domains.

Moreover, using Western blotting, we confirm that all three variants are stable (Fig. 2b). Together, these results show that the R25Q variant causes an interaction-specific disruption – the *PTS-PTS* homodimeric interaction is lost due to a local structural alteration in the corresponding interacting domain. It has been previously shown that the enzymatic activity of the R25Q variant of *PTS* is reduced, but not completely abolished compared to the activity of WT *PTS*(Oppliger et al., 1995; Thony et al., 1994). Our results suggest a molecular mechanistic basis for this reduction – since the dimerization of *PTS* is important for its enzymatic activity(Oppliger et al., 1995; Thony et al., 1994), the pathogenic R25Q that disrupts the *PTS* homodimer reduces this activity. However, since the variant is stable, *PTS* still maintains part of its activity.

Pathogenic variation loci associated with interaction strength

To understand the biophysical mechanisms by which in-frame pathogenic variations alter specific interactions, we examined the relationship between the spatial distribution of the variations and the strength of the interactions they perturb. Here, we explored the biophysical strength of an interaction – the stronger the interaction, the higher the free energy difference

between the bound and unbound states of the proteins;(Noskov and Lim, 2001; Shi et al., 2006) by calculating the buried surface area of all the interactions in the atomic-resolution human interactome network. The most direct measure of interaction strength is the equilibrium association constant (K_a, inverse of the equilibrium dissociation constant K_d). However, it is difficult to measure K_a in a high-throughput fashion and the amount of experimental K_a data is limited to a handful of human protein-protein interactions.

It has been suggested that the strength of an interaction can be measured by its buried surface area in the co-crystal structure(Jones and Thornton, 1996). To validate this postulate, we classified all interactions in the network into three distinct categories on the basis of their buried surface area – low, medium and high. Using a genome-wide microarray analysis that measures the expression levels of human genes at different time points in the cell cycle(Whitfield et al., 2002), we calculated the enrichment in co-expression of proteins involved in these interactions. We found that interactions with high buried surface area are significantly more likely to be co-expressed than interactions with low buried surface area (P = 0.015, Figure 3A). It is well known that strong, stable interactions are more likely to be co-expressed than weak, transient interactions(von Mering et al., 2002; Yu et al., 2008). Our result confirms that protein-protein interactions mediated by high buried surface area are indeed stronger. Moreover, we calculated the fraction of these binary interactions independently for the three categories detected in stable protein complexes. We found that interactions with high buried surface area are significantly enriched in stable complexes, further supporting the conclusion that these are stronger interactions (Figure 3B). Finally, we calculated the correlation between K_a and buried surface area using SKEMPI, a database of binding free energy changes for interactions with supporting co-crystal structures(Moal and Fernandez-Recio, 2012). For all interactions in SKEMPI

involving wild-type human proteins, we calculated the correlation between K_a values and the buried surface area. We find that there is a significant correlation (ρ = 0.63, $P < 10^{-3}$ using a permutation test) between K_a and buried surface area, confirming that the latter is an appropriate surrogate for interaction strength.

Next, we determined the distribution of in-frame variations in different parts of the protein as a function of the strength of the interaction. We found that variations at the interface tend to disrupt strong interactions (odds ratio = 1.10, P = 0.005) whereas those in the rest of the protein outside the interacting domains tend to be enriched in weak interactions (odds ratio = 1.24, $P < 10^{-3}$; Figures 3C-3E). As a control, we also computed the distribution of SNPs in different parts of the protein as a function of interaction strength. We found that SNPs at the interface and away from the interface are both randomly distributed with respect to interaction strength. Our results therefore suggest that there is a relationship between the location of the disease variation and the biophysical strength of the interactions it disrupts. Because pathogenic variations are enriched at the interaction interface and interface variations selectively affect biophysically strong interactions, we surmise that many strong interactions within stable protein complexes involved in key cellular functions are likely to be preferentially disrupted in human disease. This provides a molecular-level biophysical explanation for the results of previous studies which have suggested that protein complexes are useful predictors for discovering unknown disease genes(Fraser and Plotkin, 2007).

Significant alterations in structural and biochemical properties of amino acids involved in human inherited disease

To systematically explore the structural properties of human pathogenic variations, we analyzed relationships between the properties of these variations and their accessibility in the protein. Amino acids may be classified as either accessible or inaccessible. Using the Janin accessibility scale(Janin, 1979), we then calculated the proportion of accessibility-altering in-frame missense disease-associated variations (i.e., point variations that cause an accessible wild-type amino acid to be changed to an inaccessible amino acid or *vice versa*) in different parts of the protein. These variations are most likely to cause dramatic changes to the configuration of the interface because the local structural configuration is drastically altered. Since disease-associated variations in different parts of the protein may exert their effects via different pathophysiological mechanisms, we normalized our results by calculating the ratio of accessibility-altering in-frame variations against a background distribution of putatively neutral SNPs that are characterized by a similar change in their accessibility. Since these SNPs are uniformly distributed throughout the protein (**Fig. 1b**), this gives us an idea of the relative propensity of disease-associated variations to be significantly accessibility-altering. We found that at both surface and buried residues, and indeed in all parts of the protein, accessibility-altering variations are significantly more likely to occur in pathogenic variations as opposed to putatively neutral variants in the general population ($P < 10^{-3}$; Figure 4A).

We also examined amino acid substitutions in terms of their change in polarity. We calculated the proportion of polarity-altering in-frame missense disease-associated variations (i.e., those that cause a polar wild-type amino acid to change to a non-polar amino acid or *vice versa*). We note that these alterations also follow a similar trend – at both surface and buried

residues, and in all regions of the protein, polarity-altering variations are significantly more likely to occur in disease as opposed to putatively neutral variants in the population ($P < 10^{-3}$; Figure 4A). This suggests that disease-associated variations are biochemically more destabilizing to the protein than benign variants in the population.

To further understand how disease-associated variations differ in terms of their biochemical properties from changes that have been fixed over the course of evolutionary time, we calculated the relative enrichment of all possible pairs of amino-acid changes for disease-associated in-frame missense variations over those that have occurred during evolution. We obtained the probabilities of amino acid changes occurring during evolution from a recently updated version of the Dayhoff matrices(Kosiol and Goldman, 2005). We compared these amino-acid changes to in-frame disease variations occurring throughput the protein (Figure 4B). We found that disease-associated variations generally tend to alter accessibility of the wild-type amino acid whereas evolutionary changes tend to preserve it ($P = 0.010$; Figure 4C**)**. Our findings contrast with previous reports of significant correlations between amino acid variations in genetic disease and evolution(Wu et al., 2007). To further understand the specific differences in the distribution of variations in different parts of the protein, we determined which variations were enriched at least 2-fold at the interaction interface compared to other regions of the protein (Figure 4D). We found that these interface variations are significantly more likely to change the accessibility of the amino acid involved ($P = 0.034$), with the most dramatic changes occurring with those variations with the highest enrichment.

By way of an example, a K143I variation at the interaction interface of RNASEH2B and RNASEH2C has been shown to be associated with a human auto-inflammatory disorder, Aicardi-Goutières syndrome(Reijns et al., 2011). This variation causes a major change in

structural and biochemical properties, leading to a significant structural modification at the interface that specifically alters the wild-type interaction (Figure 4E). These results further validate our finding that pathogenic variations tend to be more disruptive than random evolutionary changes, with those occurring at the protein interface causing the most drastic changes, enough to perturb even strong interactions.

3.4 DISCUSSION

In this study, we build and use an atomic-resolution human protein interactome network to improve our understanding of the structural principles and molecular mechanisms of pathogenic variations that perturb protein-protein interactions leading to disease. We find that in-frame variations are significantly enriched both at the interaction interface as well as in the remainder of the corresponding interacting domain. Thus, it is not just the residues at the interface which serve as the key mediators of interactions(Hu et al., 2000; Jones and Thornton, 1996), variations outside the interface but within the interacting domain are capable of altering protein-protein interactions. Our findings suggest that it is the alteration of specific interactions by in-frame variations within the entire interacting domain that is a major molecular determinant of human inherited disease. Moreover, we show that there are important biochemical and biophysical differences between variations at the interface and those located in the remainder of the protein molecule. Specifically, we find that the locations of pathogenic variations are associated with the strength of interactions – those at the interface tend to selectively disrupt stronger interactions. One mechanistic explanation for such a phenomenon is the tendency for variations enriched at the interface (as compared to other parts of the protein) to cause the most dramatic changes in

their structural and biochemical properties. Analyses at the level of individual amino acids are only possible with atomic-resolution interactome networks. Our findings suggest that the structurally guided prioritization of pathogenic variations identified in large-scale sequencing studies using an atomic-resolution network might be useful in the context of informing follow-up experiments.

The coverage of the atomic-resolution human protein interactome network is limited by the number of co-crystal structures currently available in PDB. As more co-crystal structures become available(Chandonia and Brenner, 2006), the same principles developed here can be readily applied to reveal additional specific structural mechanisms underlying pathogenic variations. Our work further underscores the importance of the exploration of all possible domain architectures by structural genomics consortia(Editorial, 2007). Using our methodology on a more complete set of structural folds is likely to generate reliable direct atomic-resolution target sites for structurally-aided rational drug design, and has the potential to overcome the difficulties routinely encountered due to the paucity of well-elucidated structural targets(Tanrikulu and Schneider, 2008; Xie and Bourne, 2011).

3.5 MATERIALS AND METHODS

Calculating atomic-resolution interface residues for human protein interactions

To calculate atomic-resolution interaction interfaces, we systematically examined a comprehensive list of 7,340 PDB co-crystal structures and were able to determine atomic-resolution interaction interfaces for 3,398 unique human protein-protein interactions between 2,890 proteins. To define the interface, we used a water molecule of diameter 1.4Å as a probe

and calculated the relative solvent accessible surface areas of the interacting pair as well as the individual proteins involved in the interaction(Hubbard and Thornton, 1993). All calculations were performed using Naccess(Hubbard and Thornton, 1993). Residues whose relative accessibilities changed by more than 1Å^2 were considered as potential interface residues. Amino acids at the interface reside on the surfaces of the corresponding proteins, but tend to become buried in the co-crystal structure as the two proteins bind to each other. It follows that these residues should experience a significant decrease in accessible surface area when the bound and the unbound states of the protein chains are compared (Franzosa and Xia, 2011). In most cases, our calculations incorporated multiple instances of the same interaction from different chains within the same PDB structure or entirely different PDB structures representing the same interaction. This allows us to accurately determine the exact interface, and normalize differences due to specific crystallization conditions(Chayen and Saridakis, 2008). We take the union over all such instances subject to the constraint that the particular protein pair contains at least five interface residues for both interacting proteins. This ensures that all the interfaces included in our calculations represent significant regions of molecular contact, eliminating potential crystal contacts. Furthermore, 1,689/3,398 (49.7%) interactions used in this study have been detected by at least one other assay and were reported independently in a separate publication. This confirms that interactions used in this study are not only real but also reproducible using other assays.

To further refine the set of identified interface residues, we required that they be necessarily present on the surface of the protein. To determine which residues were on the surface, we calculated the fraction of surface area for each residue in the individual protein chains that was accessible to the water molecule probe defined above(Hubbard and Thornton, 1993). If more than 15% of the total surface area for a particular residue was accessible to the water molecule

probe, we defined that particular amino acid to be on the surface, otherwise it was considered to be buried. Using these two criteria, for each interaction we obtained a set of 141,686 residues that represent the interface for 3,398 interactions from 7,340 atomic-resolution co-crystal structures. The fraction of homomeric interactions to heteromeric interactions is ~2:1 as the PDB is enriched for homodimers as compared to heterodimers.

Identifying interacting domains for each interaction

We generated a list of putative interacting domains utilizing the "homology modeling approach" as described earlier(Meyer et al., 2013) using both 3did(Stein et al., 2011) and iPfam(Finn et al., 2005). However, some of the domain pairs identified as interacting by 3did and iPfam for a particular protein pair may not have been supported by the corresponding co-crystal structure as they may have been inferred from other co-crystal structures. Therefore, to avoid potential false positives, we additionally required that these domains should contain at least one interface residue for them to be considered as interacting domains. Moreover, the set of interacting domains inferred by 3did and iPfam were not always complete. For our analysis, we took advantage of our own atomic-resolution interface calculations to identify a comprehensive set of interacting domains for each co-crystal structure, and included interacting domains not identified by 3did or iPfam if they had five or more interface residues.

Compiling a comprehensive list of pathogenic variations and SNPs

We compiled a comprehensive list of 94,476 pathogenic variations from HGMD(Stenson et al., 2013; Wang et al., 2012) as described earlier (Wang et al., 2012). We updated our earlier lists with a newer version of the HGMD dataset [HGMD Professional v.2012.2]. Specifically, we

used in-frame variations (both missense variations and in-frame microinsertions and microdeletions) classified as 'DM' in HGMD. For further analysis, we employed a total of 17,306 variations in 673 genes for which we were able to define at least one atomic-resolution interaction interface. We also compiled a set of non-synonymous SNPs from the Exome Sequencing Project (Fu et al., 2013) from which we derived a dataset of 94,084 SNPs in 2,829 genes for which we were able to define at least one atomic-resolution interaction interface.

Using our publicly available supplementary website, http://www.yulab.org/Supp/AtomInt, researchers can query interface residues for their favorite interaction.

Criteria used to choose *PTS-PTS* homodimeric interaction for experimental validation

The following criteria were used to choose the *PTS-PTS* homodimeric interaction for experimental validation of the effects of pathogenic variants within and outside the interacting domain:

a. the interaction is supported by a co-crystal structure
b. the wild-type *PTS* clone is available in our library.
c. the wild-type interaction (*PTS-PTS*) is amenable to testing in our yeast two-hybrid system
d. there is a pathogenic variation in the interacting domain but outside interface residues.
e. there is a different pathogenic variation outside both the interface residues and the interacting domain.

Generation of PTS variants

Wild-type *PTS* is obtained from the human ORFeome v8.1 collection(Yang et al., 2011). To generate the alleles R25Q and R9C corresponding to two different pathogenic variations, sequence-verified single-colony wild-type *PTS* and corresponding mutagenic primers (designed according to the protocol accompanying the Stratagene QuikChange Site-Directed Mutagenesis Kit #200518) were aliquoted together. Mutagenesis PCR was then performed as specified by the New England Biolabs (NEB) PCR protocol for Phusion polymerase (M0530L), noting that PCR was limited to 18 cycles. The samples were then digested by *Dpn*I (NEB R0176L) according to the manufacturer's manual. After digestion, samples were transformed into competent *E. coli* and then individually streaked onto LB plates containing spectinomycin to obtain single colonies. The generated clones were verified by Sanger sequencing.

Yeast two-hybrid

Y2H was done as previously described(Wang et al., 2012). WT *PTS* and both pathogenic variant alleles were transferred by Gateway LR reactions into our Y2H pDEST-AD and pDEST-DB vectors. DB-X and AD-Y plasmids were transformed individually into the Y2H strains *MAT*α Y8930 and *MAT***a** Y8800, respectively. Each of the DB-X *MAT*α transformants (wild-type and variants) were then mated against corresponding AD-Y *MAT***a** transformants (wild-type and variants), including inoculation of AD-Y and DB-X yeast cultures, mating on YEPD media (incubated overnight at 30 °C), and replica-plating onto selective Synthetic Complete media lacking leucine, tryptophan, and histidine, and supplemented with 1 mM of 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT), SC-Leu-His+3AT plates containing 1 mg/l cycloheximide (SC-Leu-His+3AT+CHX), SC-Leu-Trp-Adenine (Ade) plates, and SC-Leu-Ade+CHX plates to test for CHX-sensitive expression of the *LYS2::GAL1-HIS3* and *GAL2-ADE2* reporter genes. The plates were incubated overnight at 30 °C and replica-cleaned the following day. Plates were then

incubated for another three days, after which positive colonies were scored as those that grow on SC-Leu-Trp-His+3AT and/or on SC-Leu-Trp-Ade, but not on SC-Leu-His+3AT+CHX or on SC-Leu-Ade+CHX. Disruption of an interaction by a variation was defined as significant reduction of growth when compared to the Y2H phenotype of the wild-type *PTS-PTS* interaction.

Western blotting

Wild-type and both *PTS* variants were cloned into MSCV-N-FLAG-HA-IRES-Puro vector(Behrends et al., 2010) and transfected into HEK293T cells to express HA-tagged wild-type and mutated proteins. HEK293T cells were maintained in complete DMEM medium supplemented with 10% FBS. Cells were transfected with Lipofectamine 2000 (Invitrogen) at a 5:1 (µl/µg) ratio with DNA and harvested 24 hrs after transfection. Cells were gently washed three times in PBS and then resuspended using 200 µl 1% NP-40 lysis buffer [1% Nonidet P-40, 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1× EDTA-free Complete Protease Inhibitor tablet (Roche 05056489001)] and kept on ice for 30 mins. Extracts were cleared by centrifugation for 10 min at 13,000 rpm at 4 °C. 25 µl of extracts were mixed with 6X loading buffer and subjected to SDS-PAGE. Proteins were then transferred from the gel onto PVDF membranes (GE Healthcare RPN303F). Anti-HA (Sigma H9658) and anti-γ-tubulin (Sigma T5192) were used at 1:3,000 dilutions for immunoblotting analysis. Blotting signal was developed with Novex ECL HRP chemiluminescent substrate reagent kit (Invitrogen WP20005) and captured with Amersham Hyperfilm MP (GE Healthcare 28906843).

3.6 FIGURE LEGENDS

Figure 3.1 Atomic-resolution structural analysis of pathogenic variations. (a) Odds ratios for the distribution of in-frame variations in different locations on proteins in our atomic-resolution interactome network. $^{**}P < 10^{-3}$. *P*-values calculated using *Z*-tests for the log odds ratios. (b) Odds ratios for the distribution of non-synonymous SNPs in different locations on proteins in our atomic-resolution interactome network. (c) Enrichment of in-frame variations in buried residues. $^{**}P < 10^{-3}$ Error bars indicate ± SE. (d) Different mechanistic modes of disruption for variations in different structural environments – variations at the surface are likely to cause interaction-specific disruptions, whereas those buried in the core of the protein are likely to destabilize the entire protein.

Figure 3.2 (a) Crystal structure (PDB id: 3I2B) depicting a R25Q variation in the *PTS-PTS* interacting domain but not at an interface residue and a R9C variation outside the interaction interface. (b) Y2H assay illustrating that the R25Q variation disrupts the *PTS-PTS* interaction whereas the R9C variation does not affect the interaction.

Figure 3.3 Loci of disease variations associated with interaction strength. (a) Co-expression profiles for interactions with low, medium and high buried surface areas. (b) Enrichment of interactions with low, medium and high buried surface areas in stable complexes. (c) Odds ratios for the distribution of in-frame variations at the interface in interactions with low, medium and high buried surface areas. $^{*}P < 10^{-3}$ (d) Odds ration of in-frame variations in the remainder of the interacting domain in interactions with low, medium and high buried surface areas. (e) Odds ratio

of in-frame variations in the rest of the protein in interactions with low, medium and high buried surface areas. *$P < 10^{-3}$. Error bars indicate ± SE.

Figure 3.4 Alterations of biochemical properties of individual amino acids in disease. (a) Enrichment of disease variations that alter the structural (accessibility) and biochemical (polarity) properties of amino acids as compared to SNPs. *$P < 10^{-3}$. Error bars indicate ± SE. (b) Relative enrichment of all pairs of amino-acid changes in human pathological variations as compared to changes which occurred, and which were fixed, during the course of evolution (gray indicates that these pathogenic variations are not observed). (c) Pairs of amino-acid changes enriched in pathogenic missense variations and changes that occurred during evolution (shaded pairs undergo significant change in biochemical properties). (d) Pairs of amino-acid changes enriched at the atomic-resolution interaction interface (shaded pairs undergo significant change in biochemical properties). (e) An example of the alteration of the interaction interface between RNASEH2B and RNASEH2C by a variation (K143I) in RNASEH2C that significantly alters biochemical properties (in the circular panel, the blue residue is the wild-type K and the red residue is the pathogenic variant I).

3.7 REFERENCES

Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res *36*, D419-425.

Assi, S.A., Tanaka, T., Rabbitts, T.H., and Fernandez-Fuentes, N. (2010). PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces. Nucleic Acids Res *38*, e86.

Bao, L., and Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics *21*, 2185-2190.

Barabasi, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat Rev Genet *12*, 56-68.

Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. Nature *466*, 68-76.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Celli, J., Duijf, P., Hamel, B.C., Bamshad, M., Kramer, B., Smits, A.P., Newbury-Ecob, R., Hennekam, R.C., Van Buggenhout, G., van Haeringen, A.*, et al.* (1999). Heterozygous germline mutations in the p53 homolog p63 are the cause of EEC syndrome. Cell *99*, 143-153.

Chandonia, J.M., and Brenner, S.E. (2006). The impact of structural genomics: expectations and outcomes. Science *311*, 347-351.

Chayen, N.E., and Saridakis, E. (2008). Protein crystallization: from purified protein to diffraction-quality crystal. Nat Methods *5*, 147-153.

David, A., Razali, R., Wass, M.N., and Sternberg, M.J. (2012). Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat *33*, 359-363.

de Souza, N. (2012). Systems biology: A bird's-eye view of disease. Nat Meth *9*, 220-221.

Editorial (2007). Looking ahead with structural genomics. Nat Struct Mol Biol *14*, 1.

Finn, R.D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics *21*, 410-412.

Franzosa, E.A., and Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. Proc Natl Acad Sci U S A *108*, 10538-10543.

Fraser, H.B., and Plotkin, J.B. (2007). Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol *8*, R252.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Gromiha, M.M., Oobatake, M., Kono, H., Uedaira, H., and Sarai, A. (1999). Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. Protein Eng *12*, 549-555.

Haberle, J., Shchelochkov, O.A., Wang, J., Katsonis, P., Hall, L., Reiss, S., Eeds, A., Willis, A., Yadav, M., Summar, S.*, et al.* (2011). Molecular defects in human carbamoy phosphate synthetase I: mutational spectrum, diagnostic and protein structure considerations. Hum Mutat *32*, 579-589.

Hu, Z., Ma, B., Wolfson, H., and Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. Proteins *39*, 331-342.

Hubbard, S.J., and Thornton, J.M. (1993). 'NACCESS', computer program.

Janin, J. (1979). Surface and inside volumes in globular proteins. Nature *277*, 491-492.

Jones, S., and Thornton, J.M. (1996). Principles of protein-protein interactions. Proc Natl Acad Sci U S A *93*, 13-20.

Khurana, E., Fu, Y., Chen, J., and Gerstein, M. (2013). Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol *9*, e1002886.

Kosiol, C., and Goldman, N. (2005). Different versions of the Dayhoff rate matrix. Mol Biol Evol *22*, 193-199.

Lahiry, P., Torkamani, A., Schork, N.J., and Hegele, R.A. (2010). Kinase mutations in human disease: interpreting genotype-phenotype relationships. Nat Rev Genet *11*, 60-74.

Marks, D.S., Hopf, T.A., and Sander, C. (2012). Protein structure prediction from sequence variation. Nat Biotechnol *30*, 1072-1080.

Meyer, M.J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics *29*, 1577-1579.

Moal, I.H., and Fernandez-Recio, J. (2012). SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. Bioinformatics *28*, 2600-2607.

Noskov, S.Y., and Lim, C. (2001). Free energy decomposition of protein-protein interactions. Biophys J *81*, 737-750.

Oppliger, T., Thony, B., Nar, H., Burgisser, D., Huber, R., Heizmann, C.W., and Blau, N. (1995). Structural and functional consequences of mutations in 6-pyruvoyltetrahydropterin

synthase causing hyperphenylalaninemia in humans. Phosphorylation is a requirement for in vivo activity. J Biol Chem *270*, 29498-29506.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D.*, et al.* (2005). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res *33*, D247-251.

Reijns, M.A., Bubeck, D., Gibson, L.C., Graham, S.C., Baillie, G.S., Jones, E.Y., and Jackson, A.P. (2011). The structure of the human RNase H2 complex defines key interaction interfaces relevant to enzyme function and human disease. J Biol Chem *286*, 10530-10539.

Shi, Y.Y., Miller, G.A., Qian, H., and Bomsztyk, K. (2006). Free-energy distribution of binary protein-protein binding suggests cross-species interactome differences. Proc Natl Acad Sci U S A *103*, 11527-11532.

Stein, A., Ceol, A., and Aloy, P. (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res *39*, D718-723.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A.D., and Cooper, D.N. (2013). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet.

Studer, R.A., Dessailly, B.H., and Orengo, C.A. (2013). Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem J *449*, 581-594.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A.S., and Bork, P. (2001). Prediction of deleterious human alleles. Hum Mol Genet *10*, 591-597.

Tanrikulu, Y., and Schneider, G. (2008). Pseudoreceptor models in drug design: bridging ligand- and receptor-based virtual screening. Nat Rev Drug Discov *7*, 667-677.

Thony, B., Leimbacher, W., Blau, N., Harvie, A., and Heizmann, C.W. (1994). Hyperphenylalaninemia due to defects in tetrahydrobiopterin metabolism: molecular characterization of mutations in 6-pyruvoyl-tetrahydropterin synthase. Am J Hum Genet *54*, 782-792.

Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat Protoc *6*, 1341-1354.

Vidal, M. (2005). Interactome modeling. FEBS Lett *579*, 1834-1838.

Vidal, M., Cusick, M.E., and Barabasi, A.L. (2011). Interactome networks and human disease. Cell *144*, 986-998.

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

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol *30*, 159-164.

Wells, J.A., and McClendon, C.L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature *450*, 1001-1009.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O.*, et al.* (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell *13*, 1977-2000.

Wu, H., Ma, B.G., Zhao, J.T., and Zhang, H.Y. (2007). How similar are amino acid mutations in human genetic diseases and evolution. Biochem Biophys Res Commun *362*, 233-237.

Xie, L., and Bourne, P.E. (2011). Structure-based systems biology for analyzing off-target binding. Curr Opin Struct Biol *21*, 189-199.

Yang, X., Boehm, J.S., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T., Green, T.M.*, et al.* (2011). A public genome-scale lentiviral expression library of human ORFs. Nat Methods *8*, 659-661.

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

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T.*, et al.* (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature *490*, 556-560.

Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D.*, et al.* (2009). Edgetic perturbation models of human inherited disorders. Mol Syst Biol *5*, 321.

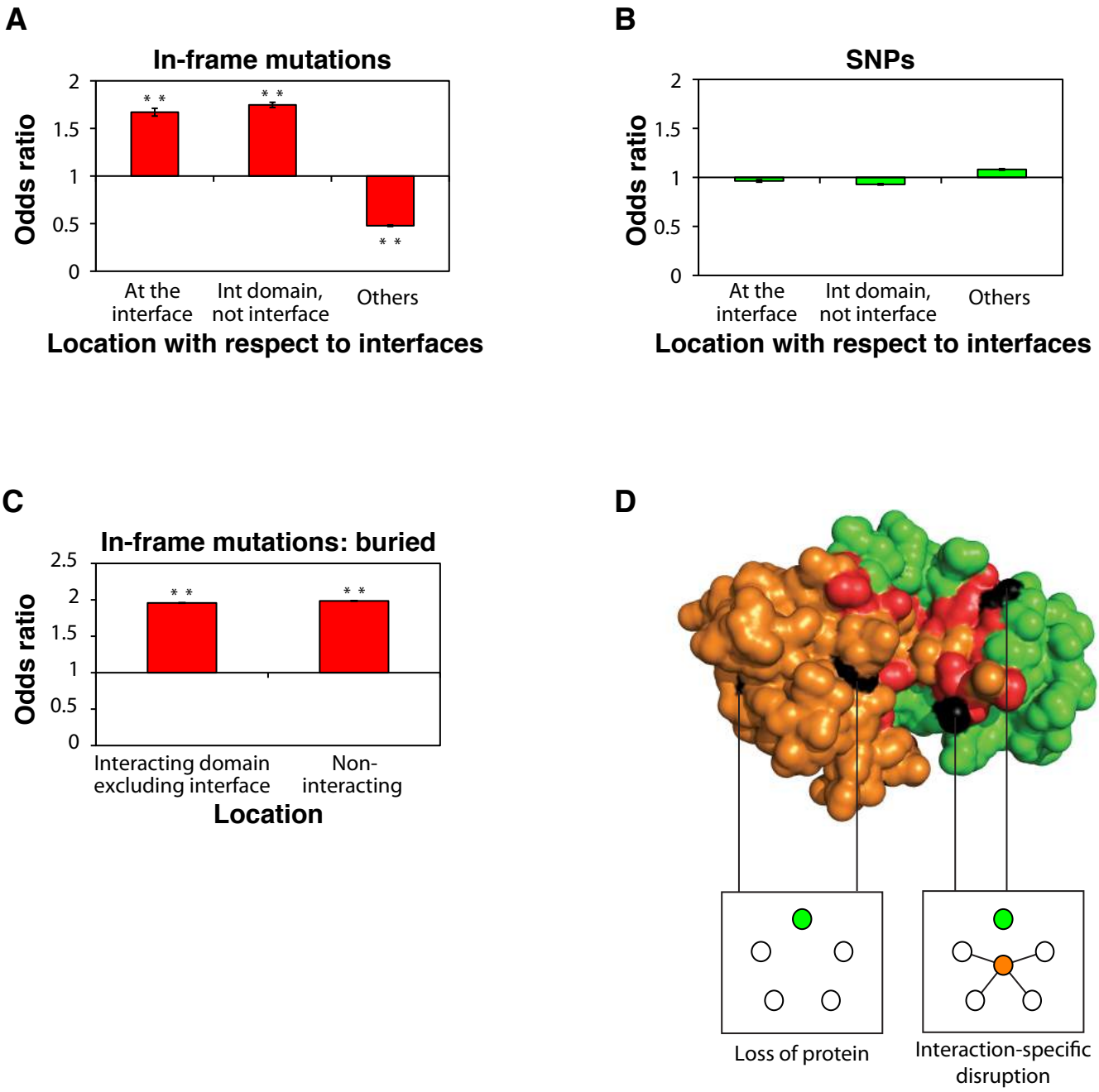
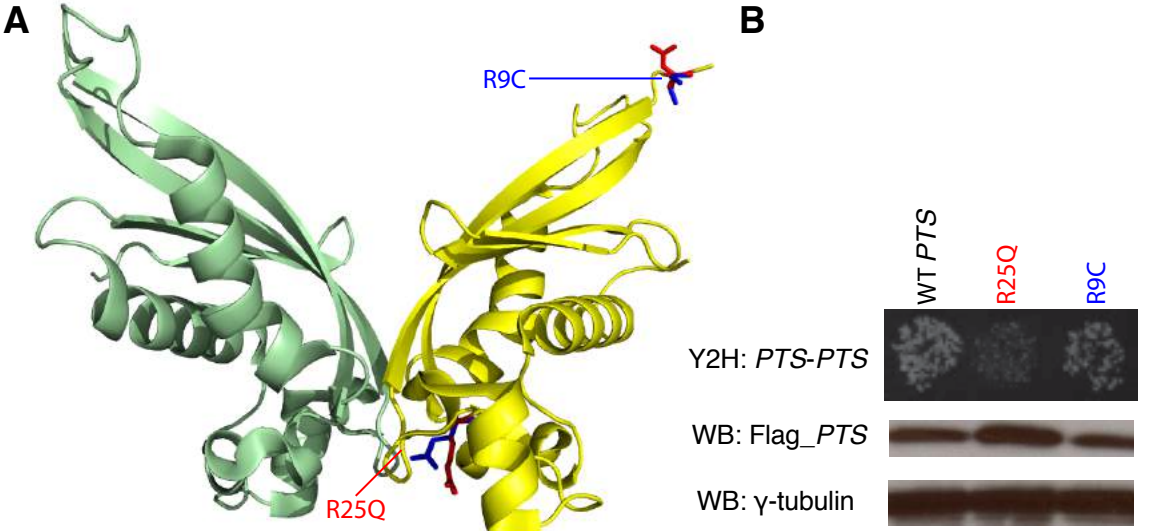

A
In-frame mutations
Odds ratio
2
1.5
1
0.5
0
* *
* *
* *
At the interface
Int domain, not interface
Others
Location with respect to interfaces
B
SNPs
Odds ratio
2
1.5
1
0.5
0
At the interface
Int domain, not interface
Others
Location with respect to interfaces
C
In-frame mutations: buried
Odds ratio
2.5
2
1.5
1
0.5
0
* *
* *
Interacting domain excluding interface
Non-interacting
Location
D
Loss of protein
Interaction-specific disruption

Figure 3.1



R25Q - not an interface residue,
but on the *PTS-PTS* interacting domain

R9C - outside the *PTS-PTS* interacting domain

Figure 3.2

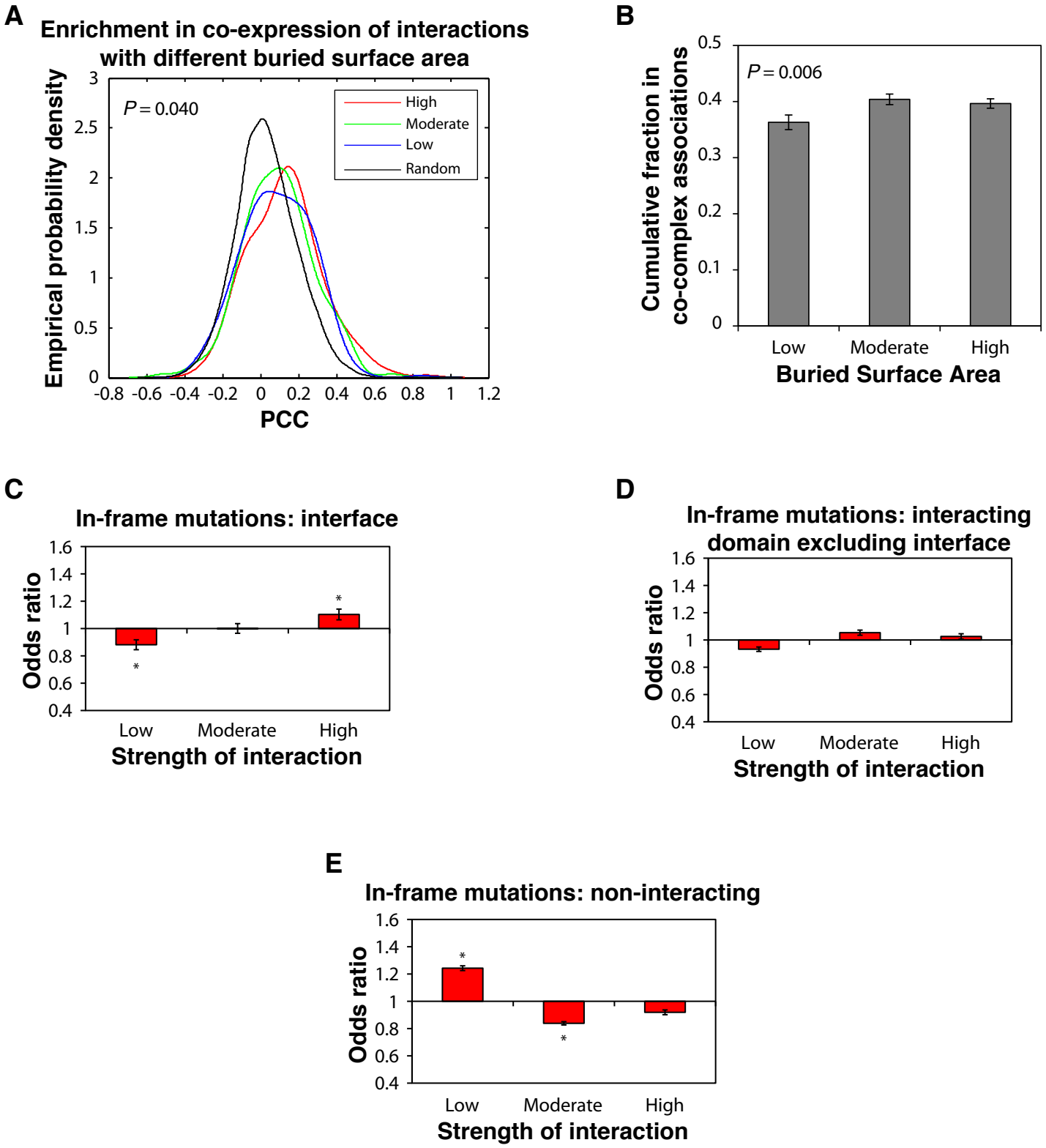
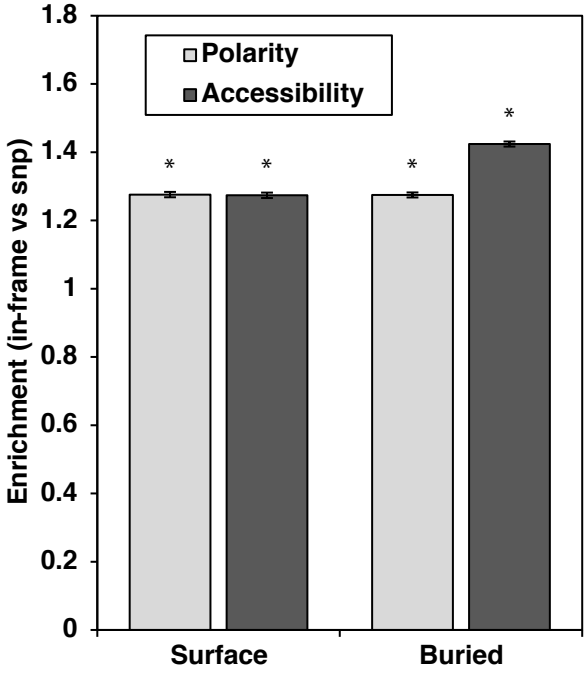
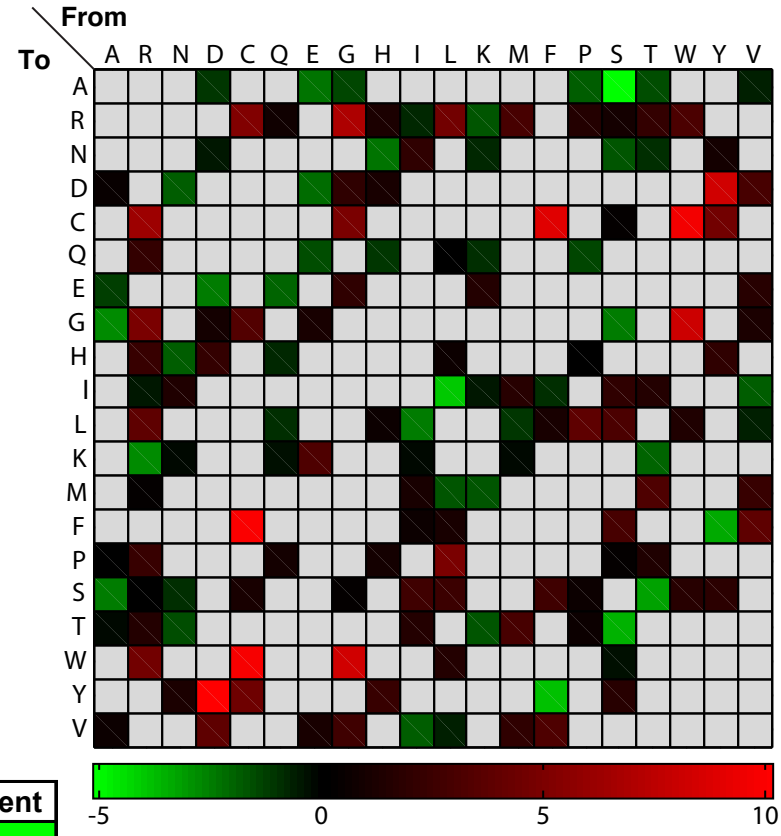

A
Enrichment in co-expression of interactions with different buried surface area
Empirical probability density
P = 0.040
High
Moderate
Low
Random
PCC
B
Cumulative fraction in co-complex associations
P = 0.006
Low
Moderate
High
Buried Surface Area
C
In-frame mutations: interface
Odds ratio
Low
Moderate
High
Strength of interaction
D
In-frame mutations: interacting domain excluding interface
Odds ratio
Low
Moderate
High
Strength of interaction
E
In-frame mutations: non-interacting
Odds ratio
Low
Moderate
High
Strength of interaction

Figure 3.3

A



B



C

	WT	Mut	Enrichment
	D	Y	1130.27
	C	F	961.38
	C	W	857.45
	W	C	753.51
	F	C	467.70
	Y	D	305.30
	G	W	285.82
	W	G	279.32
*	G	R	112.63
*	R	C	75.42
*	C	R	37.42
*	R	G	29.81
	L	P	29.05
	G	C	27.67
*	L	R	24.65
*	R	W	23.24
*	Y	C	22.93
*	C	Y	22.15

WT	Mut	Enrichment
S	A	0.03
L	I	0.06
F	Y	0.07
S	T	0.08
Y	F	0.10
T	S	0.10
A	G	0.14
R	K	0.15
S	G	0.17
D	E	0.18
A	S	0.18
I	L	0.18

$P = 0.010$

D

	WT	Mut	Enrichment
*	K	I	4.47
*	Q	L	4.01
	W	L	3.39
	P	Q	3.10
	K	R	2.72
	F	Y	2.13
	P	H	2.13
	R	K	2.13
*	K	T	2.09
	I	L	1.96
*	R	H	1.90

$P = 0.034$

E

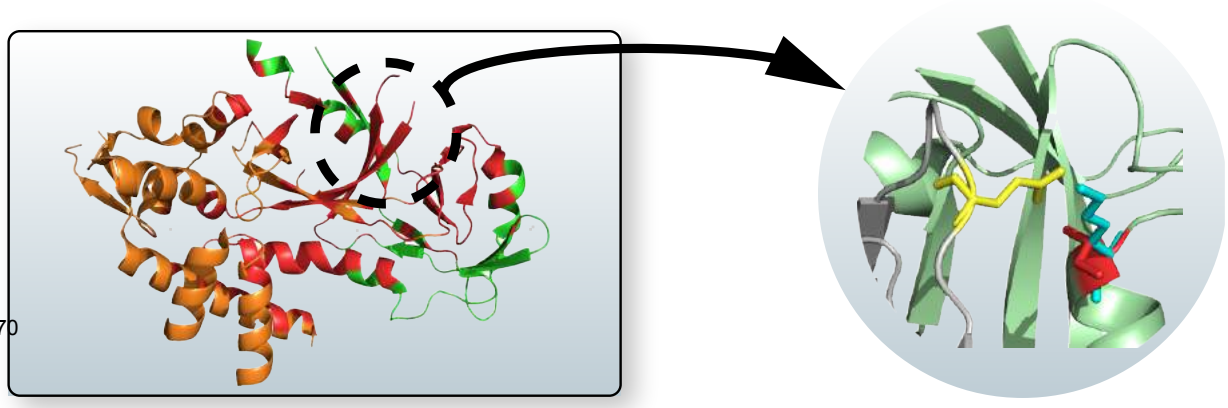


Figure 3.4

CHAPTER 4

A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations

In the following chapter, we describe a high-throughput site-directed mutagenesis pipeline to generate thousands of mutations and examine their effects on protein stability and interactions. I am a co-first author of the paper resulting from this chapter (Wei*, Das* et al PLoS Genetics 2014, *=Equal contribution) and performed all computational analyses. The first author of the Xiaomu Wei led the experiments, along with other co-first authors Robert Fragoza and Jin Liang.

4.1 ABSTRACT

Understanding the functional relevance of DNA variants is essential for all exome and genome sequencing projects. However, current mutagenesis cloning protocols require Sanger sequencing, and thus are prohibitively costly and labor-intensive. We describe a massively-parallel site-directed mutagenesis approach, "Clone-seq", leveraging next-generation sequencing to rapidly and cost-effectively generate a large number of mutant alleles. Using Clone-seq, we further develop a comparative interactome-scanning pipeline integrating high-throughput GFP, yeast two-hybrid (Y2H), and mass spectrometry assays to systematically evaluate the functional impact of mutations on protein stability and interactions. We use this pipeline to show that disease mutations on protein-protein interaction interfaces are significantly more likely than those away from interfaces to disrupt corresponding interactions. We also find that mutation pairs with similar molecular phenotypes in terms of both protein stability and interactions are significantly more likely to cause the same disease than those with different molecular phenotypes, validating the *in vivo* biological relevance of our high-throughput GFP and Y2H assays and indicating that both assays can be used to determine candidate disease mutations in the future. The general scheme of our experimental pipeline can be readily expanded to other types of interactome-mapping methods to comprehensively evaluate the functional relevance of all DNA variants, including those in non-coding regions.

4.2 INTRODUCTION

Owing to rapid advances in next-generation sequencing technologies, tens of thousands of disease-associated mutations(Stenson et al., 2009) and millions of single nucleotide polymorphisms (SNPs)(Consortium, 2012; Fu et al., 2013) have been identified in the human population. With the large number of ongoing whole-exome and whole-genome sequencing projects(Consortium, 2012; Fu et al., 2013), hundreds of thousands of new SNPs are now being discovered every month. Hence, there is an urgent need to develop high-throughput methods to sift through this deluge of sequence data and rapidly determine the functional relevance of each variant. Here, we focus on coding variants, firstly because trait- and disease-associated SNPs are significantly over-represented in nonsynonymous sites(Hindorff et al., 2009), and secondly because the vast majority of disease-associated mutations identified to date reside within coding regions(Stenson et al., 2009). We evaluate the functional impact of coding variants by examining their effects on corresponding protein-protein interactions, because most proteins carry out their functions by interacting with other proteins(Vidal et al., 2011).

Recent studies have begun to use large-scale protein interaction networks to understand human diseases and their associated mutations(Vidal et al., 2011; Zhong et al., 2009). By integrating structural details with high-quality protein networks, we created a 3D interactome network where the interface for each interaction has been structurally resolved(Wang et al., 2012). Using this 3D network, we demonstrated that in-frame disease mutations (missense mutations and in-frame insertions/deletions) are significantly enriched at the interaction interfaces of the corresponding proteins(Wang et al., 2012). Our results indicate that alteration of specific interactions is very important for the pathogenesis of many disease genes, highlighting

the importance of 3D structural models of protein interactions in understanding the functional relevance of coding variants. However, many important questions still remain unanswered – for example, what fraction of protein-protein interactions is altered by disease mutations to cause the corresponding disorders? Furthermore, do structural details of the interacting proteins, especially the position of the mutation relative to the interaction interface, affect the ability of a given disease mutation to alter a specific interaction?

To address these questions, we decided to focus on proteins with known disease mutations that participate in interactions with available co-crystal structures in the Protein Data Bank (PDB)(Berman et al., 2000). To detect the alteration of the interactions by disease mutations, it is necessary to first detect the interactions of the wild-type proteins using an assay of choice. This turns out to be a major bottleneck because all high-throughput interaction-detection assays have very limited sensitivity(Braun et al., 2009; Yu et al., 2008). Our assay of choice is Y2H because there are over 16,000 human protein interactions detected by our version of Y2H that can serve as the reference interactome for comparison(HI2012, 2012; Rual et al., 2005; Venkatesan et al., 2009; Yu et al., 2011), the largest for any assay performed to date. In total, there are 217 interactions detected by our version of Y2H with available co-crystal structures; 51 of these also have known missense disease mutations on corresponding proteins in the Human Gene Mutation Database (HGMD)(Stenson et al., 2009) and the corresponding interactions for the wild-type proteins can be detected in our experiments with strong Y2H-positive phenotypes (see Materials and Methods). Here, we focused on missense mutations because they are intrinsically more likely to generate interaction-specific disruptions(Zhong et al., 2009). We established a high-throughput comparative interactome-scanning pipeline to clone disease mutations and examine

their molecular phenotypes (Figure 4.1). The methodologies established here can be readily applied to any non-synonymous variant in the coding region, including nonsense mutations.

4.3 RESULTS

Clone-seq: a massively parallel site-directed mutagenesis pipeline using next-generation sequencing

The first step of our pipeline is a massively parallel approach, termed Clone-seq, designed to leverage the power of next-generation sequencing to generate a large number of mutant alleles using site-directed mutagenesis in a rapid and cost-effective manner. Current protocols for site-directed mutagenesis require picking individual colonies and sequencing each colony using Sanger sequencing to identify the correct clone(Suzuki et al., 2005). This standard approach is both labor-intensive and expensive; therefore, it does not scale up to genome-wide surveys. In Clone-seq, we put one colony of each mutagenesis attempt into one pool (Figure 4.1A; in other words, each pool contains one and only one colony for each desired mutation) and combine multiple pools through multiplexing for one Illumina sequencing run(Salehi-Ashtiani et al., 2008). Colonies for generating different mutations of the same gene can be put into the same pool, which can be easily distinguished computationally when processing the sequencing results. This is true even for mutations occurring at the same site (Figure 4.2A).

For the 51 selected interactions, we chose 27 disease-associated mutations of residues at the interface ("interface residue"), 100 mutations in the rest of the interface domain ("interface domain") and 77 mutations away from the interface ("away from the interface"; Figure 4.3A,B).

These interfaces were determined using solvent accessible surface area calculations as previously described(Das et al., 2014b; Khurana et al., 2013) on 7,340 co-crystal structures (Materials and Methods). To set up our Clone-seq pipeline, we first started with 39 mutations from these 204 and picked 4 colonies for each mutation. As a reference, we also pooled together all the wild-type alleles in our human ORFeome library to be sequenced together with the 4 pools of the mutagenesis colonies. In total, there were 40.1 million Illumina HiSeq 1×100 bp reads for our Clone-seq samples for an average of > 2,500× coverage on all desired mutation sites. Therefore, our Clone-seq pipeline has the capacity to generate > 3,000 mutations in one full lane of a HiSeq run with 1×100 bp reads, drastically improving the throughput and decreasing overall sequencing costs by at least 10-fold.

Fig. 4.2A presents a schematic of the criteria we use to determine which clones contain the desired mutation and can be used for subsequent steps. For example, in pool 1, all reads (ignoring sequencing errors) confirm that genes I and II each contain the desired mutation – T116A and G298T, respectively. For gene III, we want to generate two separate clones with two separate mutations – III_{A41T} and III_{C194T}. Since half the reads contain T41 (instead of A41) and the other half contain T194 (instead of C194), and we normalize DNA concentrations across all samples, we can infer that both mutant clones were generated successfully. In contrast, for gene IV, we see that while half the reads contain A511 (instead of G511), all the reads are wild-type at C74. Thus, we infer that while the IV_{G511A} clone is successfully generated, the IV_{C74T} clone is not. For gene V, although both mutant clones are successfully generated, half the reads contain an additional mutation, C436G. Since it is impossible to know which of the two clones for V contains this unwanted mutation, neither clone is usable. Similarly, we can determine mutant clones I_{T116A}, III_{A41T}, III_{C194T}, IV_{C74T}, IV_{G511A}, V_{T53G}, and V_{G272A} as usable clones in pool *n*.

Based on these criteria, we developed the *S* score calculation and used it to determine successful mutagenesis attempts (Materials and Methods**).** Out of 156 colonies for 39 mutations, 125 of them contain the desired mutations ($S > 0.8$), an overall 80% PCR-mutagenesis success rate. In fact, we were able to pick correct clones for all 39 mutant alleles using only the first two pools in Clone-seq. All 78 clones from the first two pools, from which the correct ones were selected for use in subsequent steps, were also Sanger sequenced for verification. 55 Clone-seq positive results with $S > 0.8$ were all confirmed and there is a clear separation in the *S* scores between the successful and failed mutagenesis attempts (Figure 4.2B).

One major advantage of our Clone-seq pipeline is that it allows us to carefully examine whether other unwanted mutations have been inadvertently introduced during PCR-mutagenesis in comparison with the corresponding wild-type alleles, since we obtain reads spanning the entire gene. We found that there are on average 4-5 unwanted mutations introduced in each pool of 39 colonies. This corresponds to a 0.013% PCR error rate (Materials and Methods), in agreement with previous studies(Vandenbroucke et al., 2011). The detection of unwanted mutations, especially those distant from the mutation of interest, is achieved in traditional site-directed mutagenesis pipelines by Sanger sequencing through the gene of interest. This is costly and labor-intensive, especially because multiple sequencing runs are needed for one long gene. However, since Clone-seq yields reads spanning the entire gene, we were able to determine which of the generated clones definitely do not have unwanted mutations in the full length of their sequences as illustrated in Figure 4.2A (Materials and Methods), and we pick only these clones for subsequent assays.

To further test our Clone-seq pipeline, we applied it to generate clones for 113 SNPs on 66 genes from the recently published Exome Sequencing Project dataset(Fu et al., 2013). Using the

same approach as described above, we sequenced 4 colonies each for the 113 alleles of interest using one third of a 1×100 bp MiSeq run. We obtained 4.7 million reads for these 113 alleles. With a threshold of $S > 0.8$, we were able to determine that 370 out of the 452 colonies (82%) contain the desired mutation, in perfect agreement with the PCR-mutagenesis success rate obtained earlier. We were able to choose colonies that contain only the desired mutation for all 113 alleles. Because the whole MiSeq run produced 17.7 million reads and we only used 4.7 million for generating the 113 mutant clones, the capacity of our Clone-seq pipeline using one full lane of a 1×100 bp HiSeq run is estimated to be >3,000, exactly the same as our previous assessment.

Finally, we generated the remaining 165 disease mutations (of the 204) and 717 other coding variants from the Exome Sequencing Project and the Catalog of Somatic Mutations in Cancer(Forbes et al., 2011) using a full 1×100 bp HiSeq run, including 40 mutations on a single gene – *MLH1*. Using 111.2 million reads for these 882 alleles, we found that 2,958 of the 3,528 colonies (84%) contain the desired mutation, again in excellent agreement with our previously obtained PCR-mutagenesis success rate. There was at least one colony with only the desired mutation for all 882 alleles, including all 40 MLH1 mutations. Therefore, our Clone-seq pipeline can generate a large number of mutations (>40) even for a single gene. In fact, to generate even more mutations for one gene, we can implement a two-round barcoding approach: generate groups of 40 mutations and barcode them differently for one HiSeq run. Ten such groups will enable us to generate ~400 mutations for a single gene. Since the average coverage of these 882 alleles is > 300×, the capacity of our Clone-seq pipeline using one full lane of a 1×100 bp HiSeq run is estimated to be >3,000, again in agreement with our previous two estimates.

Overall, our pipeline has been significantly optimized to make it very efficient. We established a web tool (http://www.yulab.org/Supp/MutPrimer) to design mutagenesis primers both individually and in batch. MutPrimer can design ~1,000 primers for ~500 mutations in one batch in less than one second. All of the 2,068 primers for the 1,034 mutations in this study were generated by MutPrimer. All mutagenesis PCRs are performed in batch using automatic 96-well procedures. Since single colony picking after bacterial transformation of mutagenesis PCR product is a rate-limiting step, we rigorously optimized this step and found that adding 10 μL mutagenesis PCR products to 100 μL competent cells and plating 50 μL transformed cells give the best transformation yield and well-separated single colonies. Furthermore, rather than individually streaking transformed cells onto agar plates one sample at a time, we were able to significantly increase throughput by spreading colonies using glass beads onto four sector agar plates which are partitioned into four non-contacting quadrants (Materials and Methods). In this manner, a 96-well plate of transformed bacteria can be plated out onto 24 four-sector agar plates in ~15 minutes. Traditional site-directed mutagenesis pipelines require miniprepping each of the selected colonies and sequencing them separately by Sanger sequencing. To drastically improve the throughput of our Clone-seq pipeline, we pooled together the bacteria stock of a single colony for each mutagenesis attempt to perform one single maxiprep, which makes the library construction step much more efficient and amenable to high-throughput. Furthermore, existing variant calling pipelines(McKenna et al., 2010) cannot be applied to our Clone-seq results because the expected allelic ratios built into these pipelines are a function of the ploidy of the organism. However, in our Clone-seq pipeline there is no concept of ploidy. We pool together many mutations for one gene in the same pool (e.g., 40 mutations for *MLH1*) and different genes

often have different numbers of mutations. Our *S* score calculation and unwanted mutation detection pipeline was designed according to our pooling strategy (Materials and Methods).

In total, we have used the novel Clone-seq pipeline successfully to generate 1,034 (39 + 113 + 882) mutant clones without any additional unwanted mutations, confirming the scalability, accuracy, and throughput of our Clone-seq pipeline.

A high-throughput GFP assay to determine the impact of mutations on protein stability

For the 204 mutations on proteins with co-crystal structures, we first examined whether the mutant proteins can be stably expressed in human cells. To do this, we tagged every wild-type and mutant protein with GFP at the C-terminus using high-throughput Gateway cloning (Figure 4.1B). The GFP constructs were transfected into HEK293T cells and fluorescence intensities were measured by a plate reader (Figure 4.3C; Materials and Methods). All fluorescence intensity readings were also confirmed manually under a microscope. Compared with the corresponding wild-type proteins, the expression levels of 3 of the 27 "interface residue" mutants, 8 of the 99 "interface domain" mutants and 6 of the 77 "away from the interface" mutants are significantly diminished (Figure 4.3C; Materials and Methods; Table 4.1). To validate these findings, we also performed Western blotting for 8 random mutants that are stably expressed and 8 random mutants with significantly diminished expression levels (Figure 4.4A). Western blotting results confirm our GFP intensity readings.

A high-throughput Y2H assay to determine the impact of mutations on protein interactions

Next, we investigated whether these mutations could affect protein-protein interactions using Y2H (Figure 4.1C; Materials and Methods). We found that 21 of the 27 (78%) "interface residue" mutations, 57 of the 100 (57%) "interface domain" mutations, and only 22 of the 77 (29%) "away from the interface" mutations disrupt the corresponding interactions, thereby demonstrating a clear difference (Figure 4.4B; $P = 3 \times 10^{-6}$ between "interface residue" and "interface domain" and $P = 8 \times 10^{-10}$ between "interface domain" and "away from the interface") in terms of ability to interfere with protein-protein interactions between mutations at different structural loci within the same protein. Furthermore, comparing with the GFP results, we found that all destabilizing mutations were shown to disrupt the corresponding interactions in our Y2H experiments. By considering only the mutations that do not affect protein expression based on the GFP experiments, we found the same difference: 13 out of 18 (72%) "interface residue" stable mutations, 42 out of 83 (51%) "interface domain" stable mutations, and only 9 out of 52 (17%) "away from the interface" stable mutations disrupt the corresponding interactions (Figure 4.4B; $P = 2 \times 10^{-5}$ between "interface residue" and "interface domain" and $P = 9 \times 10^{-13}$ between "interface domain" and "away from the interface"; Table 4.1). Since these interfaces are obtained from actual co-crystal structures, our results suggest that accurate structural information can help determine the functional impact of mutations on protein-protein interactions. Wild-type proteins corresponding to 113 of the 153 stably expressed mutant proteins also interact with other proteins as determined by our Y2H experiments (114 interactions in total, termed "other interactions"); however, for these interactions, there are currently no co-crystal structures available in the PDB. Using these other interactions, we calculated the likelihood of a given mutation disrupting a specific interaction without any structural information to be 32% (Figure 4.4B).

Relationships between measured molecular phenotypes and corresponding disease phenotypes

We then analyzed whether the molecular phenotypes measured by our high-throughput GFP and Y2H assays are correlated with corresponding disease phenotypes. We first examined how mutation pairs on the same gene affect protein stability and its relationship to their corresponding diseases. We find that pairs of mutations that are either both stable or both unstable cause the same disease in 68% and 70% of cases, respectively. However, pairs comprising one stable and one unstable mutation cause the same disease in only 30% of cases ($P = 6 \times 10^{-9}$ and 8×10^{-10}, respectively, Figure 4.5A). For example, we find that the mutations R727C and L844F on the spindle checkpoint kinase Bub1b both cause the protein to become unstable and lose all its interactors. These mutations are both associated with the same disease, mosaic variegated aneuploidy, an autosomal recessive disorder that causes predominantly trisomies and monosomies of different chromosomes(Hanks et al., 2004; Suijkerbuijk et al., 2010). Since our GFP assay shows that these two mutations cause loss of protein product, our results are consistent with Matusuura et al.'s finding that a more than 50% decrease in Bub1b activity leads to abnormal mitotic spindle checkpoint function and mosaic variegated aneuploidy(Matsuura et al., 2006).

We then examined whether mutation pairs on the same gene disrupt the same set or different sets of interactions (i.e., their interaction disruption profiles) and investigated whether their disruption profiles correlates with disease phenotypes. We found that mutation pairs with the exact same disruption profile are significantly more likely to cause the same disease than those

with different profiles (70% and 61% respectively, $P = 3 \times 10^{-5}$, Figure 4.5B). For example, we found that two mutations on Smad4, R361C and Y353S, disrupt its interactions with Smad3 and Smad9 while leaving the interactions with Lmo4 and Rassf5 unaltered (Figure 4.5C). These two mutations both cause juvenile polyposis coli(Houlston et al., 1998; Roth et al., 1999), a disease is known to be caused by disruption of the core Smad/Bmp signaling pathways(Massague, 2008). Our Y2H results clearly demonstrate that the R361C and Y353S mutations disrupt the Smad4-Smad3 and Smad4-Smad9 interactions (Figure 4.5C) leading to disruption of core Smad signaling pathways. However, the mutation N13S on Smad4 does not disrupt any of these interactions (Figure 4.5C) and is associated with a different disease, pulmonary arterial hypertension. Our results agree with Nasim et al.'s finding that the N13S mutation does not alter downstream Smad signaling(Nasim et al., 2011). Our findings provide support for the hypothesis that the N13S mutation either impacts pathways outside the core Smad signaling network or are pathogenic only when combined with other environmental and genetic factors(Machado, 2012).

Overall, these results show that mutation pairs with similar molecular phenotypes in terms of both protein stability and interactions are significantly more likely to cause the same disease than those with different molecular phenotypes. This confirms that the molecular phenotypes measured by our high-throughput GFP and Y2H assays are biologically relevant *in vivo*. Furthermore, by comparing the molecular phenotypes, in particular the protein interaction disruption profiles, of mutations/variants to those of known disease mutations, potential candidate mutations for a variety of diseases can be identified.

A high-throughput mass spectrometry assay to determine the impact of mutations on protein interactions

While we use only those interactions that are supported by co-crystal structures to estimate the fraction of interactions that are disrupted by mutations at different structural loci, the described procedures can also be applied to interactions with predicted interfaces and structural models(Meyer et al., 2013; Mosca et al., 2013; Tuncbag et al., 2011; Zhang et al., 2012). This is of particular importance because over 90% of known interactions do not currently have corresponding co-crystal structures(Das et al., 2014a; Mosca et al., 2013). For example, Mlh1 is known to interact with Pms2, both of which are well-studied DNA mismatch repair genes frequently mutated in hereditary nonpolyposis colorectal cancer(Peltomaki and Vasen, 1997). Although the structural basis of the Mlh1-Pms2 interaction still remains unknown, both our previous 3D reconstruction of the human interactome network(Meyer et al., 2013; Wang et al., 2012) and the newly-established Interactome3D(Mosca et al., 2013) database suggest that the HATPase_c domain is part of the interface for Mlh1's interaction with Pms2. Previous work has shown that a point mutation (I107R) on the HATPase_c domain of Mlh1 is associated with colorectal cancer and disrupts the Mlh1-Pms2 interaction(Kondo et al., 2003; Peltomaki and Vasen, 1997; Wang et al., 2012). First, using Y2H, we were able to confirm the disruption. Next, we developed a high-throughput-amenable mass spectrometry pipeline using Stable Isotope Labeling by Amino acids in Cell culture (SILAC)(Ong et al., 2002; Ong and Mann, 2006), which was designed to reveal both lost/weakened and gained/enhanced interactions of the target proteins (Figure 4.1D)(Ohouo et al., 2010). We added an HA-tag to the N-terminus of both wild-type and mutant Mlh1, as well as to GFP as a control, and performed four SILAC experiments: wild-type Mlh1 (heavy) vs. GFP control (light), mutant Mlh1 (heavy) vs. GFP control (light), wild-type (heavy) vs. mutant (light) Mlh1, and mutant (heavy) vs. wild-type (light) Mlh1 (Figure

4.6A; Materials and Methods). Interactors of wild-type/mutant Mlh1 are defined as those that bind wild-type/mutant Mlh1 more than 2× stronger than GFP control (Materials and Methods). For a lost/weakened interaction, we required that the interaction be more than 2× stronger with wild-type Mlh1 than with mutant Mlh1 as confirmed both in wild-type (heavy) vs. mutant (light) and in mutant (heavy) vs. wild-type (light) experiments; we further required that the interaction be detected in the wild-type vs. control experiment (Figure 4.6A; Materials and Methods). For a gained/enhanced interaction, we required that the interaction be more than 2× stronger with mutant Mlh1 than with wild-type Mlh1 as confirmed both in wild-type (heavy) vs. mutant (light) and in mutant (heavy) vs. wild-type (light) experiments; we further required that the interaction be detected in the mutant vs. control experiment (Figure 4.6A; Materials and Methods). We were able to detect Pms2 as the only specifically weakened interactor caused by the mutation (Figures 4.6B,C; E = -1.77; P = 3 x 10^{-4}), in agreement with our Y2H results and previous studies(Kondo et al., 2003; Wang et al., 2012). Additionally, we were able to detect Hspa8 as the only specifically enhanced interactor of the mutant protein (Figures. 4.6B,C; E = 2.71; P = 7 x 10^{-8}). Two other known interactors of Mlh1, Pms1 (Figures 4.6B,C; E = -0.32; P = 0.21)(Leung et al., 2000) and Brip1 (Figures 4.6B,C; E = 0.18; P = 0.32)(Peng et al., 2007), were also detected, although their interactions with Mlh1 are not affected by this particular mutation (Materials and Methods).

Hspa8 was not previously known to interact with Mlh1 and the impact of the Mlh1 I107R mutation on its interactions with Pms1 and Brip1 has not been reported in the literature. To verify our SILAC results, we performed *in vivo* co-immunoprecipitation using HA-tagged wild-type and mutant Mlh1 and tagged Hspa8 and Brip1 with V5 (Materials and Methods). Our co-immunoprecipitation results confirm that Hspa8 only weakly interacts with wild-type Mlh1, but

the interaction is dramatically enhanced by a single amino acid substitution (I107R) (Figure 4.6D, lanes 3 and 4), whereas the interaction between Mlh1 and Brip1 is not affected by this mutation (Figure 4.6D, lanes 6 and 7; Materials and Methods). Hspa8 is a constitutively expressed member of the heat shock protein 70 family(Goldfarb et al., 2006). It functions as a chaperone to facilitate protein folding(Goldfarb et al., 2006) and also functions as an ATPase in the disassembly of clathrin-coated vesicles during membrane trafficking(DeLuca-Flaherty et al., 1990). A recent study reported that Hspa8 is specifically recruited to reovirus viral factories, independent of its chaperone function(Kaufer et al., 2012). Therefore, our SILAC results suggest that Hspa8 may play an important role in colorectal cancer and that its function could be independent of its role as a chaperone.

4.4 DISCUSSION

We have successfully developed the first massively parallel site-directed mutagenesis pipeline, Clone-seq, using next-generation sequencing. Our Clone-seq pipeline is entirely different from previously described random mutagenesis approaches(Araya et al., 2012; Fowler et al., 2010; Pitt and Ferre-D'Amare, 2010; Starita et al., 2013). Clone-seq is used to generate a large number of specific mutant clones with desired mutations; each individual mutant clone has a separate stock and different clones can therefore be used separately for completely different downstream assays. In random mutagenesis, a pool of sequences containing different mutations for one gene is generated using error-prone PCR or error-prone DNA synthesis. Therefore, it is not possible to separate one mutant sequence from another and the whole pool can only be used for the same assay(s) together. Furthermore, it is not possible to control which or how many mutations are

generated on each DNA sequence. In fact, to improve coverage, most random mutagenesis pipelines generate on average two or more mutations on each DNA sequence(Fowler et al., 2010), which makes it impossible to distinguish the functional impact of each individual mutation on the same sequence. Site-directed mutagenesis and random mutagenesis are designed for different goals: if one wants to generate all possible mutations for a certain protein without the need to separate different clones, it would be more favorable to use random mutagenesis; whereas if one needs to have separate clones for each mutation, site-directed mutagenesis is required. As a result, the two approaches are complementary and not comparable.

While there are highly efficient methods for random mutagenesis(Araya et al., 2012; Fowler et al., 2010; Pitt and Ferre-D'Amare, 2010; Starita et al., 2013), current protocols for site-directed mutagenesis are low-throughput and become prohibitively expensive if a large number of clones needs to be generated. Clone-seq directly addresses the necessity for a high-throughput site-directed mutagenesis pipeline. It is a robust, cost-effective and efficient method that can be used to generate a total of ~3,000 distinct mutant clones in one full lane of a 1×100 bp HiSeq run. Clone-seq is suitable both for generating mutations across many genes as well as a large number of mutations on a few genes. The former situation is applicable when one wants to generate many mutations/variants from large-scale studies (e.g., whole-genome or whole-exome sequencing) since they typically identify mutations/variants on a large number of genes(Atlas, 2012; Stransky et al., 2011). The latter situation usually arises in a study focused on a single pathway with a few genes of interest (e.g., an alanine-scanning mutagenesis to determine functional sites on a gene of interest(Cunningham and Wells, 1989)).

Integrating with Clone-seq, we also established a comprehensive comparative interactome-scanning pipeline, including high-throughput GFP, Y2H, and mass spectrometry assays, to

systematically evaluate the impact of human disease mutations on protein stability and interactions. We examine each mutation individually, rather than looking at their combinatorial effects because these inherited germline disease mutations are extremely rare. Therefore, the probability of having even two of these in the same individual becomes infinitesimally small. Our results reveal that the overall likelihood of a given disease mutation disrupting a specific interaction is 32%. Accurate structural information of these interactions obtained from co-crystal structures greatly improves our understanding of the impact of disease mutations: 13 out of 18 (72%) "interface residue" stable mutations, 42 out of 83 (51%) "interface domain" stable mutations, and only 9 out of 52 (17%) "away from the interface" stable mutations disrupt the corresponding interactions, unveiling a clear dependence of the molecular phenotypes of disease mutations on their structural loci. These estimates are not affected by the false negative rate of our Y2H assay as we only use those interactions for which we can detect the wild-type interaction with strong Y2H phenotypes. Thus, any observed disruption is due to the mutation of interest and not an assay false negative. Furthermore, our Y2H pipeline has been shown to be of high quality and has an experimentally measured false positive rate of ~5% or lower in different organisms(Consortium, 2011; Das et al., 2013; Venkatesan et al., 2009; Yu et al., 2008). In addition, the interactions used to understand the relationship between molecular phenotypes and structural loci of disease mutations are all supported by co-crystal structures, therefore these interactions are not assay false positives. We also find that the molecular phenotypes detected by our GFP and Y2H assays correlate with known disease phenotypes, confirming the *in vivo* biological significance of our measurements.

Moreover, as shown by the Mlh1 example (**Fig. 6**), our comparative interactome-scanning pipeline can also be used with predicted structural models(Meyer et al., 2013; Mosca et al., 2013;

Tuncbag et al., 2011; Zhang et al., 2012). The consequent experimental results will clearly be affected by the quality of these predictions, which is not part of our pipeline. In fact, our experimental interactome-scanning pipeline can be applied to evaluate or improve these predicted models by testing mutations at different loci of a protein of interest and examining how these mutations disrupt different interactions of this protein.

Our comparative interactome-scanning pipeline described and validated here can be applied to experimentally determine in a high-throughput fashion the impact on protein stability and protein-protein interactions for thousands of DNA coding variants and disease mutations, which can directly lead to hypotheses of concrete molecular mechanisms for follow-up studies. Furthermore, the elucidation of molecular phenotypes of disease mutations is also vital for selecting actionable drug targets and ultimately for making therapeutic decisions. Finally, the general scheme of our pipeline can be readily expanded to other interactome-mapping methods, particularly other protein-protein(Braun et al., 2009), protein-DNA(Berger et al., 2006; Reece-Hoyes et al., 2011), protein-RNA(Yakhnin et al., 2012), and protein-metabolite interaction assays(Bandyopadhyay et al., 2012), to comprehensively evaluate the functional relevance of all DNA variants, including those in non-coding regions.

4.5 MATERIALS AND METHODS

Selecting interactions with mutations on and away from the interface

To calculate atomic-resolution interaction interfaces, we systematically examined a comprehensive list of 7,340 PDB co-crystal structures. To define the interface, we used a water molecule of diameter 1.4 Å as a probe and calculated the relative solvent accessible surface areas of the interacting pair as well as the individual proteins involved in the interaction. Residues whose relative accessibilities change by more than 1 $Å^2$ are considered as potential interface residues, because amino acids at the interface reside on the surfaces of the corresponding proteins, but will tend to become buried in the co-crystal structure as the two proteins bind to each other(Franzosa and Xia, 2011). So, for these residues, there should be a significant decrease in accessible surface area when we compare the bound and unbound states of the protein chains.

To identify interface domains, we required at least one of the following criteria to hold:

1. 3did(Stein et al., 2011) or iPfam(Finn et al., 2005) have identified the domain pair as interacting and each of the interface domains contains at least one interface residue based on our calculations.
2. The domain pair contains 5 or more interface residues for each protein according to our calculations.

We then identified the subset of these interactions that contain at least one disease mutation and are amenable to our version of Y2H(HI2012, 2012; Rual et al., 2005; Venkatesan et al., 2009;

Yu et al., 2011). Subsequently, we performed a pairwise retest of all these interactions and selected the ones that yield strong Y2H phenotypes, because subsequent steps involve detecting a significant decrease in these phenotypes.

Primer design for site-directed mutagenesis

Primers for site-directed mutagenesis were selected based on a customized version of the protocol accompanying the Stratagene QuikChange Site-Directed Mutagenesis Kit (200518). The following criteria are used:

1. The primer should be of length 30-50 bp and should contain the mutation of interest in the center or one base away.
2. The GC content of the primer should be ≥ 40% and the primer should start and end with a G or a C.
3. The T_m for the primer should be ≥ 78 °C. T_m was calculated using the following expression:

$$T_m = 81.5 + 0.41 \times (\%GC) - \frac{675}{N} - \%mismatch$$

where *N* is the primer length in bases, *%GC* is the percentage of G or C nucleotides in the primer, and *%mismatch* is the percentage of mismatched bases in the primer. Values for *%GC* and *%mismatch* are whole numbers.

For cases where no primer satisfies all three criteria simultaneously, we relaxed criterion 2 to GC content ≥ 30%.

We established a supplementary web tool (http://www.yulab.org/Supp/MutPrimer) to design mutagenesis primers individually or in bulk.

Construction of mutant alleles using high-throughput site-directed mutagenesis PCR

All wild-type clones were obtained from the human ORFeome v8.1 collection(Yang et al., 2011). To generate mutant alleles, sequence-verified single-colony wild-type clones and their corresponding mutagenic primers were aliquoted into individual wells of 96-well PCR plates. Mutagenesis PCR was then performed as specified by the New England Biolabs (NEB) PCR protocol for Phusion polymerase (M0530L), noting that PCR was limited to 18 cycles. The samples were then digested by *Dpn*I (NEB R0176L) according to the manufacturer's manual. After digestion, samples were transformed into competent *E. coli*. Since single colony picking after bacterial transformation of mutagenesis PCR product is a rate-limiting step, we rigorously optimized this step. First, we tried different volumes of competent cells for transformation and found that single colony yields peak when ~100 µL of competent cells are used. It is also necessary to use ~10 µL of mutagenesis PCR product: any lower volume of PCR product results in significantly reduced colony yields, while higher volumes of PCR product do not increase yield. Finally, colony picking was done using four-sector agar plates (VWR 25384-308) that are partitioned into four non-contacting quadrants with glass beads poured onto each plate quadrant. Each bead-filled quadrant was inoculated with ~50 µL of transformed bacteria. This was then

spread by lightly shaking the four-sector agar plate. Our optimized transformation protocol results in a large number of well-separated single colonies that can be easily picked the next day. Upon recovery, single colonies from each quadrant were then picked and arrayed into 96-deepwell plates filled with 300 μL of antibiotic media. Four colonies per allele were picked for next-generation sequencing.

DNA library preparation for Illumina sequencing

DNA library preparation was performed using NEBNext DNA Library Prep Master Mix Set for Illumina (NEB E6040S) according to the manufacturer's manual. Briefly, 5 μg of pooled plasmid DNA (~100 μL, all samples were normalized to the same concentration) was sonicated to ~200 bp fragments. The fragmented DNA was first mixed with NEBNext End Repair Enzyme for 30 mins at 20 °C. Blunt-ended DNA was then incubated with Klenow Fragment for 30 mins at 37 °C for dA-Tailing. Subsequently, NEBNext Adaptor was added to dA-Tailed DNA. Adaptor-ligated DNA (~300 bp) was size-selected on a 2% agarose gel. Size-selected DNA was then mixed with one of the NEBNext Multiplex Oligos (NEB E7335S) and Universal PCR primers for PCR enrichment. At each step, DNA was purified using a QIAquick PCR purification kit (Qiagen 28104). Multiplexed DNA samples were combined and analyzed in one lane of a 1×100 bp run by Illumina HiSeq 2500.

Identifying successful instances of site-directed mutagenesis based on next-generation sequencing

The mutant colonies were barcoded and pooled as shown in Fig. 1a. The multiplexed colonies were then run on an Illumina sequencer (2 HiSeq runs and 1 MiSeq run) to give 1×100 bp reads. These reads were then de-multiplexed and mapped to the genes of interest using the BWA "aln" algorithm(Li and Durbin, 2009). For each allele, we identified all reads that mapped to the position of the mutation of interest (R_{all}) and those that actually contained the desired mutation (R_{mut}). We then calculated a normalized score (S) that quantifies the fraction of reads containing the desired mutation:

$$S = \frac{R_{mut}}{\frac{1}{k}R_{all}} = \frac{k \times R_{mut}}{R_{all}}$$

where k is the number of different mutations for the same gene.

For 39 mutations, we Sanger sequenced two mutant colonies per mutagenesis attempt to quantify the correlation between S and observation of the desired mutation. We found that all clones with $S > 0.44$ are confirmed to be correct via Sanger sequencing with a clear separation between those that are correct and those that are not (Figure 4.2b). However, to further ensure that the clones we picked were correct, we require $S > 0.8$ for a colony to be scored as containing the desired mutation.

Identifying unwanted mutations

One major advantage of our Clone-seq pipeline over traditional site-directed mutagenesis protocols using Sanger sequencing(Suzuki et al., 2005) is that we can now carefully examine whether there are other unwanted mutations inadvertently introduced during the PCR process, in comparison with the corresponding wild-type alleles. It is essential to use clones with no unwanted mutations for downstream experiments, as the presence of these will make it impossible to determine whether the observed disruption is due to the desired or other undesirable mutation(s).

We use samtools "mpileup"(Li et al., 2009) to obtain read counts for different alleles at each nucleotide for all the clones. We calculate the background sequencing error rate by calculating the average fraction of non-reference alleles across all nucleotides where we did not attempt to introduce a mutation. Any site that has a significantly higher fraction of non-reference alleles (using a *P* value cutoff of 0.2 from a cumulative binomial test) is considered to have an unwanted mutation. A lenient *P* value cutoff (0.2 as opposed to the more traditionally used 0.05 or 0.01) implies more stringent filtering in this case because we want to eliminate type II errors i.e., we want to identify all unwanted mutations at the cost of discarding a few clones that actually do not have any unwanted mutations.

We identified an average of 4-5 unwanted point mutations per pool. The overall per-base point mutation rate of Phusion polymerase was calculated to be ~ 10^{-4}. NEB's advertised error rate for Phusion polymerase varies from 4.4 – 9.5 x 10^{-7} per PCR cycle. Since we perform 18 PCR cycles, the expected overall error rate is ~ 10^{-5}. Our calculated mutation is within an order of magnitude of this advertised error rate. It is slightly higher than the advertised rate as we use stringent filtering criteria as described above.

GFP assay

All wild-type and mutant clones were moved into the pcDNA-DEST47 vector with a C-terminal GFP tag using automated Gateway LR reactions in a 96-well format. After bacterial transformation, minipreps were prepared on a Tecan Freedom Evo 200, and DNA concentrations were determined by OD 260/280 with a Tecan Infinite M1000 plate reader in 96-well format. A 100 ng aliquot of each expression clone plasmid was used for transfection into HEK293T cells in 96-well plates using Lipofectamine 2000 (Invitrogen 11668019) according to the manufacturer's instructions. At approximately 48 hrs post-transfection, cells were processed with Tecan M1000. Fluorescence intensities were measured at 395 nm for excitation and 507 nm for emission, according to Invitrogen's manual. As negative controls, the fluorescence intensities corresponding to cells transfected with the empty vector were measured. The normalized fluorescence intensity was calculated as:

$$I_{norm} = I - I_{background}$$

where I corresponds to the measured intensity and $I_{background}$ corresponds to the average intensity of the empty vector controls for each plate. All I_{norm} values greater than K are considered to correspond to stable protein expression. K corresponds to the range (maximum – minimum) of background fluorescence intensities of the empty vector controls for each plate. For this study, all fluorescence intensity readings were also confirmed manually under a microscope. All transfection and GFP experiments were repeated 3 times.

Y2H assay

Y2H was performed as previously described(Wang et al., 2012). All wild-type/mutant clones were transferred by Gateway LR reactions into our Y2H pDEST-AD and pDEST-DB vectors. All DB-X and AD-Y plasmids were transformed individually into the Y2H strains *MAT*α Y8930 and *MAT***a** Y8800, respectively. Each of the DB-X *MAT*α transformants (wild-type and mutants) were then mated against corresponding AD-Y *MAT***a** transformants (wild-type and mutants) individually using automated 96-well procedures, including inoculation of AD-Y and DB-X yeast cultures, mating on YEPD media (incubated overnight at 30 °C), and replica-plating onto selective Synthetic Complete media lacking leucine, tryptophan, and histidine, and supplemented with 1 mM of 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT), SC-Leu-His+3AT plates containing 1 mg/l cycloheximide (SC-Leu-His+3AT+CHX), SC-Leu-Trp-Adenine (Ade) plates, and SC-Leu-Ade+CHX plates to test for CHX-sensitive expression of the *LYS2::GAL1-HIS3* and *GAL2-ADE2* reporter genes. The plates containing cycloheximide select for cells that do not have the AD plasmid due to plasmid shuffling. Growth on these control plates thus identifies spontaneous auto-activators(Walhout and Vidal, 2001). The plates were incubated overnight at 30 °C and "replica-cleaned" the following day. Plates were then incubated for another three days, after which positive colonies were scored as those that grow on SC-Leu-Trp-His+3AT and/or on SC-Leu-Trp-Ade, but not on SC-Leu-His+3AT+CHX or on SC-Leu-Ade+CHX. Disruption of an interaction by a mutation was defined as at least 50% reduction of growth consistently across both reporter genes, when compared to Y2H phenotypes of the corresponding wild-type allele as benchmarked by 2-fold serial dilution experiments. All Y2H experiments were repeated 3 times.

Construction of plasmids

Wild-type *MLH1*, *HSPA8,* and *BRIP1* entry clones are from the human ORFeome v8.1 collection(Yang et al., 2011). Using Gateway LR reactions, wild-type *MLH1*, mutant *MLH1* (I107R), and GFP were transferred into the pMSCV-N-FLAG-HA-PURO vector(Behrends et al., 2010); *HSPA8* and *BRIP1* were transferred into the pcDNA-DEST40 vector that contains a C-terminal V5 tag (Invitrogen 12274-015).

Analysis of interacting proteins by SILAC and LC-MS/MS

HEK293T cells were grown in SILAC media comprising SILAC DMEM (Thermo Scientific) and 10% dialyzed FBS (JR Scientific) supplemented with either 0.1 mg/ml L-lysine and L-arginine (light media) or 0.1 mg/ml L-lysine 13C6, 15N2 and L-arginine 13C6, 15N4 (heavy media). Heavy- or light-media cultured HEK293T cells were transfected using Lipofectamine 2000 (Invitrogen) in three 10 cm plates. 48 hrs after transfection, cells were washed three times in cold PBS and then resuspended in 5 ml RIPA buffer [1% NP-40, 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM EDTA, 1× EDTA-free Complete Protease Inhibitor tablet (Roche)]. Cells were lysed for 30 mins on ice before centrifuging at 13,000 rpm for 10 mins. Cell lysates were incubated with 60 µL EZview Red Anti-HA Affinity Gel (Sigma-Aldrich) for 3 hrs. After 3 washes with RIPA buffer, bound proteins were eluted with 3 resin volumes elution buffer (100 mM Tris-HCl pH 8.0, 1% SDS). Eluted proteins from light and heavy media were mixed together, reduced with 5 mM DTT, alkylated with 15 mM of iodoacetamide, and then precipitated with 3 volumes PPT solution (50% acetone, 49.9% ethanol, 0.1% acetic acid).

Proteins from pull-down experiments were solubilized with 50 µL Urea/Tris solution (8 M Urea, 50 mM Tris-HCl pH 8.0) and 150 µL NaCl/Tris (50 mM Tris-HCl pH 8.0, 150 mM NaCl) followed by the addition of 1 µg Trypsin Gold (Promega). Protein digestion was performed overnight at 37 °C after which trifluoroacetic acid and formic acid were added to a final concentration of 0.2%. Peptides were de-salted with Sep-Pak C18 columns (Waters Corporation), dried in a speed-vac, and reconstituted in 85 µL of a solution containing 80% acetonitrile and 1% formic acid. Samples were fractionated by Hydrophilic Interaction LIquid Chromatography (HILIC) using a TSK gel Amide-80 column (Tosoh Bioscience). HILIC fractions were dried in a speed-vac, reconstituted in 0.1% trifluoroacetic acid, and analyzed by LC-MS/MS using a 125 µM ID capillary column packed in-house with 3 µm C18 particles (Michrom Bioresources) and a Q-Exactive mass spectrometer (Thermo Fisher Scientific) coupled with a Nano LC-Ultra system (Eksigent). Xcalibur 2.2 software (Thermo Fischer Scientific) was used for the data acquisition and Q-Exactive was operated in the data-dependent mode. Survey scans were acquired in the Orbitrap mass analyzer over the range of 380 to 2000 m/z with a mass resolution of 70.000 (at m/z 200). Up to the top 10 most abundant ions with a charge state higher than 1 and less than 5 were selected within an isolation window of 2.0 m/z. Selected ions were fragmented by Higher-energy Collisional Dissociation (HCD) and the tandem mass spectra were acquired in the Orbitrap mass analyzer with a mass resolution of 17.500 (at m/z 200). The fragmentation spectra were searched by using the SEQUEST software on a SORCERER system (Sage-N Research) and a human database downloaded from the International Protein Index (version 3.80). In all database searches, trypsin was designated as the protease, allowing for one non-tryptic end and two missed-cleavages. The following parameters were used in the database search: a mass accuracy of 15 ppm for the precursor ions, differential

modification of 8.0142 Daltons for lysine and 10.00827 Daltons for arginine. Results were filtered based on probability score to achieve a 1% false positive rate. The Xpress software, part of the Trans-Proteomic Pipeline (Seattle Proteome Center), was used to process the raw data and quantify the light/heavy peptide isotope ratios. Results were also manually inspected.

Identifying loss and gain of interactors for Mlh1

We performed four SILAC experiments using both wild-type and mutant Mlh1, as well as GFP as a control: wild-type (heavy) vs. control (light) [WT_Control]; mutant (heavy) vs. control (light) [Mutant_Control]; wild-type (heavy) vs. mutant (light) [WT_Mutant]; and mutant (heavy) vs. wild-type (light) [Mutant_WT].

We use the following variables and define four ratios for all subsequent calculations. In the WT_Control experiment, the relative abundance of protein p pulled down by wild-type Mlh1 to protein p pulled down by GFP (WT_p) is quantified by the inverse of the geometric mean of r_{wc} reads with Xpress values X_i. In the Mutant_Control experiment, the relative abundance of protein p pulled down by mutant Mlh1 (I107R) to protein p pulled down by GFP (Mut_p) is quantified by the inverse of the geometric mean of r_{mc} reads with Xpress values Y_i. In the WT_Mutant experiment, the relative abundance of protein p pulled down with mutant Mlh1 (I107R) to protein p pulled down by wild-type Mlh1 is quantified by the geometric mean of r_{wm} reads with Xpress values P_i. The amount of mutant Mlh1 (I107R) to wild-type Mlh1 is quantified by the geometric mean of t_{wm} reads with Xpress values C_j. In the Mutant_WT experiment, the relative abundance of protein p pulled down with mutant Mlh1 (I107R) to protein p pulled down by wild-

type Mlh1 is quantified by the inverse of the geometric mean of r_{mw} reads with Xpress values Q_j. The amount of mutant Mlh1 (I107R) to wild-type Mlh1 is quantified by the inverse of the geometric mean of t_{mw} reads with Xpress values D_i.

$$WT_p = \sqrt[r_{wc}]{\prod_{i=1}^{r_{wc}} \frac{1}{X_i}}$$

$$Mut_p = \sqrt[r_{mc}]{\prod_{i=1}^{r_{mc}} \frac{1}{Y_i}}$$

$$FC_{wm} = \frac{\sqrt[r_{wm}]{\prod_{i=1}^{r_{wm}} P_i}}{\sqrt[t_{wm}]{\prod_{j=1}^{t_{wm}} C_j}}$$

$$FC_{mw} = \frac{\sqrt[t_{mw}]{\prod_{i=1}^{t_{mw}} D_i}}{\sqrt[r_{mw}]{\prod_{j=1}^{r_{mw}} Q_j}}$$

where both FC_{wm} and FC_{mw} denote the fold change in protein abundance as the normalized ratio of the amount of protein pulled down with mutant Mlh1 to that with wild-type Mlh1.

To identify interactors that are lost/weakened due to the I107R mutation, we required the following criteria to hold simultaneously:

1. The protein has to be identified as an interactor of wild-type Mlh1: $WT_p > 2$, $r_{wc} \geq 5$.
2. The protein has to be identified as a lost interactor based on both Mutant_WT: $FC_{mw} < 0.5$, $r_{mw} \geq 5$, and WT_Mutant: $FC_{wm} < 0.5$, $r_{wm} \geq 5$.

The first criterion ensures that the protein identified is a true interactor of wild-type Mlh1. The second criterion ensures that the loss of interaction is significant and reliably observed across both WT_Mutant and Mutant_WT experiments.

Similarly, to identify interactors that are gained/enhanced due to the I107R mutation, we required the following criteria to hold simultaneously:

1. The protein has to be identified as an interactor of mutant Mlh1 (I107R): $Mut_p > 2$, $r_{mc} \geq 5$.
2. The protein has to be identified as a gained interactor based on both Mutant_WT: $FC_{mw} > 2$, $r_{mw} \geq 5$, and WT_Mutant: $FC_{wm} > 2$, $r_{wm} \geq 5$.

The first criterion ensures that the protein identified is a true interactor of the I107R mutant of Mlh1. The second criterion ensures that the gain of interaction is significant and reliably observed across both WT_Mutant and Mutant_WT experiments.

We also identify interactors of Mlh1 that are unaffected by the I107R mutation using the following criteria:

1. The protein has to be identified as an interactor of both wild-type Mlh1: $WT_p > 2$, $r_{wc} \geq 5$, and mutant Mlh1 (I107R): $Mut_p > 2$, $r_{mc} \geq 5$.
2. The protein has to be identified as an unchanged interactor based on both Mutant_WT: $0.5 < FC_{mw} < 2$, $r_{mw} \geq 5$, and WT_Mutant: $0.5 < FC_{wm} < 2$, $r_{wm} \geq 5$.

Integrating both WT_Mutant and Mutant_WT experiments, we calculated a weighted average of the individual fold changes:

$$E = \frac{r_{mw} \times \log_2(FC_{mw}) + r_{wm} \times \log_2(FC_{wm})}{r_{mw} + r_{wm}}$$

P values are calculated using a two-sided Kolmogorov-Smirnov test (with bootstrapping).

Cell culture, co-immunoprecipitation, and Western blotting

HEK293T cells were maintained in complete DMEM medium supplemented with 10% FBS. Cells were transfected with Lipofectamine 2000 (Invitrogen) at a 6:1 (μL/μg) ratio with DNA in 6-well plates and were harvested 24 hrs after transfection. Cells were gently washed three times in PBS and then resuspended using 200 μL 1% NP-40 lysis buffer [1% Nonidet P-40, 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1× EDTA-free Complete Protease Inhibitor tablet (Roche)] and kept on ice for 20 mins. Extracts were cleared by centrifugation for 10 mins at 13,000 rpm at 4

°C. 15 µL EZview Red Anti-HA Affinity Gel (Sigma-Aldrich) and 100 µL protein lysate were used for each co-immunoprecipitation reaction. The samples were rotated gently at 4 °C for 2 hrs. HA beads were then washed three times with protein lysis buffer, treated with 6× protein sample buffer, and subjected to SDS-PAGE. Proteins were then transferred from the gel onto PVDF (Amersham) membranes. Anti-HA (Sigma H9658), anti-V5 (Invitrogen 46-0705), anti-β-tubulin (Promega G7121), and anti-GFP (Santa Cruz sc-9996) antibodies were used at 1:3,000 dilutions for immunoblotting analysis.

4.6 FIGURE AND TABLE LEGENDS

Figure 4.1. Schematic of our comparative interactome-scanning pipeline.

Our pipeline begins with Clone-seq (a), a massively-parallel low-cost site-directed mutagenesis pipeline leveraging next-generation sequencing. This is followed by a high-throughput GFP assay (b) to determine protein stability, and a high-throughput Y2H assay (c), along with SILAC-based mass spectrometry (d) to determine the impact of DNA coding variants on protein interactions.

Figure 4.2. Identifying usable clones from Clone-seq.

(a) Schematic illustrating criteria used to determine which of the clones generated by our Clone-seq pipeline are usable for further assays – green ticks indicate usable clones, while red crosses indicate clones that cannot be used. (b) Variation of *S* across different mutagenesis attempts that either contain or do not contain the desired mutation as confirmed by Sanger sequencing.

Figure 4.3. Examples of disease mutations in different structural loci of protein-protein interactions and examples of our GFP assay results.

(a) Crystal structure (PDB id: 3W4U) depicting a D100Y mutation (on Hbb) at an interface residue and a F104L mutation in the interface domain for the Hbb-Hbz interaction. (b) Crystal structure (PDB id: 1G3N) depicting a V31L mutation (on Cdkn2c) away from the Cdkn2c-Cdk6 interaction interface. (c) GFP assays that determine the stability of wild-type Rrm2b and the R41P and L317V mutations on Rrm2b that are at an interface residue and away from the interface for the Rrm2b-Rrm2b interaction; GFP assays that determine the stability of wild-type

Hprt1 and the C206Y mutation on Hprt1 that is away from the interaction interface of Hprt-Hprt1. Empty vector was used as a negative control.

Figure 4.4. Effect of disease mutations on protein stability and protein-protein interactions.

(a) Western blotting with anti-GFP antibody confirming the protein expression levels of wild-type Rrm2b, Actn2, Hprt1, Pnp, Tpk1, Gnmt, Gale, Fbp1, Klhl3, Tp53, Pnp, Smad4, and corresponding mutant alleles. β-tubulin and γ-tubulin were used as loading controls. Red denotes "interface residue" mutations, orange denotes "interface domain" mutations and blue denotes "away from the interface" mutations. (b) Likelihood of disruption of interactions by "interface residue", "interface domain" and "away from the interface" mutations – overall and for stable mutants only; likelihood of a disease mutation disrupting a given interaction in the absence of structural information. Error bars indicate +SE. (N = 204 mutations)

Figure 4.5. Relationships between molecular phenotypes and disease phenotypes.

(a) Fraction of mutation pairs on the same gene that cause the same disease: for the same and different effects on protein stability. (b) Fraction of mutation pairs on the same gene that cause the same disease: for the same and different interaction disruption profiles. Error bars indicate +SE. (c) Crystal structure (PDB id: 1U7F) depicting the Y353S and R361C mutations (on Smad4) at interface residues for the Smad4-Smad3 interaction. (d) Y2H analysis of the effects of Smad Y353S, R361, and N13S mutations on its interactions with Smad3, Lmo4, Rassf5, and Smad9. Western blotting with anti-GFP antibody confirming the protein expression levels of wild-type Smad4 and its 3 mutant alleles – Y353S, R361C and N13S. γ-tubulin was used as a loading control.

Figure 4.6. Identifying interactions of Mlh1 that are affected by the I107R mutation using SILAC-based mass spectrometry.

(a) Schematic illustrating criteria used to identify interactions that are lost/weakened, unchanged, and gained/enhanced due to the I107R mutation on Mlh1. Blue denotes samples cultured in light media and black denotes samples cultured in heavy media. (b) Scatter plot illustrating fold change (*FC*; log scale) in the amount of protein pulled down by wild-type Mlh1 and mutant Mlh1 (I107R). Values are computed based on the wild-type (heavy) vs. mutant (light) (X-axis) and mutant (heavy) vs. wild-type (light) (Y-axis) experiments. Green denotes enhancement of interaction, red denotes weakening of interaction, and gold denotes no change. Mlh1 is shown in grey. (c) Fold changes and read counts (*r*) for interactors of Mlh1 that can be reliably identified as weakened, unchanged, and enhanced due to the I107R mutation. (d) Anti-HA immunoprecipitation followed by Western blotting with anti-V5 antibody confirming that the Mlh1-Brip1 interaction remains unchanged and that the Mlh1-Hspa8 interaction is dramatically enhanced due to the I107R mutation.

Table 4.1. Summary of GFP and Y2H assay results for all the mutations tested in our interactome-scanning pipeline. For the GFP assay: “1” indicates a stable mutation, “0” indicates an unstable mutation, and “–” indicates inconclusive results due to weak signal for the wild-type protein. For the Y2H assay: “1” indicates no disruption and “0” indicates disruption of the corresponding interaction.

4.7 REFERENCES

Araya, C.L., Fowler, D.M., Chen, W., Muniez, I., Kelly, J.W., and Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proc Natl Acad Sci U S A *109*, 16858-16863.

Atlas, T.C.G. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330-337.

Bandyopadhyay, A., Saxena, K., Kasturia, N., Dalal, V., Bhatt, N., Rajkumar, A., Maity, S., Sengupta, S., and Chakraborty, K. (2012). Chemical chaperones assist intracellular folding to buffer mutational variations. Nat Chem Biol *8*, 238-245.

Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. Nature *466*, 68-76.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol *24*, 1429-1435.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.S.*, et al.* (2009). An experimentally derived confidence score for binary protein-protein interactions. Nat Methods *6*, 91-97.

Consortium, A.I.M. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Consortium, T.G.P. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Cunningham, B.C., and Wells, J.A. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science *244*, 1081-1085.

Das, J., Fragoza, R., Lee, H.R., Cordero, N.A., Guo, Y., Meyer, M.J., Vo, T.V., Wang, X., and Yu, H. (2014a). Exploring mechanisms of human disease through structurally resolved protein interactome networks. Molecular bioSystems *10*, 9-17.

Das, J., Lee, H.R., Sagar, A., Fragoza, R., Liang, J., Wei, X., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014b). Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks. Hum Mutat.

Das, J., Vo, T.V., Wei, X., Mellor, J.C., Tong, V., Degatano, A.G., Wang, X., Wang, L., Cordero, N.A., Kruer-Zerhusen, N.*, et al.* (2013). Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. Sci Signal *6*, ra38.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

DeLuca-Flaherty, C., McKay, D.B., Parham, P., and Hill, B.L. (1990). Uncoating protein (hsc70) binds a conformationally labile domain of clathrin light chain LCa to stimulate ATP hydrolysis. Cell *62*, 875-887.

Finn, R.D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics *21*, 410-412.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res *39*, D945-950.

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nat Methods *7*, 741-746.

Franzosa, E.A., and Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. Proc Natl Acad Sci U S A *108*, 10538-10543.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Goldfarb, S.B., Kashlan, O.B., Watkins, J.N., Suaud, L., Yan, W., Kleyman, T.R., and Rubenstein, R.C. (2006). Differential effects of Hsc70 and Hsp70 on the intracellular trafficking and functional expression of epithelial sodium channels. Proc Natl Acad Sci U S A *103*, 5817-5822.

Hanks, S., Coleman, K., Reid, S., Plaja, A., Firth, H., Fitzpatrick, D., Kidd, A., Mehes, K., Nash, R., Robin, N.*, et al.* (2004). Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. Nat Genet *36*, 1159-1161.

HI2012 (2012). http://interactomedfciharvardedu/indexphp?page=login&lg=/H_sapiens/indexphp?page=newrelease.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A *106*, 9362-9367.

Houlston, R., Bevan, S., Williams, A., Young, J., Dunlop, M., Rozen, P., Eng, C., Markie, D., Woodford-Richens, K., Rodriguez-Bigas, M.A.*, et al.* (1998). Mutations in DPC4 (SMAD4)

cause juvenile polyposis syndrome, but only account for a minority of cases. Hum Mol Genet *7*, 1907-1912.

Kaufer, S., Coffey, C.M., and Parker, J.S. (2012). The cellular chaperone hsc70 is specifically recruited to reovirus viral factories independently of its chaperone function. J Virol *86*, 1079-1089.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A.*, et al.* (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. Science *342*, 1235587.

Kondo, E., Suzuki, H., Horii, A., and Fukushige, S. (2003). A yeast two-hybrid assay provides a simple way to evaluate the vast majority of hMLH1 germ-line mutations. Cancer Res *63*, 3302-3308.

Leung, W.K., Kim, J.J., Wu, L., Sepulveda, J.L., and Sepulveda, A.R. (2000). Identification of a second MutL DNA mismatch repair complex (hPMS1 and hMLH1) in human epithelial cells. J Biol Chem *275*, 15728-15732.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Machado, R.D. (2012). The molecular genetics and cellular mechanisms underlying pulmonary arterial hypertension. Scientifica *2012*, 106576.

Massague, J. (2008). TGFbeta in Cancer. Cell *134*, 215-230.

Matsuura, S., Matsumoto, Y., Morishima, K., Izumi, H., Matsumoto, H., Ito, E., Tsutsui, K., Kobayashi, J., Tauchi, H., Kajiwara, Y.*, et al.* (2006). Monoallelic BUB1B mutations and defective mitotic-spindle checkpoint in seven families with premature chromatid separation (PCS) syndrome. Am J Med Genet A *140*, 358-367.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

Meyer, M.J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics *29*, 1577-1579.

Mosca, R., Ceol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nat Methods *10*, 47-53.

Nasim, M.T., Ogo, T., Ahmed, M., Randall, R., Chowdhury, H.M., Snape, K.M., Bradshaw, T.Y., Southgate, L., Lee, G.J., Jackson, I.*, et al.* (2011). Molecular genetic characterization of SMAD signaling molecules in pulmonary arterial hypertension. Hum Mutat *32*, 1385-1389.

Ohouo, P.Y., Bastos de Oliveira, F.M., Almeida, B.S., and Smolka, M.B. (2010). DNA damage signaling recruits the Rtt107-Slx4 scaffolds via Dpb11 to mediate replication stress response. Mol Cell *39*, 300-306.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics *1*, 376-386.

Ong, S.E., and Mann, M. (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). Nat Protoc *1*, 2650-2660.

Peltomaki, P., and Vasen, H.F. (1997). Mutations predisposing to hereditary nonpolyposis colorectal cancer: database and results of a collaborative study. The International Collaborative Group on Hereditary Nonpolyposis Colorectal Cancer. Gastroenterology *113*, 1146-1158.

Peng, M., Litman, R., Xie, J., Sharma, S., Brosh, R.M., Jr., and Cantor, S.B. (2007). The FANCJ/MutLalpha interaction is required for correction of the cross-link response in FA-J cells. EMBO J *26*, 3238-3249.

Pitt, J.N., and Ferre-D'Amare, A.R. (2010). Rapid construction of empirical RNA fitness landscapes. Science *330*, 376-379.

Reece-Hoyes, J.S., Barutcu, A.R., McCord, R.P., Jeong, J.S., Jiang, L., MacWilliams, A., Yang, X., Salehi-Ashtiani, K., Hill, D.E., Blackshaw, S.*, et al.* (2011). Yeast one-hybrid assays for gene-centered human gene regulatory network mapping. Nat Methods *8*, 1050-1052.

Roth, S., Sistonen, P., Salovaara, R., Hemminki, A., Loukola, A., Johansson, M., Avizienyte, E., Cleary, K.A., Lynch, P., Amos, C.I.*, et al.* (1999). SMAD genes in juvenile polyposis. Genes, chromosomes & cancer *26*, 54-61.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N.*, et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.

Salehi-Ashtiani, K., Yang, X., Derti, A., Tian, W., Hao, T., Lin, C., Makowski, K., Shen, L., Murray, R.R., Szeto, D.*, et al.* (2008). Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing. Nat Methods *5*, 597-600.

Starita, L.M., Pruneda, J.N., Lo, R.S., Fowler, D.M., Kim, H.J., Hiatt, J.B., Shendure, J., Brzovic, P.S., Fields, S., and Klevit, R.E. (2013). Activity-enhancing mutations in an E3

ubiquitin ligase identified by high-throughput mutagenesis. Proc Natl Acad Sci U S A *110*, E1263-1272.

Stein, A., Ceol, A., and Aloy, P. (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res *39*, D718-723.

Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S., and Cooper, D.N. (2009). The Human Gene Mutation Database: 2008 update. Genome Med *1*, 13.

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A.*, et al.* (2011). The mutational landscape of head and neck squamous cell carcinoma. Science *333*, 1157-1160.

Suijkerbuijk, S.J., van Osch, M.H., Bos, F.L., Hanks, S., Rahman, N., and Kops, G.J. (2010). Molecular causes for BUBR1 dysfunction in the human cancer predisposition syndrome mosaic variegated aneuploidy. Cancer Res *70*, 4891-4900.

Suzuki, Y., Kagawa, N., Fujino, T., Sumiya, T., Andoh, T., Ishikawa, K., Kimura, R., Kemmochi, K., Ohta, T., and Tanaka, S. (2005). A novel high-throughput (HTP) cloning strategy for site-directed designed chimeragenesis and mutation using the Gateway cloning system. Nucleic Acids Res *33*, e109.

Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat Protoc *6*, 1341-1354.

Vandenbroucke, I., Van Marck, H., Verhasselt, P., Thys, K., Mostmans, W., Dumont, S., Van Eygen, V., Coen, K., Tuefferd, M., and Aerssens, J. (2011). Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. Biotechniques *51*, 167-177.

Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Vidal, M., Cusick, M.E., and Barabasi, A.L. (2011). Interactome networks and human disease. Cell *144*, 986-998.

Walhout, A.J., and Vidal, M. (2001). High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. Methods *24*, 297-306.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol *30*, 159-164.

Yakhnin, A.V., Yakhnin, H., and Babitzke, P. (2012). Gel mobility shift assays to detect protein-RNA interactions. Methods Mol Biol *905*, 201-211.

Yang, X., Boehm, J.S., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T., Green, T.M.*, et al.* (2011). A public genome-scale lentiviral expression library of human ORFs. Nat Methods *8*, 659-661.

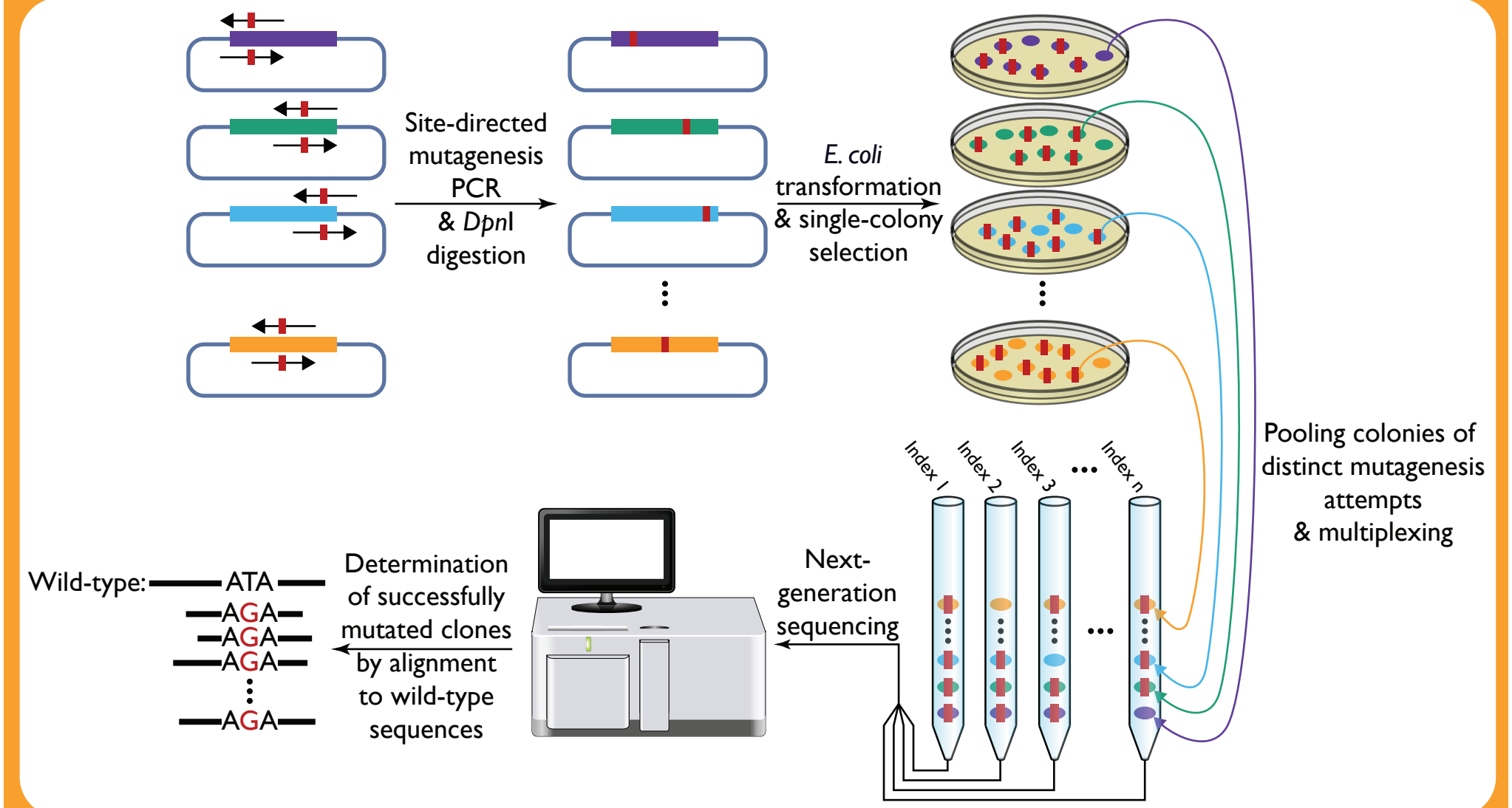
Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X.*, et al.* (2011). Next-generation sequencing to generate interactome datasets. Nat Methods *8*, 478-480.

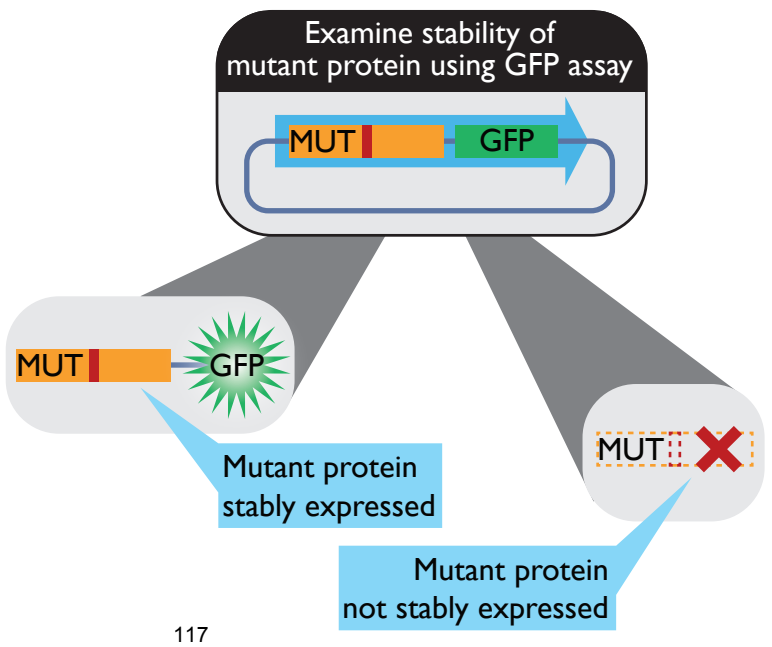
Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T.*, et al.* (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature *490*, 556-560.

Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D.*, et al.* (2009). Edgetic perturbation models of human inherited disorders. Mol Syst Biol *5*, 321.

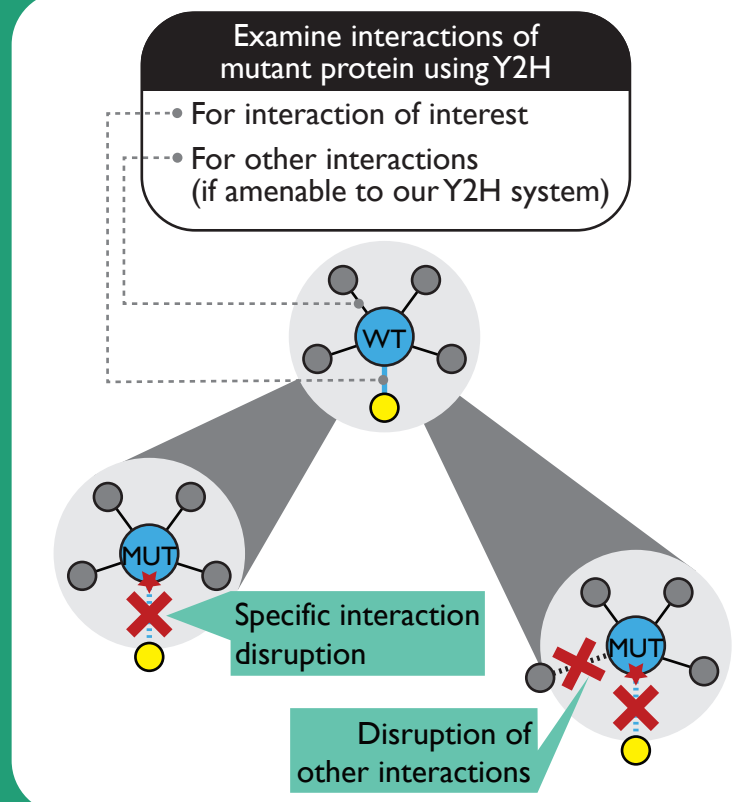
a. Clone-seq



b. GFP



c. Y2H



d. SILAC

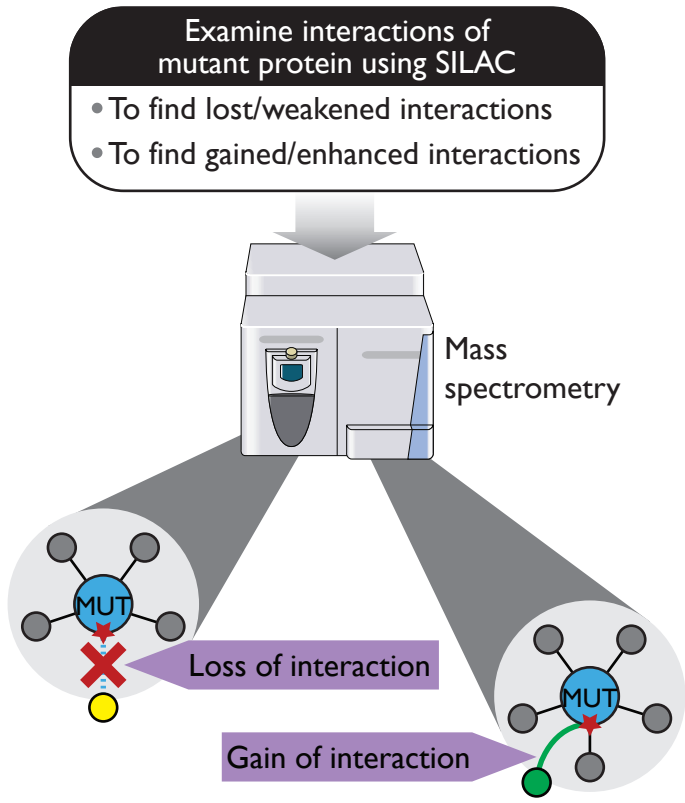


Figure 4.1

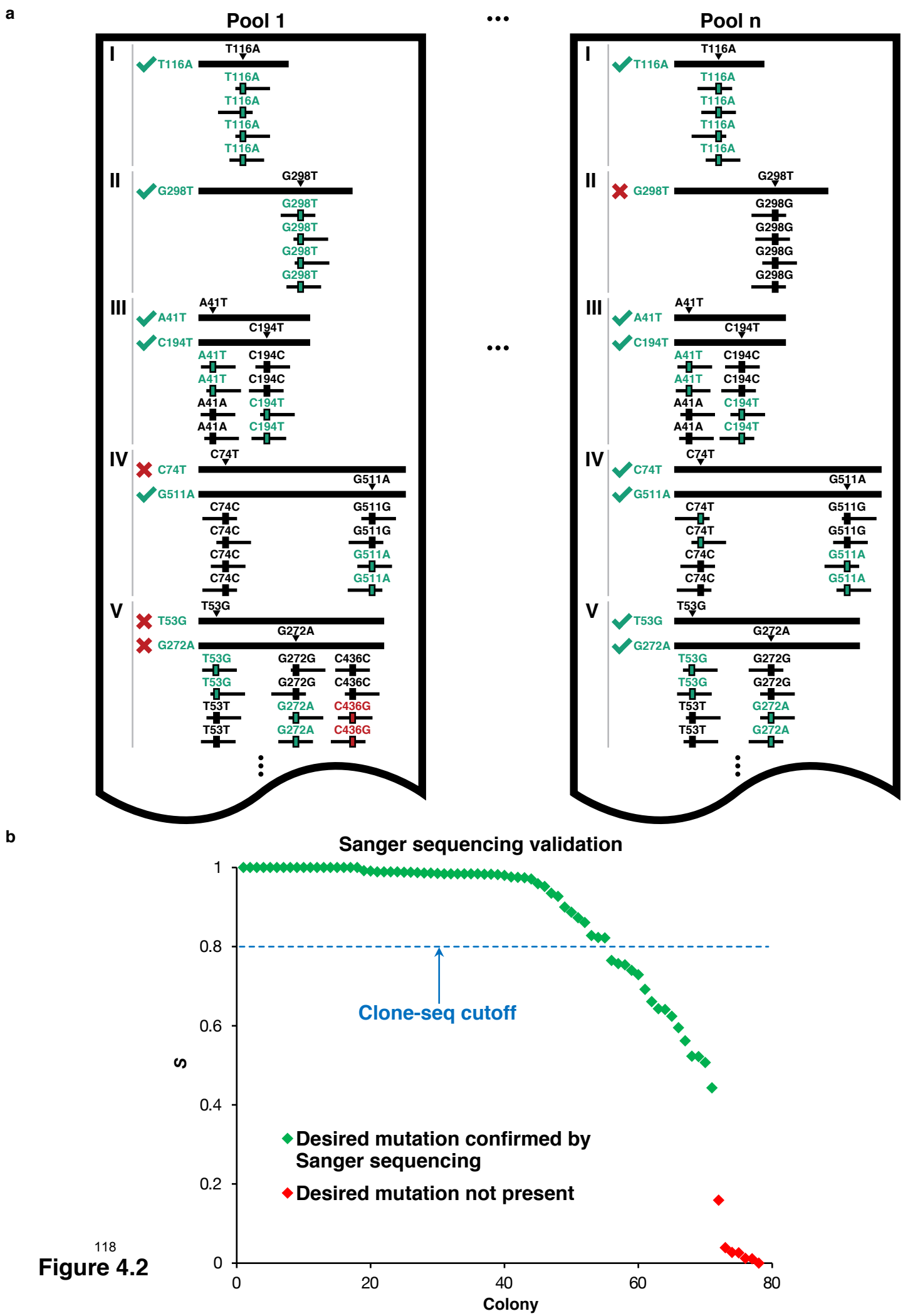
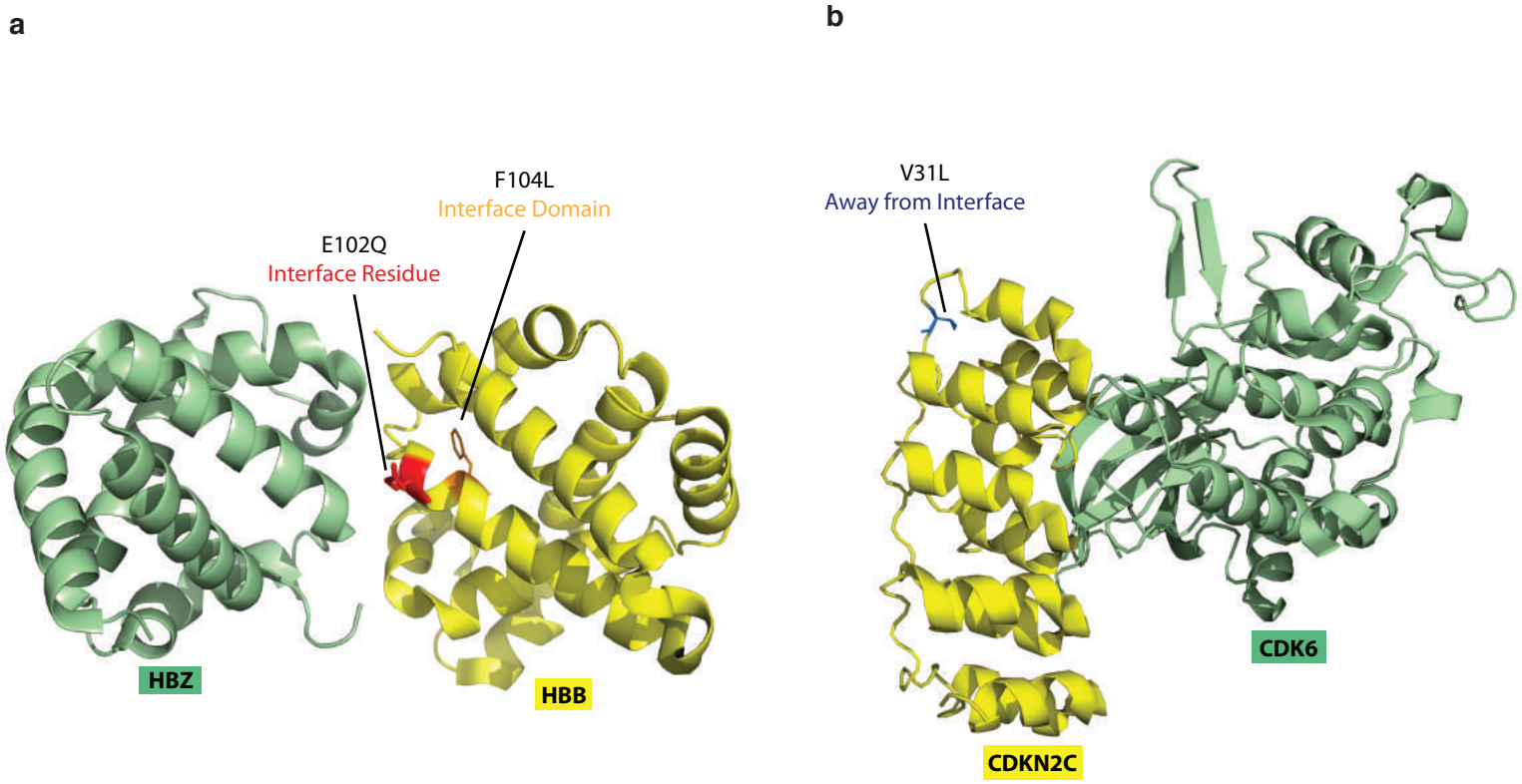

a
Pool 1
Pool n
I
II
III
IV
V
T116A
G298T
G298G
A41T
C194T
A41A
C194C
C74T
C74C
G511A
G511G
T53G
T53T
G272A
G272G
C436C
C436G
b
Sanger sequencing validation
Clone-seq cutoff
s
Colony
Desired mutation confirmed by Sanger sequencing
Desired mutation not present
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80

Figure 4.2


a
F104L
Interface Domain
E102Q
Interface Residue
HBZ
HBB
b
V31L
Away from Interface
CDK6
CDKN2C

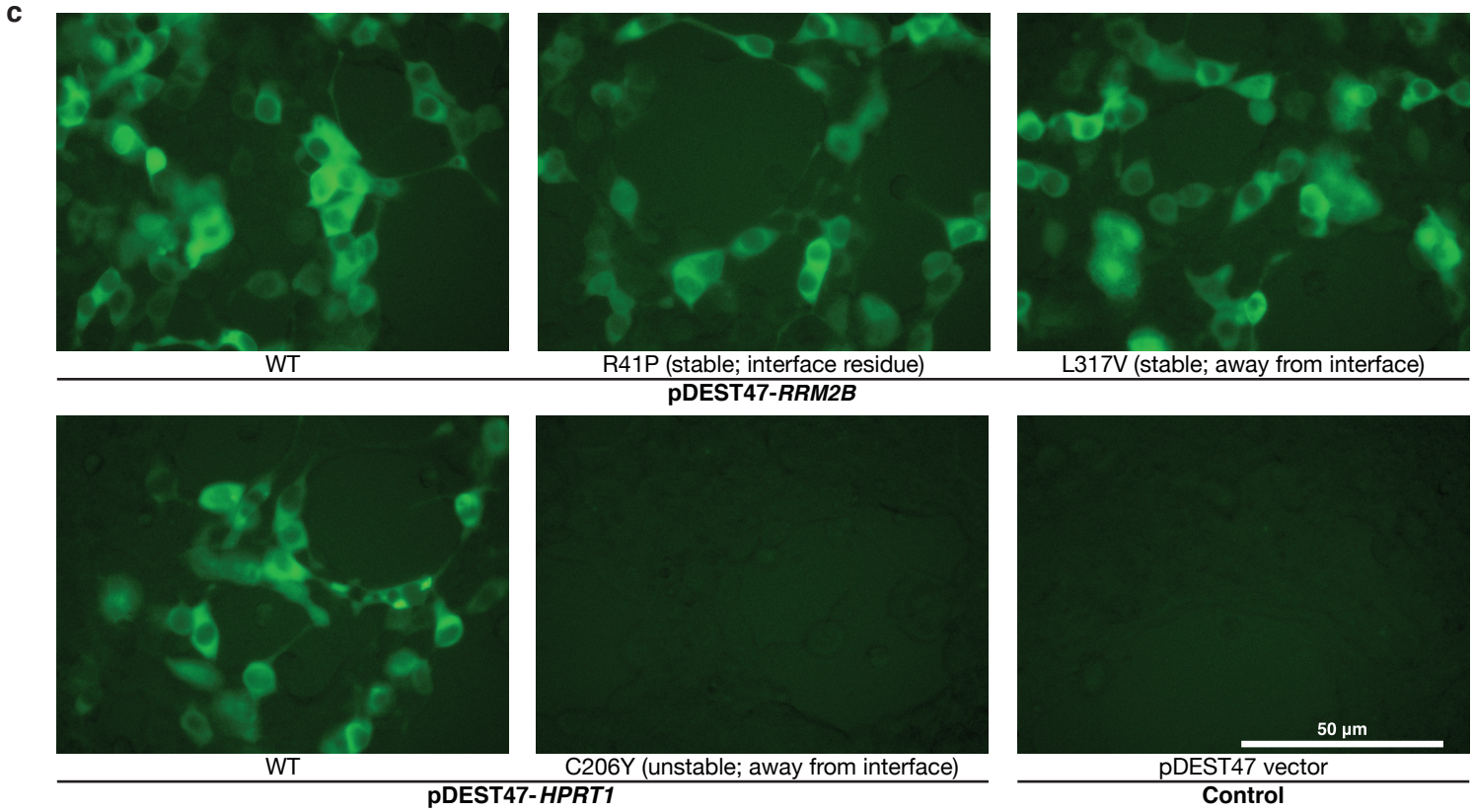
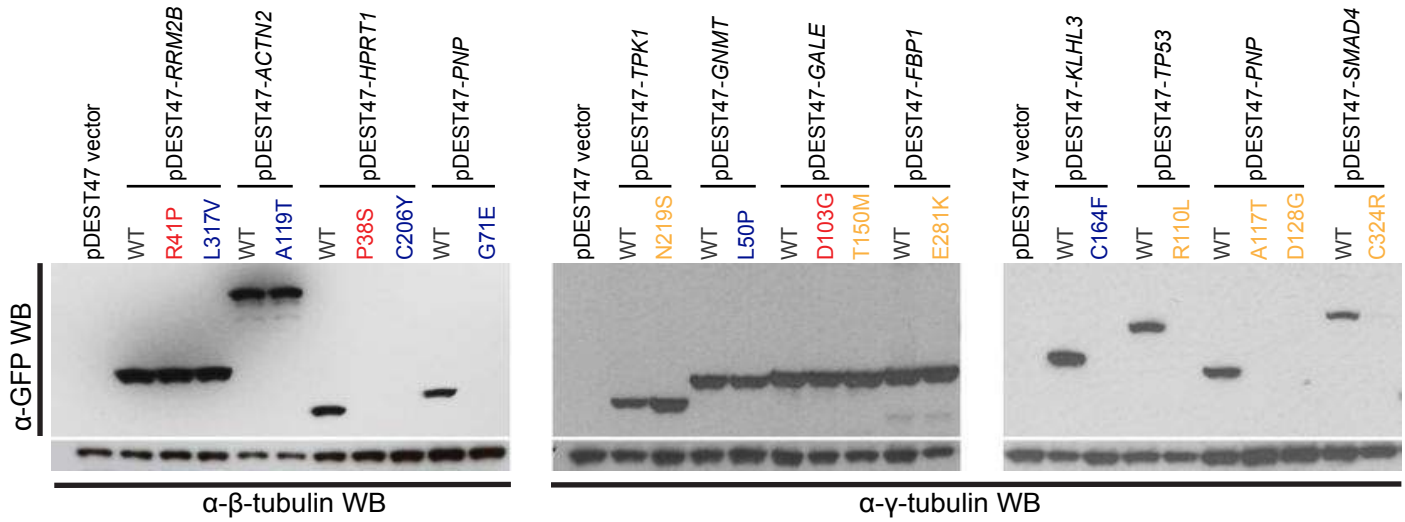

c
WT
R41P (stable; interface residue)
L317V (stable; away from interface)
pDEST47-RRM2B
WT
C206Y (unstable; away from interface)
pDEST47 vector
50 µm
pDEST47-HPRT1
Control

Figure 4.3

a


pDEST47 vector
pDEST47-RRM2B
WT
R41P
L317V
pDEST47-ACTN2
WT
A119T
pDEST47-HPRT1
WT
P38S
C206Y
pDEST47-PNP
WT
G71E
pDEST47 vector
pDEST47-TPK1
WT
N219S
pDEST47-GNMT
WT
L50P
pDEST47-GALE
WT
D103G
T150M
pDEST47-FBP1
WT
E281K
pDEST47 vector
pDEST47-KLHL3
WT
C164F
pDEST47-TP53
WT
R110L
pDEST47-PNP
WT
A117T
D128G
pDEST47-SMAD4
WT
C324R
α-GFP WB
α-β-tubulin WB
α-γ-tubulin WB

b

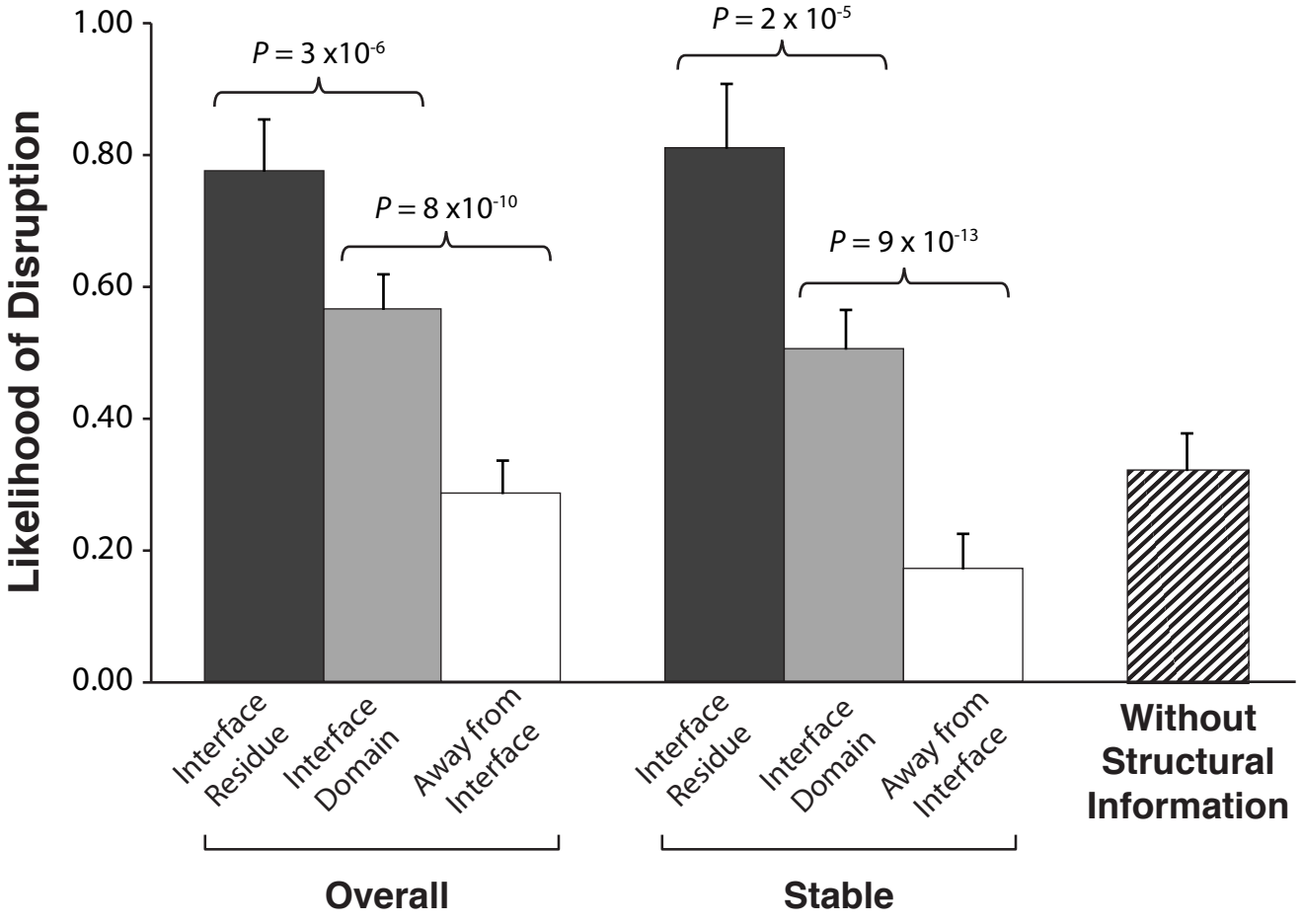

Likelihood of Disruption
1.00
0.80
0.60
0.40
0.20
0.00
P = 3 x10-6
P = 8 x10-10
P = 2 x 10-5
P = 9 x 10-13
Interface Residue
Interface Domain
Away from Interface
Interface Residue
Interface Domain
Away from Interface
Without Structural Information
Overall
Stable

Figure 4.4

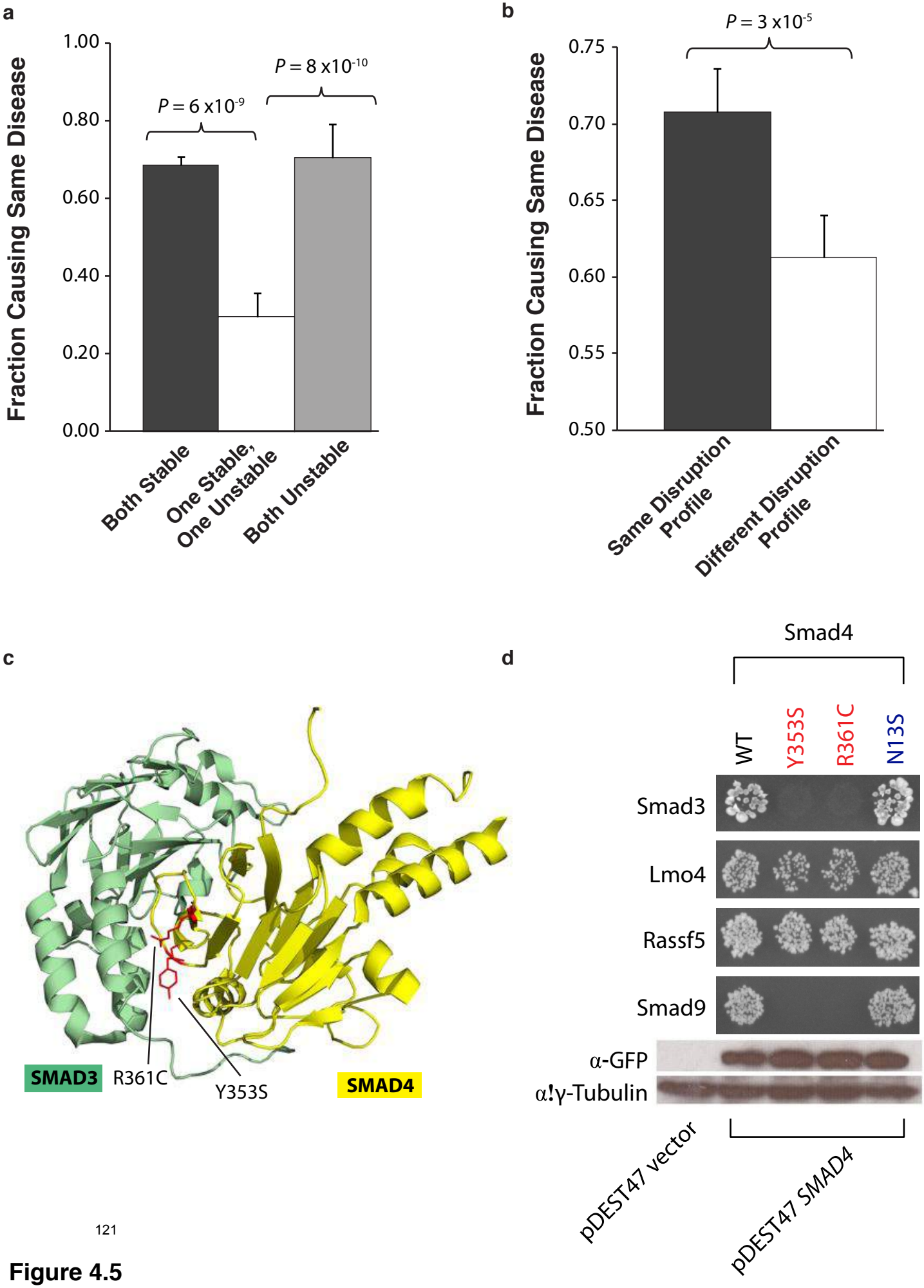
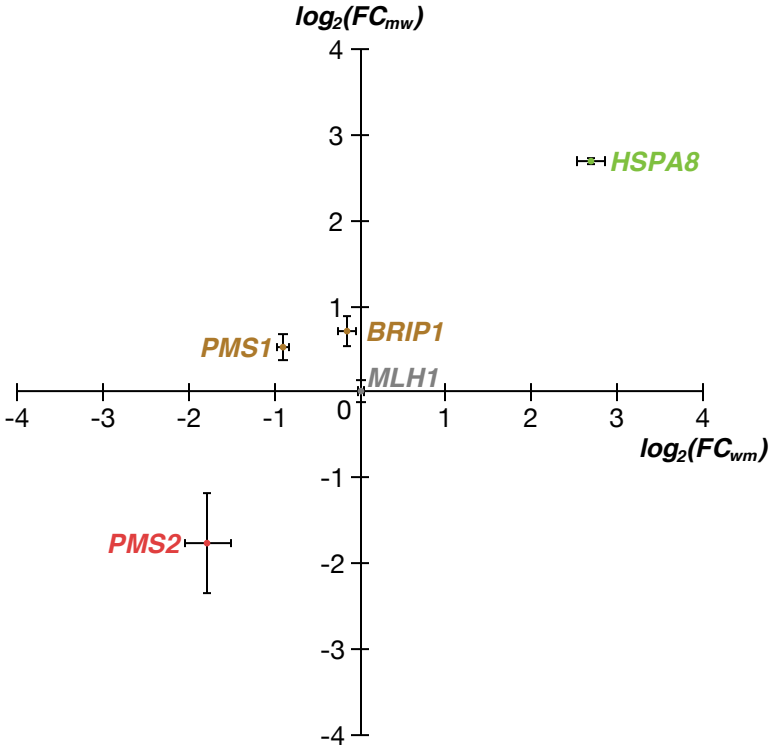

a
Fraction Causing Same Disease
1.00
0.80
0.60
0.40
0.20
0.00
P = 6 x10-9
P = 8 x10-10
Both Stable
One Stable, One Unstable
Both Unstable
b
Fraction Causing Same Disease
0.75
0.70
0.65
0.60
0.55
0.50
P = 3 x10-5
Same Disruption Profile
Different Disruption Profile
c
SMAD3
R361C
Y353S
SMAD4
d
Smad4
WT
Y353S
R361C
N13S
Smad3
Lmo4
Rassf5
Smad9
α-GFP
α!γ-Tubulin
pDEST47 vector
pDEST47 SMAD4

Figure 4.5

a

	SILAC (light vs. heavy)			
	MUT / WT	MUT / WT	WT / GFP	MUT / GFP
Lost / Weakened	-	-	+	n/a
Unchanged	≈	≈	+	+
Gained / Enhanced	+	+	n/a	+

b



c

Protein	FC_{wm}	r_{wm}	FC_{mw}	r_{mw}	WT_p	r_{wc}	Mut_p	r_{mc}
PMS2	0.29	10	0.29	22	18.69	37	17.66	3
PMS1	0.53	53	1.45	37	14.04	55	10.11	25
BRIP1	0.90	66	1.64	40	13.92	40	15.01	63
HSPA8	6.50	56	6.56	126	0.64	79	3.36	201

d

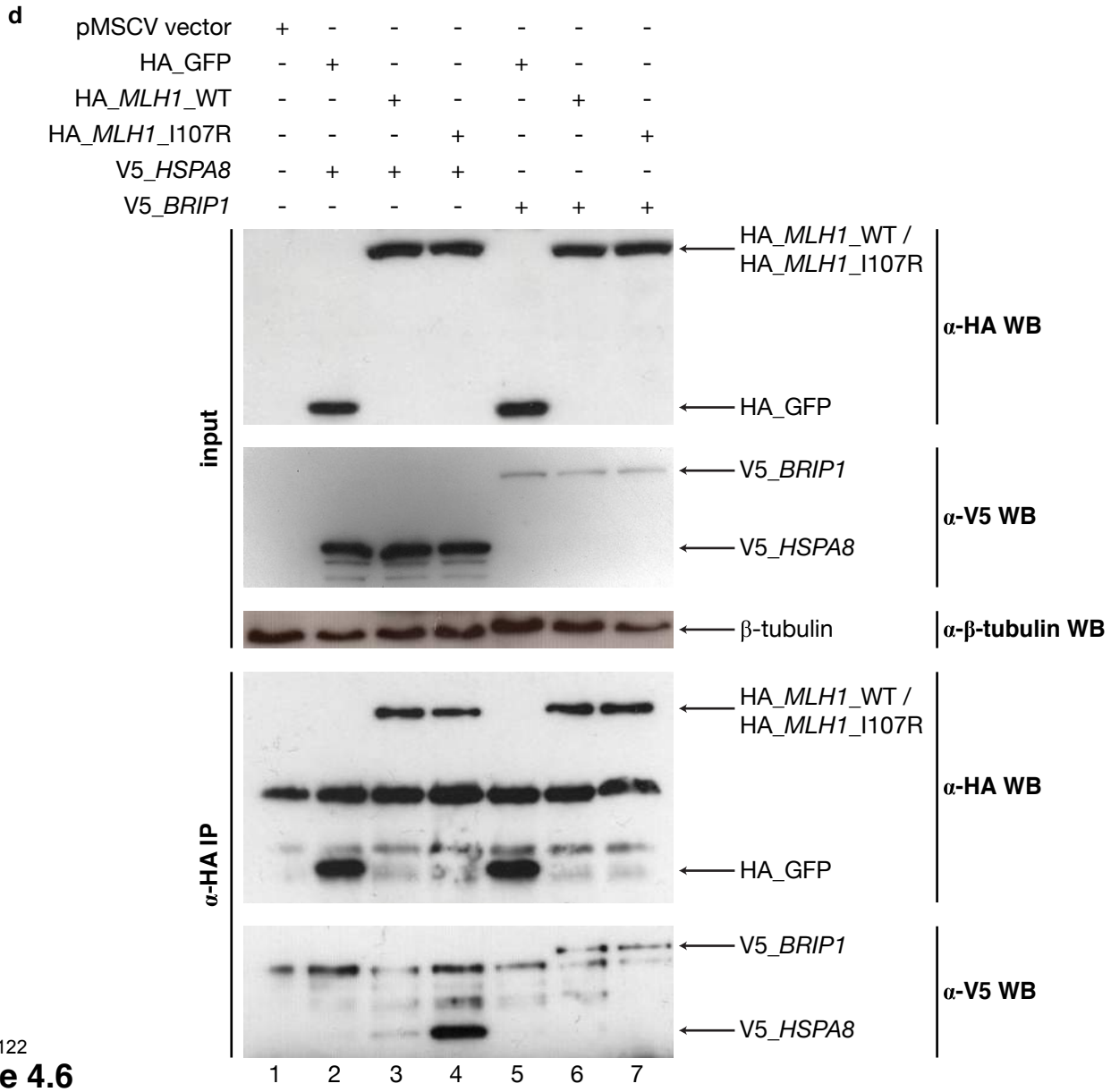


Figure 4.6

Table 4.1

Interface residue

TargetProt_entrez	Target_uniprot	Interactor_entrez	Mutation	GFP score	Y2H score
60	P60709	60	R196C	1	0
95	Q03154	95	R197W	1	0
95	Q03154	95	R378W	1	1
1967	Q14232	1967	V183F	1	0
1967	Q14232	1967	P278R	1	0
2203	P09467	2203	N213K	1	0
2582	Q14376	2582	D103G	1	0
2582	Q14376	2582	R169W	1	1
2752	P15104	2752	R324C	1	1
3043	P68871	3050	D100Y	-	0
3043	P68871	3050	P101A	-	0
3043	P68871	3050	E102Q	-	0
3043	P68871	3050	N103Y	-	1
3251	P00492	3251	P38S	0	0
3945	P07195	3945	R172P	1	0
4088	P84022	9372	T330A	1	1
4089	Q13485	4088	G352R	0	0
4089	Q13485	4088	Y353S	1	0
4089	Q13485	4088	R361C	1	0
4089	Q13485	4088	L533R	1	0
4860	P00491	4860	F159V	0	0
5631	P60891	5631	D183H	1	0
5805	Q03393	5805	T106M	-	0
7329	P63279	7341	V25M	1	0
50484	Q7LG56	50484	R41P	1	0
50484	Q7LG56	50484	R121H	1	1
51135	Q9NWZ3	4615	R12C	-	0

Interface domain

TargetProt_entrez	Target_uniprot	Interactor_entrez	Mutation	GFP score	
60	P60709	60	E364K	1	1
60	P60709	60	N12D	1	0
60	P60709	60	L65V	1	1
60	P60709	60	E117K	1	0
60	P60709	60	R183W	1	0
88	P35609	88	T495M	1	1
88	P35609	88	Q349L	1	1
88	P35609	88	E583A	1	1
95	Q03154	95	E233D	1	0
95	Q03154	95	R386C	1	0
95	Q03154	95	R393H	1	1
435	P04424	435	R12Q	1	1

875	P35520	875	G85R	1	0
875	P35520	875	L101P	1	0
875	P35520	875	K102Q	1	1
875	P35520	875	C109R	1	0
875	P35520	875	A114V	1	1
875	P35520	875	G116R	1	0
1019	P11802	595	S52N	1	1
1019	P11802	595	R24H	1	1
1019	P11802	595	N41S	1	1
1967	Q14232	1967	N208Y	1	0
1967	Q14232	1967	Y275C	1	1
2203	P09467	2203	G164S	1	0
2203	P09467	2203	A177D	1	0
2203	P09467	2203	F194S	1	0
2203	P09467	2203	G260R	1	0
2203	P09467	2203	E281K	1	1
2512	P02792	2512	Q26L	1	0
2512	P02792	2512	A27V	1	0
2512	P02792	2512	T30I	1	0
2512	P02792	2512	A96T	1	0
2582	Q14376	2582	T150M	1	1
2582	Q14376	2582	K161N	1	1
2582	Q14376	2582	E165K	1	0
2582	Q14376	2582	D175N	1	1
2752	P15104	2752	R341C	1	0
3043	P68871	3050	S10C	-	0
3043	P68871	3050	F104L	-	0
3251	P00492	3251	Y105C	1	1
3251	P00492	3251	S110L	1	1
3251	P00492	3251	T124S	1	1
3611	Q13418	55742	A262V	-	1
3939	P00338	3939	K222E	1	1
4085	Q13257	9587	L84M	1	0
4088	P84022	9372	E239K	1	1
4088	P84022	9372	T261I	1	1
4088	P84022	9372	P263L	1	1
4088	P84022	9372	R279K	1	1
4088	P84022	9372	R287W	1	1
4089	Q13485	4088	C324R	0	0
4089	Q13485	4088	E330K	0	0
4089	Q13485	4088	V350D	0	0
4598	Q03426	4598	N301T	1	0
4615	Q99836	51135	L93P	1	0
4830	P15531	4830	S120G	1	0
4860	P00491	4860	G71E	0	0
4860	P00491	4860	A117T	0	0
4860	P00491	4860	D128G	0	0
4860	P00491	4860	P146L	0	0
4860	P00491	4860	G156A	0	0

5723	P78330	5723	D32N	1	1
5723	P78330	5723	M52T	1	1
5805	Q03393	5805	R25Q	-	0
5805	Q03393	5805	F100V	-	0
5805	Q03393	5805	A101V	-	0
5805	Q03393	5805	V103A	-	1
5805	Q03393	5805	A111T	-	0
6898	P17735	6898	R119W	1	0
6898	P17735	6898	C151Y	1	0
6898	P17735	6898	L201R	1	0
6898	P17735	6898	P220S	1	1
6898	P17735	6898	L273P	1	0
6898	P17735	6898	G362V	1	0
7128	P21580	7128	A125V	1	1
7157	P04637	7159	G105C	1	0
7157	P04637	7159	S106R	1	1
7157	P04637	7159	R110L	0	0
7157	P04637	7159	V122G	1	1
7157	P04637	7159	Y126C	1	0
7454	P42768	998	I294T	1	0
8772	Q13158	8772	C105W	1	0
8815	O75531	2010	A12T	1	1
11144	Q14565	11144	M200V	1	1
23568	Q9Y2Y0	402	M45R	1	0
27010	Q9H3S4	27010	N219S	1	0
27010	Q9H3S4	27010	L40P	1	1
27010	Q9H3S4	27010	N50H	1	1
27232	Q14749	27232	N141S	1	0
27232	Q14749	27232	H177N	1	0
50484	Q7LG56	50484	R110C	1	1
50484	Q7LG56	50484	F123S	1	0
50484	Q7LG56	50484	E131K	1	1
50484	Q7LG56	50484	T144I	1	1
64802	Q9HAN9	64802	R66W	1	0
64802	Q9HAN9	64802	A13T	1	1
64802	Q9HAN9	64802	A147P	1	1
64802	Q9HAN9	64802	V151F	1	0
64802	Q9HAN9	64802	L153V	1	0
64802	Q9HAN9	64802	D173G	1	0

Away from the interface

TargetProt_entrez	UniprotID	Interactor_entrez	AA_mut	GFP score	
88	P35609	88	A119T	1	1
88	P35609	88	Q9R	1	1
88	P35609	88	V115M	1	1
88	P35609	88	E628G	1	1
88	P35609	88	H775Y	1	1

331	P98170	842	G188E	-	1
331	P98170	842	R166I	-	1
331	P98170	842	W173G	-	1
331	P98170	842	V198M	-	1
331	P98170	842	C203Y	-	1
331	P98170	842	L207P	-	1
701	O60566	991	L1012P	0	0
701	O60566	991	R36Q	1	1
701	O60566	991	Y155C	1	1
701	O60566	991	R727C	0	0
701	O60566	991	L844F	0	0
875	P35520	875	L456P	1	0
875	P35520	875	R379W	1	1
875	P35520	875	K384E	1	1
875	P35520	875	M391I	1	1
875	P35520	875	P422L	1	1
875	P35520	875	P427L	1	1
958	P25942	7186	C83R	1	0
1026	Q6FI05	5111	R67L	1	1
1026	Q6FI05	5111	R84Q	1	1
1031	Q6ICV4	1021	V31L	1	1
2010	P50402	8815	P183T	-	1
2010	P50402	8815	S54F	1	1
2010	P50402	8815	D72V	1	1
2582	Q14376	2582	P293L	1	1
2582	Q14376	2582	G302D	1	1
2582	Q14376	2582	L313M	1	1
2582	Q14376	2582	G319E	1	1
2582	Q14376	2582	R335H	1	1
3043	P68871	3050	V114E	-	0
3043	P68871	3050	L115P	-	0
3043	P68871	3050	H118Y	-	0
3043	P68871	3050	E122V	-	0
3043	P68871	3050	F123S	-	0
3043	P68871	3050	V127G	-	0
3251	P00492	3251	C206Y	0	0
3251	P00492	3251	I10S	1	0
3251	P00492	3251	D12A	1	0
3251	P00492	3251	E14K	1	0
3251	P00492	3251	R167M	0	0
3251	P00492	3251	T168I	1	0
3945	P07195	3945	K7E	1	1
4088	P84022	9372	A112V	1	1
4088	P84022	9372	N197I	1	1
4089	Q13485	4088	N13S	1	1
4598	Q03426	4598	H20Q	1	0
4615	Q99836	51135	R196C	1	0
5631	P60891	5631	E43D	1	1
5805	Q03393	5805	R9C	-	1

6829	O00267	6827	E455D	1	1
7157	P04637	7159	R290L	1	1
7157	P04637	7159	K292I	1	1
7157	P04637	7159	G293W	1	1
7157	P04637	7159	K305M	1	1
7157	P04637	7159	R306P	1	1
7157	P04637	7159	P309S	1	1
7454	P42768	998	E131K	1	1
8504	P56589	5824	D347Y	-	1
26249	Q9UH77	8452	C164F	0	0
26249	Q9UH77	8452	R228G	1	0
27232	Q14749	27232	L50P	1	1
50484	Q7LG56	50484	L317V	1	1
51135	Q9NWZ3	4615	G298D	-	0
51135	Q9NWZ3	4615	A428T	-	1
55737	Q96QK1	51699	R524W	-	1
55737	Q96QK1	51699	D620N	-	1
64802	Q9HAN9	64802	N273D	1	1
64802	Q9HAN9	64802	V9M	1	1
64802	Q9HAN9	64802	H251P	1	1
64802	Q9HAN9	64802	E257K	1	1
124590	Q495M9	10083	L16V	1	1
124590	Q495M9	10083	L48P	1	1

CHAPTER 5

Genome-scale analysis of interaction dynamics reveals organization of biological networks

In the following chapter, we explore the concept of "interaction dynamics" and how it can be used to understand the topological and biological properties of networks. I am the first author of the paper resulting from this chapter (Das et al Bioinformatics 2012) and led all computational analyses. Jaaved Mohammed made a significant contribution to several of the analyses in the paper.

5.1 ABSTRACT

Analyzing large-scale interaction networks has generated numerous insights in systems biology. However, such studies have primarily been focused on highly co-expressed, stable interactions. Most transient interactions that carry out equally important functions, especially in signal transduction pathways, are yet to be elucidated and are often wrongly discarded as false positives. Here, we revisit a previously described Smith-Waterman-like dynamic programming algorithm and use it to distinguish stable and transient interactions on a genomic scale in human and yeast. We find that in biological networks, transient interactions are key links topologically connecting tightly regulated functional modules formed by stable interactions and are essential to maintaining the integrity of cellular networks. We also perform a systematic analysis of interaction dynamics across different technologies and find that high-throughput yeast two-hybrid (Y2H) is the only available technology for detecting transient interactions on a large scale.

5.2 INTRODUCTION

The protein-protein interactome of an organism is the network of all biophysically possible interactions of different proteins in that organism (Yu et al., 2008). It is of key importance to accurately map this network as most proteins function by interacting with other proteins (Pawson and Nash, 2000). Moreover, a better understanding of genotype to phenotype relationships in human disease require modeling of how disease-causing mutations might affect protein interactions and interactome properties (Goh et al., 2007; Wang et al., 2011). Currently, there are two main high-throughput technologies to generate high-quality protein-protein interactomes on a large-scale: yeast two-hybrid (Y2H), where a protein interaction reconstitutes a transcription factor which then activates expression of reporter genes (Fields and Song, 1989); and affinity purification followed by mass spectrometry (AP/MS), where proteins bound to tagged baits are co-purified and identified (Rigaut et al., 1999). High-throughput Y2H maps have been generated for yeast, fly, worm, and human, while large-scale AP/MS datasets have been generated for yeast, worm and human (Jensen and Bork, 2008; Yu et al., 2008). An alternative approach, adopted by most databases, is to obtain literature-curated (LC) interactions (Cusick et al., 2009).
It has been shown that well-controlled Y2H and AP/MS experiments are both of high quality, but of complementary nature – Y2H identifies direct binary interactions whereas AP/MS determines co-complex associations (Jensen and Bork, 2008; Yu et al., 2008). Moreover, gene expression and other functional genomics datasets are routinely integrated with protein-protein interactions to validate their biological relevance - for example, interactions between proteins encoded by co-expressed genes are often considered to be of high quality (Ge et al., 2001; Suthram et al., 2006; von Mering et al., 2002). In these analyses, gene co-expression is normally determined by a high Pearson correlation coefficient (PCC), which really means that the expression levels of the two

genes are correlated over most conditions i.e., they are globally co-expressed (Figure 5.1A). Previous studies have shown that interacting proteins within stable complexes also tend to be encoded by globally co-expressed gene pairs (Jansen et al., 2002; Yu et al., 2008). On the other hand, the regulation and coordination of the sub-cellular machinery is achieved by dynamic transient interactions for example in signal transduction pathways (Jansen et al., 2002). Proteins involved in transient interactions are not globally co-expressed. Rather, they share local blocks of co-expression. Transient interactions and their dynamics have significant biological importance but most genes in these pathways are often co-expressed only under certain conditions (Figure 5.1B). As a result, these are usually discarded as false positives (Ge et al., 2001; Suthram et al., 2006). Here, we take advantage of a novel measurement of expression relationships (Qian et al., 2001) to directly distinguish stable from transient interactions on a genome-wide scale in human and yeast and systematically analyze their topological and biological significance. We also evaluate different technologies in terms of their sensitivity in detecting interaction dynamics on a genomic scale.

5. 3 RESULTS

Expression dynamics: global vs. local co-expression

For our analysis, we created compendiums of gene expression and high-quality large-scale protein-protein interaction datasets for human and yeast. We decided to use time course datasets because four distinct kinds of expression relationships - co-expression, time-shifted, inverted, and inverted time-shifted can be determined using such datasets (Qian et al., 2001). As the cell is in a different state at each of these time points, we are in fact measuring expression under

different intra-cellular conditions. All datasets are carefully normalized to remove potential noise (Irizarry et al., 2003; Johnson et al., 2007; Luscombe et al., 2003; Yu et al., 2007b). We also compiled high-quality large-scale protein-protein interaction datasets for human and yeast spanning both high-throughput technologies - Y2H and AP/MS. We consolidated high-quality binary interactions in the literature from various databases. Although traditionally these LC interactions are considered to be of high quality, recent studies have shown that many of them, especially those supported by only one publication, in fact tend to be false positives (Cusick et al., 2009). To remove unreliable interactions from our analysis, we carefully compiled comprehensive sets of high-quality binary LC interactions supported by multiple publications (named "LC-multiple") for human and yeast. High-quality LC co-complex associations were obtained from MIPS (Mewes et al., 2011) for yeast and Reactome (D'Eustachio, 2011) for human - two databases generally considered as gold standards for complexes in the corresponding organisms (Jansen et al., 2002; Lage et al., 2007).

From these datasets, we first calculated the PCC for expression profiles corresponding to interacting protein pairs in the high-quality interaction datasets described above. For a pair of gene expression profiles, PCC reports the global correlation of expression levels across all conditions (Qian et al., 2001). A PCC value close to one indicates the pair of genes is globally co-expressed (Figure 5.1A), whereas values close to zero indicate random, uncorrelated expression patterns. We find that the different interaction datasets for both human and yeast are significantly enriched for global co-expression as opposed to random gene pairs (Figures 5.2A and 5.2B). Since PCC is a linear correlation coefficient and certain co-expression relationships could be non-linear, we also used the maximal information coefficient (MIC) (Reshef et al., 2011) to explore global expression dynamics of the different interaction datasets in human and

yeast. MIC belongs to a class of maximal information-based nonparametric exploration (MINE) statistics and has been shown to be very robust in detecting a wide range of associations both linear and not (Reshef et al., 2011). Using MIC, we re-validate the global expression dynamics captured by PCC – all the high-quality interaction datasets in both human and yeast have significantly enriched global co-expression as opposed to random gene pairs. Interacting protein pairs that have PCC greater than a certain cutoff are defined as stable interactions. However, gene pairs that are only co-expressed under certain conditions could have low and non-significant global PCC/MIC values. These often go undetected in the global nature of the computation, making global correlation an ineffective method for identifying condition-specific characteristics of transient interactions. To define dynamic co-expression relationships, we employed a Smith-Waterman-like dynamic programming algorithm as described previously (Qian et al., 2001). For each pair of genes and their expression profiles, this algorithm calculates local expression-correlation scores (LES) to find subsets of conditions with correlated expression levels (Figure 5.1B). Interacting proteins that do not pass the global PCC cutoff but have high LES are defined as transient interactions (see Methods).

Interaction dynamics: stable vs. transient

Next, in order to explore interaction dynamics across different technologies, we compared how successful different experimental techniques were in detecting stable and transient interactions. In agreement with previous studies, stable interactions within sub-cellular complexes show a strong enrichment of proteins encoded by globally co-expressed genes (Figures 5.2a and 5.2B). On the other hand, although statistically significant, the enrichment of these globally-co-expressed pairs is much less for binary interactions from both large-scale Y2H and LC sources.

This lack of global co-expression has often been used as an argument to suggest that high-throughput Y2H interactions are of low quality (Ge et al., 2001; Suthram et al., 2006; von Mering et al., 2002). However, a recent study applied orthogonal assays to experimentally confirm that these binary interactions are in fact highly reliable (Yu et al., 2008). Figures 5.2C and 5.2D show that in both human and yeast, Y2H is the only technology consistently able to identify transient interactions significantly more than random expectation. Surprisingly, binary interactions from the literature are not enriched with transient ones. Given the sociological biases within interactions from the literature (Cusick et al., 2009; Yu et al., 2008), there might be many compounding factors for this result. Stable interactions are easier to recapitulate under different experimental conditions whereas transient interactions can only be tested under specific conditions. Therefore, transient interactions are more likely to be considered as false positives and not reported in the literature. Additionally, in the post-genomic era, many candidate interaction partners are first identified based on gene expression and other genomic features favoring selection of stable interactions over transient ones. This result further highlights the importance of high-throughput Y2H because it is the only technology available to detect transient interactions, confirming that different protein interaction detection technologies capture different modes of biochemical interactions (Jensen and Bork, 2008; Yu et al., 2008).

Biological significance of transient interactions

To assess the biological significance of transient interactions as defined by our algorithm, we computed functional similarity of protein pairs involved in these interactions. We find that transient interactions are significantly enriched for proteins with similar functions and the fold enrichment is comparable to that of stable interactions in both human and yeast. These results

confirm the validity of our definition of transient interactions. We therefore provide the first method to systematically detect transient interactions on a genomic scale. Although our method might miss certain transient interactions, especially extremely transient ones that are virtually impossible to distinguish from random, our results confirm that those detected by our method are high-quality and share significant functional similarity.

A good example of transient interactions identified by Y2H is the interaction between Sfb2 and Sec23. This interaction has been confirmed *in vivo* (Peng et al., 2000). Sec23 is a subunit of the COPII complex, required for the budding of transport vesicles from endoplasmic reticulum (Miller et al., 2003). *SFB2* has a 56% sequence identity with *SEC24*, an essential component of COPII involved in cargo selection (Miller et al., 2003). Over-expression of *SFB2* can rescue the *sec24* null mutant cells (Kurihara et al., 2000). Furthermore, it has been suggested experimentally that Sfb2 may recognize different export signals from those of Sec24 and may be used under non-normal growth conditions (Miller et al., 2003; Peng et al., 2000). These results agree with the expression dynamics revealed by our new analysis – *SFB2* and *SEC23* are only co-expressed during stress response (Figure 5.3A).

Transient interactions key in maintaining network integrity

Traditionally, in network analysis, the focus has been on nodes. Hubs are crucial in maintaining the integrity of biological networks (Albert et al., 2000; Barabasi and Albert, 1999; Jeong et al., 2000). Interaction networks have two broad categories of hubs. Date hubs have low average PCCs with their interactors and hold the key in maintaining the integrity of cellular networks while party hubs have high average PCC with their interactors and are often contained in tightly organized modules (Han et al., 2004). We find that date hubs have a significant propensity to be

involved in transient interactions (Figures 5.3B and 5.3C) suggesting that these play an important role in maintain the integrity of the networks. To validate this result, we compared the edge "betweenness" of global and transient interactions. Edge betweenness can be used to detect community structure within networks (Girvan and Newman, 2002). Clusters detected by this approach tend to share similar functions (Dunn et al., 2005). We find that transient interactions for both human and yeast have a significantly higher betweenness than stable interactions (Figure 5.3D). This implies that transient interactions hold the key in maintaining the integrity of the underlying cellular network. Disrupting these will partition the interactome into disjoint clusters, unable to perform temporally and spatially well-regulated processes.

To further explore topological properties of transient interactions, we examined connectivity in response to progressive edge removal and found that selectively removing transient interactions increased characteristic path length much more sharply than selectively removing stable or random interactions (Figures 5.3E and 5.3F). Biological interactomes are small-world networks and removing a random edge is unlikely to significantly alter connectivity, as most random edges are not essential in maintaining network integrity (Albert et al., 2000). However, selectively disrupting key edges disrupts network structure and increases the characteristic path length significantly. Since removal of transient interactions causes the sharpest increase in path length, these are indeed critical for network integrity.

5.4 DISCUSSION

Here, we utilize a previously described Smith-Waterman-like dynamic programming algorithm to segregate transient interactions from stable complexes on a genomic scale directly from gene

expression data. For the first time, we distinguish their biological roles and show that though transient interactions are currently underexplored, they perform key biological functions and are essential to maintaining the integrity of cellular networks. Moreover, we find that Y2H is currently the only technology that is able to determine transient interactions on a large scale. Our findings are likely to generate significant interest in designing experiments to detect transient interactions to further explore their properties.

5.5 MATERIALS AND METHODS

Calculating PCC, MIC and LES

PCC was calculated in a massively parallel, Java program utilizing the Parallel Java framework (Kaminsky, 2010). MIC was calculated using a Java implementation provided by Reshef et al. (Reshef et al., 2011). Transient interactions for human and yeast were identified with a similar Parallel Java implementation of a Smith-Waterman-like dynamic programming algorithm to calculate LES (Qian et al., 2001).

Calculating betweenness and functional similarity

Edge betweenness was calculated using the Girvan-Newman algorithm (Girvan and Newman, 2002). Functional similarity was studied using total ancestry measure – a metric that takes the entire biological process tree and calculates the association of each gene with a biological process. For each protein pair query, it computes what fraction of all possible protein pairs that share the same set of Gene Ontology (GO) (Ashburner et al., 2000) biological pathway terms as the query pair (Yu et al., 2007a). The calculations are performed using a massively Parallel Java program (Kaminsky, 2010).

5.6 FIGURE LEGENDS

Figure 5.1. Cartoon depiction of protein–protein interaction dynamics. (A) Gene expression profiles for two proteins that are highly correlated under all conditions indicating a stable or globally co-expressed interaction. (B) Two contiguous blocks of significant co-expression indicate this pair of proteins is transiently interacting or locally co-expressed.

Figure 5.2. (A, B) Enrichment of PCC of co-expression of interacting proteins (detected by different technologies) as opposed to random gene pairs in human and yeast respectively. (C, D) Comparison of transient interactions detected per technology in human and yeast, respectively. The dashed line indicates the overall average detection of transient interactions.

Figure 5.3. (A) The expression profiles of SFB2 and SEC23 (co-expression only in the final yellow block). (B, C) Transient interactions in human are enriched in "date hubs". These have previously been shown to be vital in forming important topological links between stable functional modules. (D) Transient interactions in human and yeast have a significantly higher betweenness value–they hold the key in maintaining the integrity of cellular networks. (E, F) Characteristic path length as a measure of network connectivity after successive removal of edges of the network. Each data point represents the removal of a fixed percentage of overall nodes of the graph from each interaction type. Random removal occurs on all interactions in the network, which may include other interactions that are still uncategorized as transient or stable. Removal of transient interactions increases path length more sharply than disturbing random or stable interactions.

5.7 REFERENCES

Albert, R., Jeong, H., and Barabasi, A.L. (2000). Error and attack tolerance of complex networks. Nature *406*, 378-382.

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

Barabasi, A.L., and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509-512.

Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.R., Simonis, N., Rual, J.F., Borick, H., Braun, P., Dreze, M.*, et al.* (2009). Literature-curated protein interaction datasets. Nat Methods *6*, 39-46.

D'Eustachio, P. (2011). Reactome knowledgebase of human biological pathways and processes. Methods Mol Biol *694*, 49-61.

Dunn, R., Dudbridge, F., and Sanderson, C.M. (2005). The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics *6*, 39.

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

Ge, H., Liu, Z., Church, G.M., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet *29*, 482-486.

Girvan, M., and Newman, M.E. (2002). Community structure in social and biological networks. Proc Natl Acad Sci U S A *99*, 7821-7826.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proc Natl Acad Sci U S A *104*, 8685-8690.

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

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249-264.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. Genome Res *12*, 37-46.

Jensen, L.J., and Bork, P. (2008). Biochemistry. Not comparable, but complementary. Science *322*, 56-57.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L. (2000). The large-scale organization of metabolic networks. Nature *407*, 651-654.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

Kaminsky, A. (2010). Building parallel programs : SMPs, clusters, and Java (Boston, Mass.: Course Technology, Cengage Learning).

Kurihara, T., Hamamoto, S., Gimeno, R.E., Kaiser, C.A., Schekman, R., and Yoshihisa, T. (2000). Sec24p and Iss1p function interchangeably in transport vesicle formation from the endoplasmic reticulum in Saccharomyces cerevisiae. Mol Biol Cell *11*, 983-998.

Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tumer, Z., Pociot, F., Tommerup, N.*, et al.* (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol *25*, 309-316.

Luscombe, N.M., Royce, T.E., Bertone, P., Echols, N., Horak, C.E., Chang, J.T., Snyder, M., and Gerstein, M. (2003). ExpressYourself: A modular platform for processing and visualizing microarray data. Nucleic Acids Res *31*, 3477-3482.

Mewes, H.W., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K.F., Stumpflen, V.*, et al.* (2011). MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res *39*, D220-224.

Miller, E.A., Beilharz, T.H., Malkus, P.N., Lee, M.C., Hamamoto, S., Orci, L., and Schekman, R. (2003). Multiple cargo binding sites on the COPII subunit Sec24p ensure capture of diverse membrane proteins into transport vesicles. Cell *114*, 497-509.

Pawson, T., and Nash, P. (2000). Protein-protein interactions define specificity in signal transduction. Genes Dev *14*, 1027-1047.

Peng, R., De Antoni, A., and Gallwitz, D. (2000). Evidence for overlapping and distinct functions in protein transport of coat protein Sec24p family members. J Biol Chem *275*, 11521-11528.

Qian, J., Dolled-Filhart, M., Lin, J., Yu, H., and Gerstein, M. (2001). Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. J Mol Biol *314*, 1053-1066.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., and Sabeti, P.C. (2011). Detecting novel associations in large data sets. Science *334*, 1518-1524.

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

Suthram, S., Shlomi, T., Ruppin, E., Sharan, R., and Ideker, T. (2006). A direct comparison of protein interaction confidence assignment schemes. BMC Bioinformatics *7*, 360.

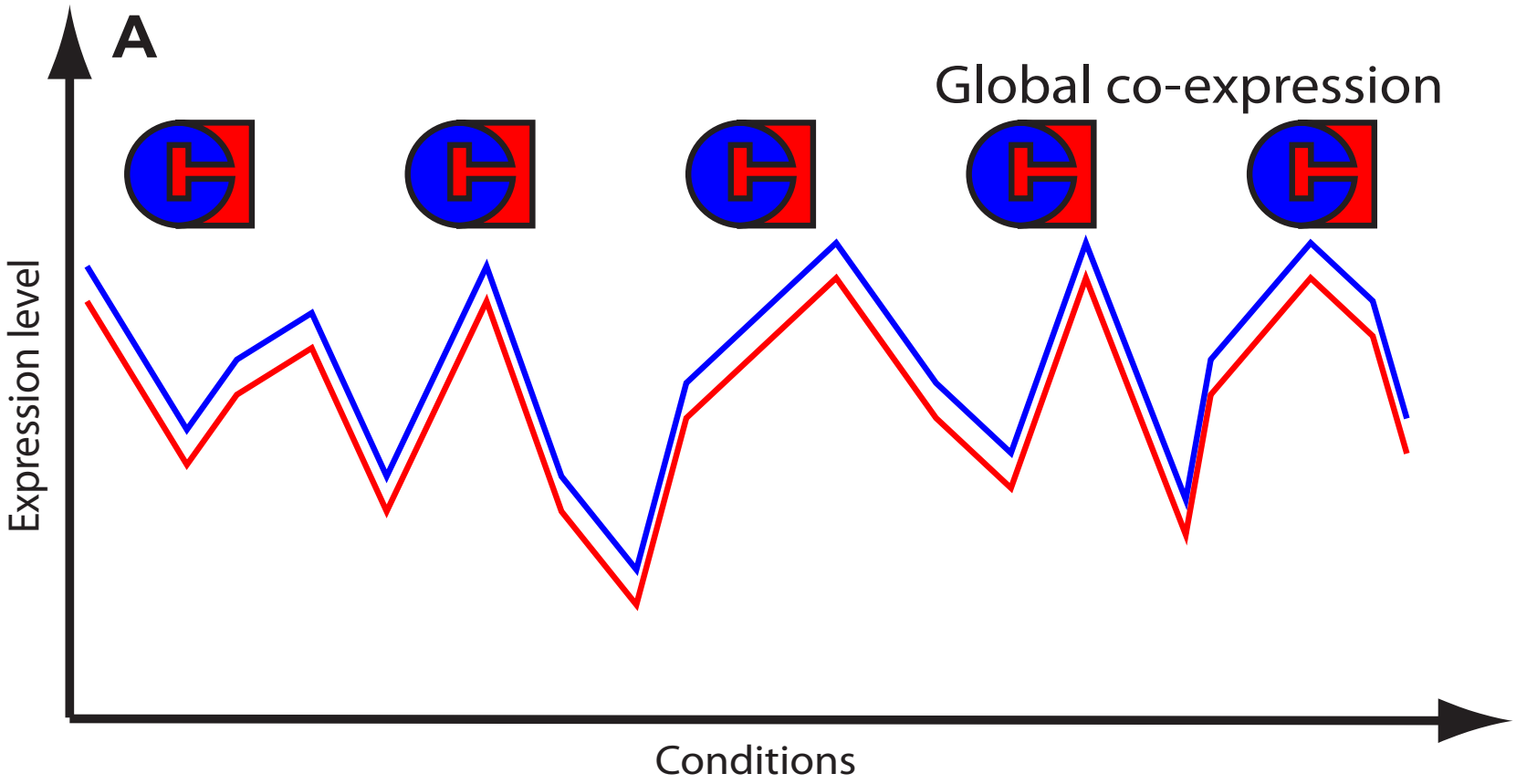
von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. Nature *417*, 399-403.

Wang, X., Gulbahce, N., and Yu, H. (2011). Network-based methods for human disease gene prediction. Brief Funct Genomics.

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

Yu, H., Jansen, R., Stolovitzky, G., and Gerstein, M. (2007a). Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. Bioinformatics *23*, 2163-2173.

Yu, H., Nguyen, K., Royce, T., Qian, J., Nelson, K., Snyder, M., and Gerstein, M. (2007b). Positional artifacts in microarrays: experimental verification and construction of COP, an automated detection tool. Nucleic Acids Res *35*, e8.


A
Global co-expression
Expression level
Conditions

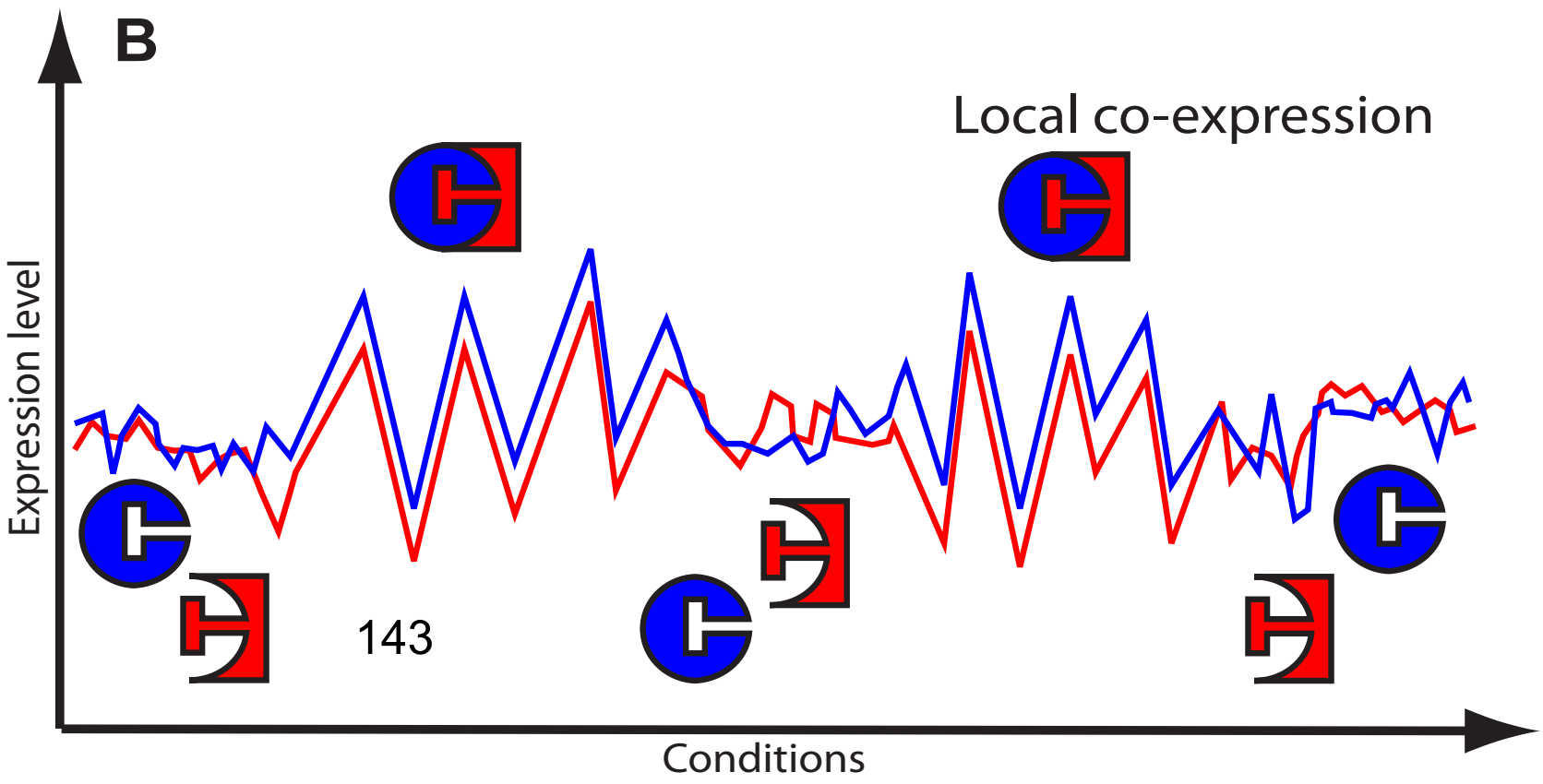

B
Local co-expression
Expression level
Conditions

Figure 5.1

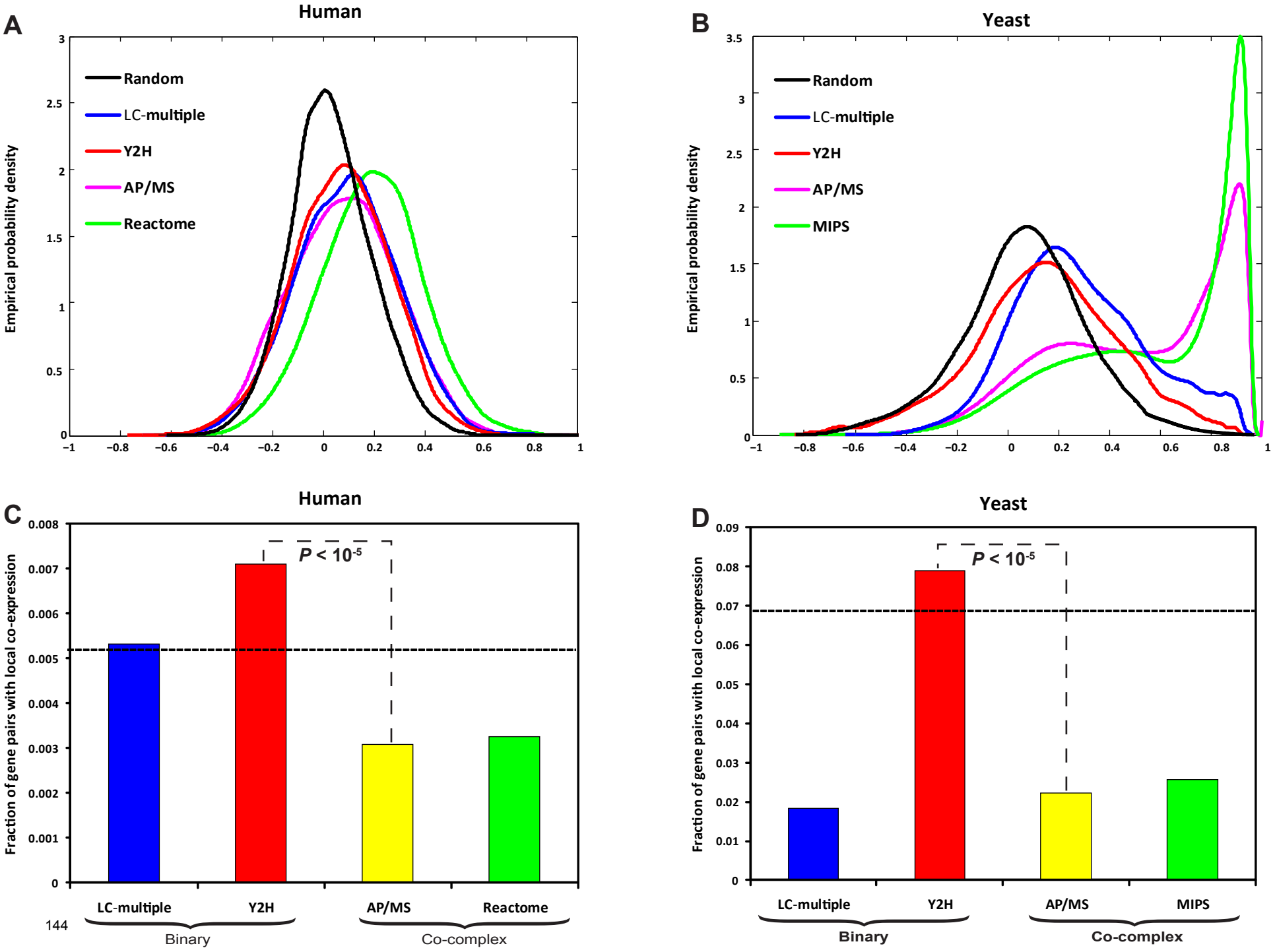

A
Human
Random
LC-multiple
Y2H
AP/MS
Reactome
Empirical probability density
B
Yeast
Random
LC-multiple
Y2H
AP/MS
MIPS
Empirical probability density
C
Human
P < 10-5
Fraction of gene pairs with local co-expression
LC-multiple
Y2H
AP/MS
Reactome
Binary
Co-complex
D
Yeast
P < 10-5
Fraction of gene pairs with local co-expression
LC-multiple
Y2H
AP/MS
MIPS
Binary
Co-complex

Figure 5.2

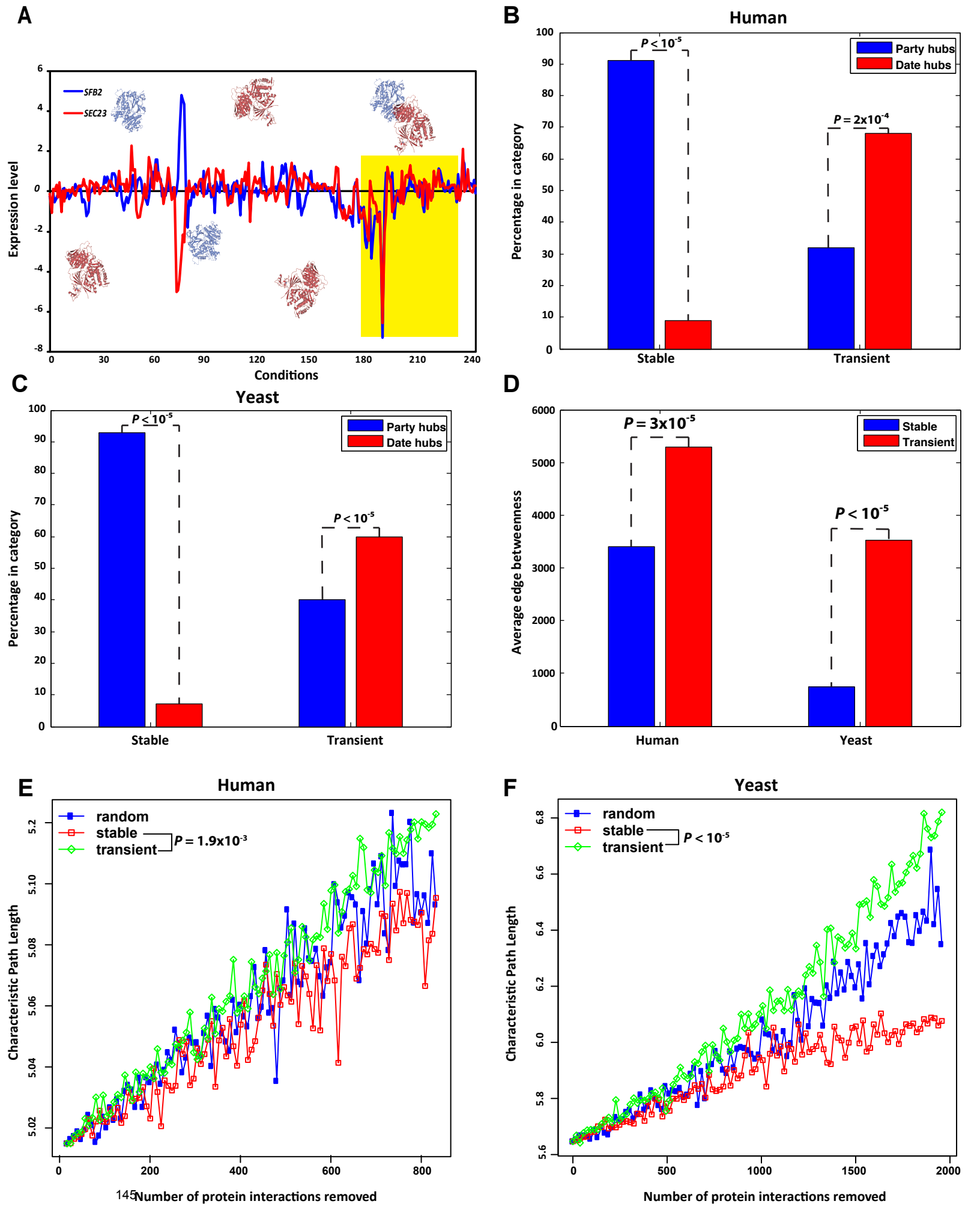

A
SFB2
SEC23
Expression level
Conditions
B
Human
Party hubs
Date hubs
P < 10-5
P = 2x10-4
Percentage in category
Stable
Transient
C
Yeast
Party hubs
Date hubs
P < 10-5
P < 10-5
Percentage in category
Stable
Transient
D
P = 3x10-5
P < 10-5
Stable
Transient
Average edge betweenness
Human
Yeast
E
Human
random
stable
transient
P = 1.9x10-3
Characteristic Path Length
Number of protein interactions removed
F
Yeast
random
stable
transient
P < 10-5
Characteristic Path Length
Number of protein interactions removed

Figure 5.3

CHAPTER 6

ENCAPP: elastic-net-based prognosis prediction and biomarker discovery for human cancers

In the following chapter, we explore how expression data can be combined with protein networks to predict cancer outcome. I am the first author of the paper resulting from this chapter (Das et al BMC Genomics 2015) and led all computational analyses. Kaitlyn Gayvert made a significant contribution to several of the analyses in the paper.

6.1 ABSTRACT

With the explosion of genomic data over the last decade, there has been a tremendous amount of effort to understand the molecular basis of cancer using informatics approaches. However, this has proven to be extremely difficult primarily because of the varied etiology and vast genetic heterogeneity of different cancers and even within the same cancer. Thus, one particularly challenging problem is to predict prognostic outcome of the disease for different patients. Here, we present ENCAPP, an elastic-net-based approach that combines the reference human protein interactome network with gene expression data to accurately predict prognosis for different human cancers. Our method identifies functional modules that are differentially expressed between patients with good and bad prognosis and uses these to fit a regression model that can be used to predict prognosis for breast, colon and ovarian cancers. Using this model, ENCAPP can also identify prognostic biomarkers with a high degree of confidence, which can be used to generate downstream mechanistic and therapeutic insights. ENCAPP is a robust method that can accurately predict prognostic outcome and identify biomarkers for different human cancers.

6.2 INTRODUCTION

The genetic complexity of cancer and its widely varying etiology and outcome make it extremely difficult to treat. It has been realized that rather than being a single disease, different cancers have widely diverse molecular bases (Hanahan and Weinberg, 2011; Lawrence et al., 2014). There has been a tremendous amount of effort in the literature to understand molecular signatures underlying cancer (Hanahan and Weinberg, 2011). A significant number of these efforts have been informatics-based approaches that try to leverage genomic information such as expression alterations, mutations in genomes, copy number changes and epigenetic modifications to elucidate the mechanistic basis of cancer (Chin et al., 2011). Global collaborative research endeavors such as The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and the International Cancer Genome Consortium (ICGC) (Hudson et al., 2010) are trying to assimilate these genome-scale datasets for different kinds of cancers across many countries.

One of the key challenges has been to use genomic information to understand the basis for different outcomes for the same cancer. However, this has been difficult because it is unclear as to which parameters contain the most information regarding disease outcome. One of the first attempts at predicting cancer prognosis using genome-scale transcriptomic datasets was undertaken by van de Vijver et al (van de Vijver et al., 2002). Using microarrays, they obtained tissue-specific gene-expression profiles for breast cancer patients. They then clustered these expression profiles and correlated them with prognostic outcome to identify a 70-gene 'prognosis profile' for breast cancer. One of the key limitations in using only expression datasets to predict cancer prognosis is the assumption of independence between genes in hypotheses testing. However, the protein products encoded by these genes are not independent but part of a complex interactome network. The dependencies of this network have been shown to be of great

importance in understanding the genetic and molecular bases of disease (Das et al., 2014a; Das et al., 2014b; Guo et al., 2013; Vidal et al., 2011; Wang et al., 2012). Chuang et al (Chuang et al., 2007), Taylor et al (Taylor et al., 2009) and Wu and Stein (Wu and Stein, 2012) used a functional interactome network to predict breast cancer prognosis. Recently, Hofree et al reported a network-based stratification approach that can use somatic mutations to predict cancer subtypes (Hofree et al., 2013). However, their method is primarily designed to work with mutation data and is less accurate for expression data (Hofree et al., 2013). Given the much wider availability of expression datasets as compared to whole genome or exome sequences, it is of paramount importance to have a robust method that can use gene expression to accurately predict prognosis across different types of cancer. To this end, in this manuscript, we report ENCAPP, an elastic-net-based cancer prognosis prediction method. We use tissue-specific gene expression data from patients along with the reference human protein interaction network to construct a regression model that can predict disease outcome for breast, colon and ovarian cancers. Our approach outperforms previous methods in terms of accuracy of prognosis prediction. Moreover, ENCAPP can also accurately identify genes that can serve as prognostic biomarkers for different cancers.

6.3 RESULTS

ENCAPP – a schematic

A reference high-quality human protein interactome was constructed as described earlier (Das and Yu, 2012). Our interactome comprises a total of 42,604 binary and co-complex interactions between 9,985 proteins. We include both kinds of interactions as they capture orthogonal layers of information – binary interactions represent direct contacts between two proteins, while co-complex associations capture co-membership of a protein complex. This network is clustered into different functional modules. We then overlay tissue-specific gene expression data from cancer patients onto these functional modules to generate 'expression modules'. We then identify ones that are differentially expressed between patients with good and bad prognosis (Figure 6.1). We use the expression modules that show the maximum difference between the prognostic outcome classes as decision boundaries to build a regression model that can predict disease prognosis (Figure 6.1). Our regression approach attempts to estimate the conditional probability of having good or bad prognosis given the patient's expression modules.

Since the data is inherently high dimensional (i.e., the number of expression modules is greater than the number of patients), ordinary least squares regression cannot be used and a regularization term is essential (see Methods). While ridge regression (L2 regularization term) (Hoerl and Kennard, 1970) uses all input variables to fit the model, the least absolute shrinkage and selection operator (LASSO, L1 regularization term) (Tibshirani, 1996) attempts to find the most optimal sparse fit. Ridge regression can lead to inflated variance but has low bias, while LASSO can have high bias but ensures low variance. To optimize the bias-variance tradeoff, the

elastic net (Bunea, 2008; Zou and Hastie, 2005) has been proposed and is our choice of regression model (see Methods).

Prognosis prediction using differentially expressed functional modules

We first examined expression data from a cohort of breast cancer patients (van de Vijver et al., 2002). Here, prognosis was defined as five-year disease-free survival. Using five-fold cross validation, we first measured prognosis prediction accuracy using only expression values from all genes and found it to be a suboptimal predictor (median AUC = 0.747, 95% CI for AUC = 0.743-0.751 Figure 6.2A, see Methods). Since proteins carry out their function by interacting with other proteins, we then used only expression values from genes whose corresponding proteins have at least one known interaction to predict prognosis. This did not significantly alter performance (median AUC = 0.745, Figure 6.2A). Taylor et al used hub groups as a measure of network topology, however we choose modules for two reasons (Figure 6.2B). First, hub groups only include interactions between the hub protein and its interactors, not those between the interactors themselves. Thus, modules contain more information. Second, in Taylor et al's model, each protein is assigned to one and only one hub group. However, since network modules can be overlapping (Ahn et al., 2010; Ravasz et al., 2002), the same protein may be assigned to multiple modules if it has multiple functions. Since numerous proteins carry out biological functions in a pleiotropic fashion, our approach captures such relationships while hub groups do not.

To identify functional modules, we tried three separate algorithms – hierarchical clustering (Ward, 1963), affinity propagation clustering (Frey and Dueck, 2007) and ClusterOne (Nepusz et al., 2012). We constructed modules from all three algorithms using default parameters (the

module creation is independent of any expression data). Using expression values from the van de Vijver dataset, we used modules generated by all three algorithms to construct expression modules and used them to predict prognosis. We find that ClusterOne has the best performance (Figure 6.2C; see Methods). One possible reason for this is that the protein interactome network is binary (1 corresponding to an interaction between two proteins, while 0 corresponds to no interaction between the two proteins) and sparse. Thus, the number of discrete values (equal to 1 + the graph diameter) the graph distance used for hierarchical clustering can take is limited. Affinity propagation clustering is more suited to identifying hub-group-like topological structures as hubs fit the definition of exemplars. On the other hand, ClusterOne was designed to identify functional modules that capture pleiotropic relationships. Thus, ClusterOne was used for all further analyses.

We then explored the contribution of the three different datasets – clinical covariates, gene expression and the protein network to predicting prognosis. Figure 6.3 presents a flowchart of our ENCAPP algorithm. We find that expression and network in combination are the most informative (Figure 6.4A, median AUC = 0.777; 95% CI for AUC = 0.773-0.780; $P < 10^{-3}$; Table 6.1) and the addition of clinical data only marginally improves the performance (Figure 6.4A, AUC = 0.786; 95% CI for AUC = 0.783-0.789; $P < 10^{-3}$; Table 6.1). ENCAPP also performs much better than an approach that just uses differential approach; we trained a generalized linear model with differentially expressed genes selected using the LIMMA package (Ritchie et al., 2015) and found that the median AUC is 0.685, significantly lower than ENCAPP ($P < 10^{-3}$). These results confirm that using interaction dynamics, a combination of gene expression data with the topological structure of the network, is a key predictor of prognosis. Our results also confirm that ENCAPP will work efficiently even in the absence of clinical

information, which can be hard to collect and thus is often unavailable. Furthermore, while we used 'death' as the outcome variable for the prognosis prediction described above, we find that it is robust to using other variables as outcome labels.

To compare the performance of our method to previous attempts, we first compared our classification accuracy (i.e., the fraction of patients for which we were able to accurately predict prognosis, see Methods) and AUC to Taylor et al (Figure 6.4B) and Chuang et al et al. Using expression data in conjunction with the protein network, ENCAPP achieves a median AUC of 0.777, significantly higher than the value of 0.71 reported by Taylor et al ($P < 10^{-3}$). We then compared the performance of ENCAPP to that of Chuang et al. At a fixed sensitivity of 90%, ENCAPP has a significantly higher accuracy (75.1% vs 70.1%, $P = 0.025$). Finally, we compared ENCAPP to the results reported by Wu and Stein (Wu and Stein, 2012). Since they do not directly report ROC curves, we adopted a slightly different approach for this comparison. We trained a generalized linear model (GLM) using expression values from Wang et al Lancet 2005 for the significant modules identified by them and attempted to predict prognosis for the Wang dataset. We found that the median AUC is 0.510. We then used the same modules and constructed the features that ENCAPP uses to train a GLM. The median AUC goes up to 0.561, significantly higher ($P < 10^{-3}$) than the earlier median AUC.

We then sought to assess the changes that cause the performance boost over previous methods. We used ENCAPP on an experimentally verified subset of the Ophid interactome used in the Taylor et al. study. We obtained a median AUC of 0.750, which is significantly higher ($P = 0.040$) than the AUC of 0.71 obtained by them. This confirms that a large portion of the increase in performance is solely due to the core methodology underlying ENCAPP – our approach captures more information regarding the topology of the protein interactome than Taylor et al

because of the differences in hub groups and modules outlined earlier. The rest of the increase is due to a higher quality protein network used in our study. The improvement in the protein network can be attributed to two factors – a methodological enhancement: we employ a series of stringent filtering steps (Das and Yu, 2012) to identify a set of high-quality interactions and an increase in the available data. Thus, ENCAPP is a robust and reliable method that combines expression data with protein network modules to accurately predict cancer prognosis; it works efficiently even in the absence of clinical data.

Robustness of ENCAPP

Is ENCAPP robust to changes of the response variable or the incompleteness of the reference protein network? To systematically test this, we first focused on how the performance of ENCAPP changes when the response variable is altered. For the van de Vijver dataset, he outcome variable (survival) is right censored, i.e., if a patient survives for >=5 years, she is considered to have good prognosis, else bad prognosis. To test the robustness of ENCAPP to the right censoring cutoff, we varied it from 3-14 years i.e, a patient is defined to have good prognosis if she survived for >=*k* years, where *k* varies from 3 to 14. We find that ENCAPP performs consistently well for all values of *k* (Figure 6.4C), with the highest median AUC being 0.778 and the lowest median AUC being 0.730. This confirms that ENCAPP is robust across a wide range of cutoff values for right censoring.

To further validate the robustness of ENCAPP to alternate definitions of prognosis, we modified the outcome definition. We defined a patient to have a good prognosis, if she does not have metastases for >= *k* years, where *k* varies from 3 to 10. Here too, ENCAPP performs consistently

well (Figure 6.4D), with the highest median AUC being 0.744 and the lowest median AUC being 0.652, confirming that it is also robust across prognosis definitions.

To address the robustness of ENCAPP to incompleteness of the protein network, we generated sets of 50 random networks for each of the following scenarios: 5%, 10%, 15% and 20% of the total edges randomly removed. We then generated modules for all these random networks using the same ClusterOne parameters as the original network. We then re-calculated the performance of ENCAPP on the van de Vijver dataset for each of these networks with a certain fraction of the edges removed. We find that ENCAPP still performs well, with median AUCs of 0.744, 0.740, 0.743 and 0.742 at 5%, 10%, 15% and 20% edge deletions respectively (Figure 6.4E), confirming that it highly robust to network incompleteness.

Pan-cancer prognosis prediction

A major challenge of prognosis prediction algorithms is to make them generically applicable to different human cancers. To examine the applicability of ENCAPP for other cancer types and sub-types, we first used it on a dataset of lymph-node negative breast cancer patients (Wang et al., 2005). Although, van de Vijver et al also examined breast cancer patients, the consensus gene signature identified was very different. Wang et al stated that the results vary so much "because of differences in patients, techniques, and materials used" (Wang et al., 2005). The van de Vijver dataset included node-negative and node-positive patients and women less than 53 years old. Moreover, prognosis for the Wang dataset is defined as metastasis-free survival. However, ENCAPP is still able to accurately predict (median AUC = 0.690; 95% CI for AUC = 0.684-0.695; $P < 10^{-3}$; Table 6.1) cancer prognosis for these patients (Figure 6.5A), confirming that its robustness across cancer sub-types.

Another key goal of prognosis prediction algorithms is to be applicable across data collected from different cohorts of patients. To test whether ENCAPP can be trained on a certain dataset and then used to predict outcome for a completely different set of patients, we used the Wang et al dataset to train the model and then predicted outcomes for the van de Vijver dataset using it. While we originally analyzed the van de Vijver dataset in terms of overall survival, clinical information on metastasis was available. Since, the Wang et al dataset uses metastasis-free survival as the prognostic outcome, we used this as the outcome for the cross-dataset prediction. ENCAPP was accurate in predicting outcomes (median AUC = 0.649; 95% CI for AUC = 0.649-0.650; P = 0.019; Table 6.1), showing that our approach is highly robust and successful in incorporating major differences in clinical parameters (Figure 6.5B). Here too, we perform better than Chuang et al who report a classification accuracy of 55.8% at 90% sensitivity (for predictions on the Wang dataset using the van de Vijver sub-network markers). ENCAPP achieves a significantly higher classification accuracy of 62.6% at 90% sensitivity (P = 0.009).

We then used ENCAPP to analyze other kinds of cancer – a colon cancer (Atlas, 2012) and an ovarian cancer (Atlas, 2011) expression dataset published by the TCGA. The ovarian cancer dataset that we analyzed consisted of platinum-resistant cancer patients, which occurs in approximately 25% of patients within 6 months of therapy. For each dataset, we looked to see how well our method could predict overall survival. ENCAPP was able to predict prognostic outcome successfully for both colon and ovarian cancer (median AUCs = 0.666 and 0.766 respectively; 95% CIs for AUC = 0.658-0.674 and 0.760-0.771 respectively; P = 0.001 and 0.097 respectively; Table 6.1) confirming that it works robustly across different cancers (Figures 6.5C, 6.5E).

Finally, we tried using ENCAPP to predict prognosis across cancer types when they are related.

We tried predicting rectal cancer prognosis (Atlas, 2012) having trained ENCAPP using colon cancer data (Atlas, 2012). ENCAPP is very successful (median AUC = 0.803; 95% CI for AUC = 0.782-0.823; Figure 6.5D; $P < 10^{-3}$; Table 6.1) at predicting rectal cancer prognosis showing that ENCAPP is able to predict prognosis across related cancers.

Identifying prognostic markers using ENCAPP

Since our elastic net approach is a combination of LASSO and ridge regression, the number of coefficients with significant regression coefficients is relatively low (Figure 6.6A, Table 6.2; see Methods). The modules whose corresponding coefficients are mathematically significant are termed 'significant modules'. To test the robustness of these 'significant' modules, we calculated the Spearman rank correlation coefficient of these significant modules across cross-validation runs and folds. We find that they are highly stable: 99.1% have a rank correlation coefficient >= 0.98. To see if these modules are also biologically significant, we examined the distribution of known cancer genes in these modules (see Methods). We found that these modules are significantly enriched for cancer genes (Figure 6.6B; $P < 0.01$ for all 4 datasets). The fact that the enrichment extends to the level of entire modules shows that the differences in expression patterns extend to the level of the modules themselves. This is conceptually consistent with previous findings that gene sets rather than genes themselves better explain dysregulation in cancer (Subramanian et al., 2005). Thus differential co-expression of these modules is a molecular determinant of different outcomes for different patients.

We also compared the average degree of proteins in these significant modules with that of cancer-associated proteins (11.2) and all proteins in the network (8.2). The average degree of proteins in significant modules is not, in general, skewed towards the average degree of cancer-

associated proteins. For the van de Vijver and Wang breast cancer datasets, the average degree of proteins in significant modules are 12.0 and 10.5 respectively, similar to the average degree of cancer-associated proteins. However, for the colon and ovarian cancer datasets, they are 8.3 and 8.8 respectively, similar to the overall average degree. These findings are also consistent with Figure 6B, which shows that the enrichment of cancer genes in significant modules for the 2002 and 2005 breast cancer datasets is higher than the enrichment for the colon and ovarian cancer datasets. This could be due to higher noise for the colon and ovarian cancer datasets or due to the list of cancer genes being incomplete with varying degrees of incompleteness for different tissue types.

To examine whether the significant modules that we find agree with what has been previously reported, we compared the significant modules that we obtained for the van de Vijver dataset with the significant modules that Wu and Stein (Wu and Stein, 2012) obtained for the same dataset. 29/85 (34.1%) of the modules are overlapping. Thus, ENCAPP does find a large number of signatures concordant with what has been reported earlier, but it also finds a significant number of potentially novel signatures. We then compared the significant modules that we obtained for the Wang dataset with the significant modules that Wu and Stein obtained for the van de Vijver dataset. This is a comparison both across methods and cancer sub-types. 25/268 (9.3%) of the modules are still overlapping, showing that there are a number of stable signatures across cancer sub-types. We also find that 3 significant modules for the van de Vijver dataset contain 13 proteins of which 5 have been previously implicated in cancer (Figure 6.6C) and 3 significant modules for the colon cancer dataset contain 9/21 known cancer genes (Figure 6.6C). A number of these genes are known to be good prognostic markers.

As a further validation, we examined prognostic biomarkers detected by ENCAPP that were unknown at the time of publication of the expression dataset, but have since been clinically validated. Conceptually, these correspond to novel biomarkers detected by ENCAPP. For example, we detected *NFKB2* and *BCL3* in a significant module for the breast cancer (2002) dataset (Figure 6.6C). In 2005, it was shown that the *NFκB* complex, of which *NFKB2* is one of the subunits, can be used a well-known prognostic marker for breast cancer (Zhou et al., 2005). More recently, it has also been shown that suppression of the *NFκB* co-factor *BCL3* correlates with poor prognosis as it inhibits apoptosis of mammary cells (Wakefield et al., 2008). *GATA2* was present in a significant module for the colon cancer (published in 2011) dataset (Figure 6C). In 2013, *GATA2* was shown to be a useful prognostic marker for colorectal cancer – patients with high expression levels of *GATA2* are likely to have worse disease-free survival outcomes than those with lower expression levels of *GATA2* (Chen et al., 2013). These confirm that the significant modules identified by ENCAPP contain numerous prognostic markers.

We also found a number of modules with proteins that have not yet been validated as prognostic biomarkers but are excellent candidates for hypothesis-driven follow-up experiments. For example, one of the significant modules for the breast cancer (2002) dataset contains *CKS1B*, *SKP2* and *DUSP1* (Figure 6.6D). It has been shown that *CKS1B* is required for the *SKP2*-mediated ubiquitination of *PSMD9* (*p27*) (Ganoth et al., 2001). A recent study shows that *PSMD9* expression is altered in breast cancer patients irrespective of the *BRCA* mutation state (Dressler et al., 2013). Together, these results suggest that this module and especially *CKS1B* and *SKP2* could be reliable prognostic markers across breast cancer subtypes as altered expression of these genes will lead to mis-regulation of *PSMD9*, whose expression is altered in breast cancer patients with or without mutations in *BRCA1*.

For the colon cancer dataset, one of the significant modules contains *FAM175B*, *BARD1, CSTF1*, *BRE*, and *UIMC1* (Figure 6.6D). It is well known that *BARD1* interacts with *BRCA1* to form a ubiquitin ligase complex (Brzovic et al., 2003; Hashizume et al., 2001) and the interaction can be disrupted by breast cancer mutations on *BRCA1* (Brzovic et al., 2003; Hashizume et al., 2001). A blood test based on *BARD1* has been proposed as a potential way to diagnose breast cancer (Irminger-Finger, 2010). *FAM175B* (*ABRO1*) and *BRE* are two of the 4 subunits of the BRISC deubiquitinating enzyme complex (Cooper et al., 2009). *BRE* has already been shown to be a reliable prognostic marker for acute myeloid leukemia (Noordermeer et al., 2012; Noordermeer et al., 2011). In the context of these studies, our results suggest that this module and especially *FAM175B*, *BARD1* and *BRE* can be potential prognostic markers for colon cancer as altered expression of these genes can modify ubiquitination activity in the cell.

6.4 DISCUSSION

Here we have described ENCAPP, a robust prognosis predictor of different human cancers. Since ENCAPP uses differentially expressed modules between patients with good and bad prognosis to accurately predict disease outcome, the decision boundaries used to make this prediction correspond to functional changes in the cell. This is potentially extremely useful in generating mechanistic hypotheses regarding cancer causation and progression that can then be experimentally tested. Conceptually, the ENCAPP algorithm uses interaction dynamics, a combination of gene expression data with the topological structure of the network, to predict prognosis. Previous studies have shown that interaction dynamics is also useful in understanding the organization and evolutionary modes of biological networks (Das et al., 2012; Das et al., 2013). Together, these suggest that approaches using interaction dynamics may be successful in elucidating the mechanistic basis of a wide range of biological phenomena, by combining two discrete layers of information – gene expression and protein networks.

Another key feature of ENCAPP is its ability to identify prognostic markers from the regression model itself. While some previous methods show examples of prognostically relevant genes identified by their method (Hofree et al., 2013; Taylor et al., 2009), the key difference is that such detections are typically anecdotal. On the other hand, we demonstrate that the significant modules in ENCAPP are systematically enriched for cancer genes. Thus, our model identifies biologically relevant genes and uses these for determining prognostic outcome. We also show that significant modules identified by ENCAPP contain known prognostic markers and hypothesize that they may contain novel biomarkers. Follow-up studies may want to validate these putative prognostic markers. Since ENCAPP identifies modules containing these genes,

any positive results emerging from such studies will directly tie in to a pathway-level understanding of the mechanistic basis of that specific cancer type.

One limitation of ENCAPP is that the accuracy of the prognosis prediction is highly dependent on the quality of the expression dataset, which is why the AUCs vary across the different cancers. Future approaches may want to combine gene expression and protein networks with other data such as somatic mutations, epigenetic modifications and copy number alterations to make the overall prediction accuracy less dependent on the quality of an individual dataset.

6.5 MATERIALS AND METHODS

Expression data and the human protein interactome network

Sample size, number of good and bad prognosis patients, and breakdown by stage and grade for the different expression datasets used are available in Table 6.3. Expression data were RMA-normalized. High-quality binary and co-complex human protein interactome networks were obtained from HINT (Das and Yu, 2012). The final network used for this study was the union of the binary and co-complex networks. It comprises 42,604 interactions between 9,985 proteins. All datasets used in this study are obtained from papers that have already been published and required no ethics approval.

Identifying functional modules using clustering

ClusterOne identifies overlapping functional modules based on the topological properties of the protein interactome network (Nepusz et al., 2012). We did a sweep for the '*s*' (size) and '*d*' (minimum cluster density) parameters in ClusterOne (Nepusz et al., 2012). The default parameters are $s = 3$ and $d = 0.35$. We examined the parameter space around these values. Since the modules were identified independently of the expression datasets, situations occasionally arose in which some modules had missing gene expression values. In these cases, a module was included only if at least 1/3 of the genes in that module had corresponding expression values. For each cancer type, we report the highest AUC value obtained in the parameter sweep.

The elastic-net-based regression model

The elastic net (Zou and Hastie, 2005) is a regularized regression model that uses a linear combination of the L1 penalty term from LASSO (Tibshirani, 1996) and the L2 penalty term from ridge regression (Hoerl and Kennard, 1970). The objective function is given by:

$$\min_{\beta_0,\beta}\left[\sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2 + \lambda P_\alpha(\beta)\right]$$

where, y_i corresponds to the prognostic outcome for the i^{th} patient (0 or 1 corresponding to good and bad prognosis respectively). x_i is a vector of a vector of features for the i^{th} patient (please see below for a detailed description of x_i). The β's are regression coefficients that we estimate. The tuning parameter λ is the weight of the regularization term and is chosen to minimize mean square error. The regularization term $P_\alpha(\beta)$ is given by:

$$P_\alpha(\beta) = \sum_{j=1}^{|x_i|}\left(\frac{1-\alpha}{2}\beta_j^2 + \alpha|\beta_j|\right)$$

Here, α is a number between 0 and 1 with α = 0 corresponding to ridge regression alone and α = 1 corresponding to LASSO alone. We choose the best α using cross validation.

For our first analysis (Figure 6.2a) that used only expression data, x_i is a vector of dimension n containing expression values for n genes for the i^{th} patient (The entire set of expression values for d patients will be a matrix of size d x n, where each row is the transpose of x_i.). For ENCAPP, x_i is a vector of dimension $2n$ containing expression values for n modules for the i^{th} patient. Each functional module m contributes 2 values – G_{im} and B_{im} to x_i:

$$P_{ik} = \frac{E_{ik} - < E_{kg} >}{\sigma_{kg}}$$

$$Q_{ik} = \frac{E_{ik} - < E_{kb} >}{\sigma_{kb}}$$

$$G_{im} = \frac{\sum_{k=1}^{s_m} P_{ik}}{s_m}$$

$$B_{im} = \frac{\sum_{k=1}^{s} Q_{ik}}{s}$$

Here, s_m is the number of genes in module m. E_{ik} corresponds to the expression value of the k^{th} gene for the i^{th} patient. $< E_{kg} >$ and $< E_{kb} >$ represent the average (mean) expression values of the k^{th} gene across all patients with good and bad prognosis respectively. σ_{kg} and σ_{kb} represent the standard deviation of the expression values of the k^{th} gene across all patients with good and bad prognosis respectively. For all the cross-validations, $< E_{kg} >$, $< E_{kb} >$, σ_{kg} and σ_{kb} are calculated using only the samples in the training set. However, while using G_{im} and B_{im} as features derived from every module generally gives the most optimum performance, we noticed that in certain cases it is possible to obtain a slight increase in performance by not averaging over each module. There all P_{ik} and Q_{ik} values are used as input. While training the ENCAPP classifier, it is necessary to check which of the two approaches performs better.

For the datasets where clinical information was also available, we incorporated it using a logistic regression model. Since the clinical data is not high dimensional, elastic net regression is not a suitable choice for it. The final predicted outcome was a weighted linear combination of the two outputs – one predicted by the elastic-net-based model (using expression and protein network data) and the other predicted by the logistic regression model. Thus, $Y_1 = f(X_1)$ and $Y_2 = g(X_2)$ and $Y = k \times Y_1 + (1 - k) \times Y_2$. Here, X_1 is the set of expression derived features, f the function

obtained from the elastic-net based classifier and Y_1 the corresponding outcome variable, X_2 the set of clinical features, f the function obtained from the logistic-regression based classifier and Y_2 the corresponding outcome variable. Y is the final outcome obtained by a linear combination of Y_1 and Y_2. An optimal value of k, the relative weight parameter is obtained by grid search.

Evaluating performance

The performance of our model is evaluated in a five-fold cross validation framework. We split the patients into five subsets such that four subsets are used for training and the fifth one is the test set. The prognostic outcomes for the training set were used to determine the regression coefficients. These coefficients were then used to predict outcomes for patients in the test set. We repeated this procedure five times so that each subset served as a test set. The predicted outcomes were compared to the actual outcomes using a receiver operating characteristic (ROC) curve (Hastie et al., 2009). The area under the ROC curve (AUC) and classification accuracy were used as measure of the quality of the prediction (Hastie et al., 2009). The cross validation is process was repeated 50 times with a set of random seeds. For all comparisons, each method was run with the same set of random seeds, which ensured that the cross-validation dataset splits were identical across methods. Thus, all observed differences are solely due to one method being superior to the other and not because of how the dataset was split into the 5 folds. *P*-values evaluating the significance of difference in performance between different methods (two sets of AUC values) were calculated using a Mann-Whitney *U* test.

Classification accuracy is measured at the optimum point on the ROC curve. This is usually the point where the slope of the curve (*S*) is given by:

$$S = \frac{c(P \mid N) - c(N \mid N)}{c(N \mid P) - c(P \mid P)} \times \frac{N}{P}$$

Here, $c(I|J)$ represents the cost of assigning class I to class J. Here, P = true positives + false negatives and N = true negatives + false positives are the total counts in the positive and negative classes, respectively. For our calculations, we chose c($P|P$) = c($N|N$) = 0. And c($N|P$)=c($P|N$). Substituting these values, we get,

$$S = \frac{N}{P}$$

Enrichment of cancer genes in modules with significant regression coefficients

To identify modules with significant regression coefficients, we examined the distribution of coefficients and chose the highest and lowest two percentile of coefficients as significant (Figure 6A). We then examined the genes in these modules and compared them to known cancer genes. A list of known cancer genes was obtained from the Cancer Gene Census (Futreal et al., 2004). This is a high-confidence list of manually curated cancer genes with orthogonal layers of evidence, including but not limited to mutation information from COSMIC (Forbes et al., 2011). The expected fraction of cancer genes identified by random is given by:

$$Ef_i = \frac{C_i}{T_i}$$

where C_i is the number of cancer genes and T_i the total number of genes in modules in the i^{th} expression dataset. The observed fraction of cancer genes in modules with significant regression coefficients is given by:

$$Of_i = \frac{X_i}{N_i}$$

where X_i is the actual number of cancer genes and N_i the total number of unique genes in these modules. Thus, the enrichment of cancer genes in modules with significant regression coefficients is given by:

$$En = \frac{Of_i}{Ef_i}$$

P-values were calculated using a cumulative binomial test.

6.6 FIGURE AND TABLE LEGENDS

Figure 6.1. Schematic of ENCAPP.

ENCAPP begins by overlaying tissue-specific gene expression data with the reference interactome network. Modules that have significant differential co-expression between patients with good and bad prognosis are used to build a regression model that can predict prognostic outcome.

Figure 6.2. Integrating gene-expression data with protein interactome networks

(A) Receiver operating characteristic (ROC) curves for prognosis prediction using expression data alone.

(B) Illustration of hub groups and networks modules.

(C) ROC curves comparing the performance of three module-detection algorithms – hierarchical clustering, affinity propagation clustering and ClusterOne.

Figure 3. Flowchart illustrating the different steps in ENCAPP.

The inputs to ENCAPP are RMA-normalized expression data and modules from a reference human protein interactome network. These are then combined into features that are input to an elastic-net based regression model. The performance of the model is evaluated using cross-validation.

Figure 4. Predicting breast cancer prognosis using differentially expressed functional modules

(A) ROC curves for prognosis prediction of patients in the breast cancer (2002) dataset using clinical data alone, expression data alone, expression data with the protein network and all 3 datasets together.

(B) Comparison of the performance of ENCAPP with Taylor et (values shown are those obtained in the absence of clinical information).

(C) Boxplots showing performance of ENCAPP at different right censoring cutoffs *k* used for determining prognostic outcome: for each boxplot, good prognosis is defined as survival for >=*k* years and bad as death within *k* years.

(D) Boxplots showing performance of ENCAPP at different right censoring cutoffs *k* used for determining prognostic outcome; here a different outcome definition is used: for each boxplot, good prognosis is defined as no metastasis for >=*k* years and bad as metastasis within *k* years.

(E) Boxplots showing performance of ENCAPP using random networks that have 5%, 10%, 15% and 20% of the total edges in the original network randomly removed.

Figure 5. Prognosis prediction for different cancer types and subtypes

(A) ROC curves for prognosis prediction of patients in the breast cancer (2005) dataset using expression data alone and expression data with the protein network.

(B) ROC curves for prognosis prediction of patients in the breast cancer (2002) dataset using data from the breast cancer (2005) dataset.

(C) ROC curves for prognosis prediction of patients in the colon cancer dataset using clinical data alone, expression data alone, expression data with the protein network and all 3 datasets together.

(D) ROC curves for prognosis prediction of patients in the rectal cancer dataset using data from the colon cancer dataset.

(E) ROC curves for prognosis prediction of patients in the ovarian cancer dataset using expression data alone and expression data with the protein network.

Figure 6. Prognostic biomarker discovery using ENCAPP.

(A) Distribution of regression coefficients for different human cancers. The red shaded area corresponds to the top 10 percentile. Significant modules are defined as those with coefficients in the red shaded area.

(B) Enrichment of known cancer genes in the significant modules for the breast cancer (2002), breast cancer (2005), colon cancer and ovarian cancer datasets.

(C) Examples of significant modules for the breast cancer (2002) and colon cancer datasets. Known cancer genes are depicted in red.

(D) Examples of novel biomarker prediction for the breast cancer (2002) and colon cancer datasets.

Table 6.1. Summary of AUCs and *p* values for the different datasets.

Table 6.2. Number of different modules identified by each clustering method and list of significant modules identified by ENCAPP for the breast cancer (2002), breast cancer (2005),

colon cancer and ovarian cancer datasets. All genes in a particular module are listed in a single row. Each module is listed in a separate row.

Table 6.3. Sample size, number of good and bad prognosis patients, breakdown by stage and grade for the different datasets..

6.7 REFERENCES

Ahn, Y.Y., Bagrow, J.P., and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. Nature *466*, 761-764.

Atlas, T.C.G. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Atlas, T.C.G. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330-337.

Brzovic, P.S., Keeffe, J.R., Nishikawa, H., Miyamoto, K., Fox, D., 3rd, Fukuda, M., Ohta, T., and Klevit, R. (2003). Binding and recognition in the assembly of an active BRCA1/BARD1 ubiquitin-ligase complex. Proc Natl Acad Sci U S A *100*, 5646-5651.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l(1) and l(1) + l(2) penalization. Electron J Stat *2*, 1153-1194.

Chen, L., Jiang, B., Wang, Z., Liu, M., Ma, Y., Yang, H., Xing, J., Zhang, C., Yao, Z., Zhang, N.*, et al.* (2013). Expression and prognostic significance of GATA-binding protein 2 in colorectal cancer. Medical oncology *30*, 498.

Chin, L., Hahn, W.C., Getz, G., and Meyerson, M. (2011). Making sense of cancer genomic data. Genes Dev *25*, 534-555.

Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. Mol Syst Biol *3*, 140.

Cooper, E.M., Cutcliffe, C., Kristiansen, T.Z., Pandey, A., Pickart, C.M., and Cohen, R.E. (2009). K63-specific deubiquitination by two JAMM/MPN+ complexes: BRISC-associated Brcc36 and proteasomal Poh1. EMBO J *28*, 621-631.

Das, J., Fragoza, R., Lee, H.R., Cordero, N.A., Guo, Y., Meyer, M.J., Vo, T.V., Wang, X., and Yu, H. (2014a). Exploring mechanisms of human disease through structurally resolved protein interactome networks. Molecular bioSystems *10*, 9-17.

Das, J., Lee, H.R., Sagar, A., Fragoza, R., Liang, J., Wei, X., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014b). Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks. Hum Mutat *35*, 585-593.

Das, J., Mohammed, J., and Yu, H. (2012). Genome-scale analysis of interaction dynamics reveals organization of biological networks. Bioinformatics *28*, 1873-1878.

Das, J., Vo, T.V., Wei, X., Mellor, J.C., Tong, V., Degatano, A.G., Wang, X., Wang, L., Cordero, N.A., Kruer-Zerhusen, N.*, et al.* (2013). Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. Sci Signal *6*, ra38.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

Dressler, A.C., Hudelist, G., Fink-Retter, A., Gschwantler-Kaulich, D., Pfeiler, G., Rosner, M., Hengstschlager, M., and Singer, C.F. (2013). Tuberin and p27 expression in breast cancer patients with or without BRCA germline mutations. Journal of cancer research and clinical oncology *139*, 1349-1355.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res *39*, D945-950.

Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972-976.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177-183.

Ganoth, D., Bornstein, G., Ko, T.K., Larsen, B., Tyers, M., Pagano, M., and Hershko, A. (2001). The cell-cycle regulatory protein Cks1 is required for SCF(Skp2)-mediated ubiquitinylation of p27. Nature cell biology *3*, 321-324.

Guo, Y., Wei, X., Das, J., Grimson, A., Lipkin, S.M., Clark, A.G., and Yu, H. (2013). Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle. Am J Hum Genet *93*, 78-89.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

Hashizume, R., Fukuda, M., Maeda, I., Nishikawa, H., Oyake, D., Yabuki, Y., Ogata, H., and Ohta, T. (2001). The RING heterodimer BRCA1-BARD1 is a ubiquitin ligase inactivated by a breast cancer-derived mutation. J Biol Chem *276*, 14537-14540.

Hastie, T., Tibshirani, R., and Friedman, J.H. (2009). The elements of statistical learning : data mining, inference, and prediction, 2nd edn (New York, NY: Springer).

Hoerl, A.E., and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics *12*, 55-67.

Hofree, M., Shen, J.P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. Nat Methods *10*, 1108-1115.

Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S.*, et al.* (2010). International network of cancer genome projects. Nature *464*, 993-998.

Irminger-Finger, I. (2010). BARD1, a possible biomarker for breast and ovarian cancer. Gynecologic oncology *117*, 211-215.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495-501.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods *9*, 471-472.

Noordermeer, S.M., Monteferrario, D., Sanders, M.A., Bullinger, L., Jansen, J.H., and van der Reijden, B.A. (2012). Improved classification of MLL-AF9-positive acute myeloid leukemia patients based on BRE and EVI1 expression. Blood *119*, 4335-4337.

Noordermeer, S.M., Sanders, M.A., Gilissen, C., Tonnissen, E., van der Heijden, A., Dohner, K., Bullinger, L., Jansen, J.H., Valk, P.J., and van der Reijden, B.A. (2011). High BRE expression predicts favorable outcome in adult acute myeloid leukemia, in particular among MLL-AF9-positive patients. Blood *118*, 5613-5621.

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

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J.L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol *27*, 199-204.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met *58*, 267-288.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J.*, et al.* (2002). A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med *347*, 1999-2009.

Vidal, M., Cusick, M.E., and Barabasi, A.L. (2011). Interactome networks and human disease. Cell *144*, 986-998.

Wakefield, A., Piggott, L., Croston, D., Jiang, W.G., and Clarkson, R. (2008). Suppression of the NF-κB cofactor Bcl3 inhibits mammary epithelial cell apoptosis and, in breast tumours, correlates with poor prognosis. Breast Cancer Research *10*, O4.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol *30*, 159-164.

Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J.*, et al.* (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet *365*, 671-679.

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association *58*, 236.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113-1120.

Wu, G., and Stein, L. (2012). A network module-based method for identifying cancer prognostic signatures. Genome Biol *13*, R112.

Zhou, Y., Eppenberger-Castori, S., Marx, C., Yau, C., Scott, G.K., Eppenberger, U., and Benz, C.C. (2005). Activation of nuclear factor-kappaB (NFkappaB) identifies a high-risk subset of hormone-dependent breast cancers. The international journal of biochemistry & cell biology *37*, 1130-1144.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *67*, 301-320.

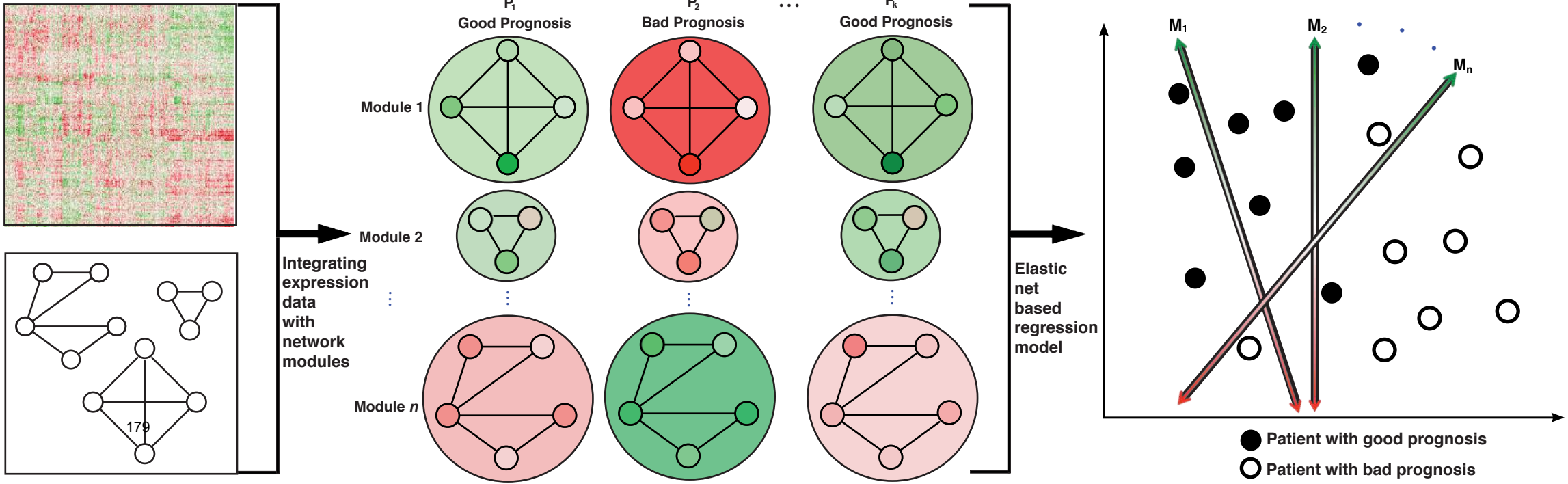

179
Integrating expression data with network modules
P1
Good Prognosis
P2
Bad Prognosis
Pk
Good Prognosis
Module 1
Module 2
Module n
Elastic net based regression model
M1
M2
Mn
Patient with good prognosis
Patient with bad prognosis

Figure 6.1

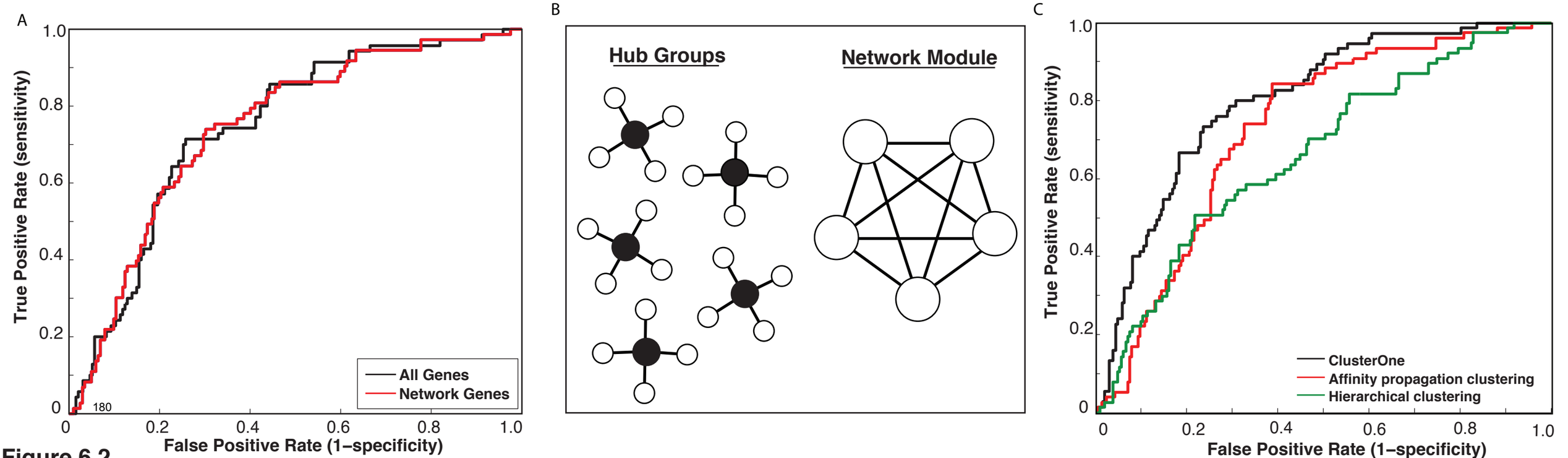
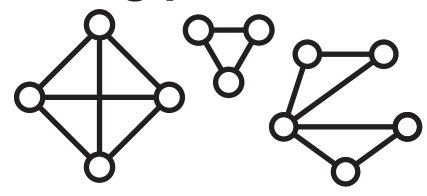

A
True Positive Rate (sensitivity)
False Positive Rate (1–specificity)
All Genes
Network Genes
180
B
Hub Groups
Network Module
C
True Positive Rate (sensitivity)
False Positive Rate (1–specificity)
ClusterOne
Affinity propagation clustering
Hierarchical clustering

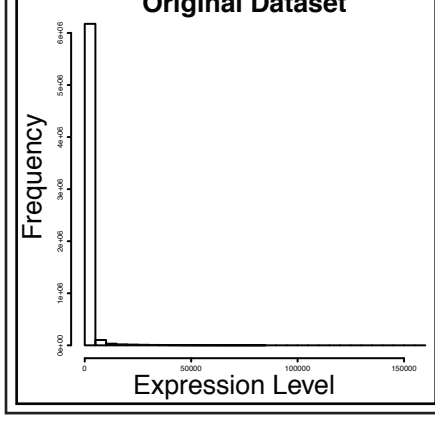
Figure 6.2

Input Expression Data

Modules obtained by clustering protein network



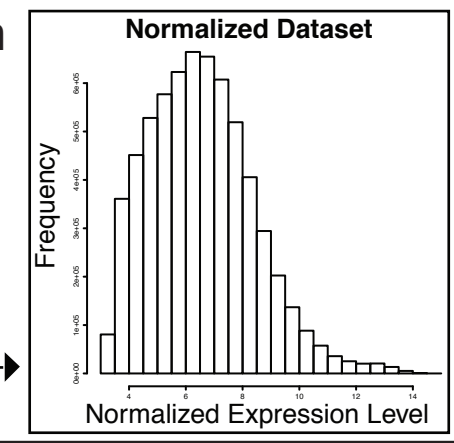
Original Dataset



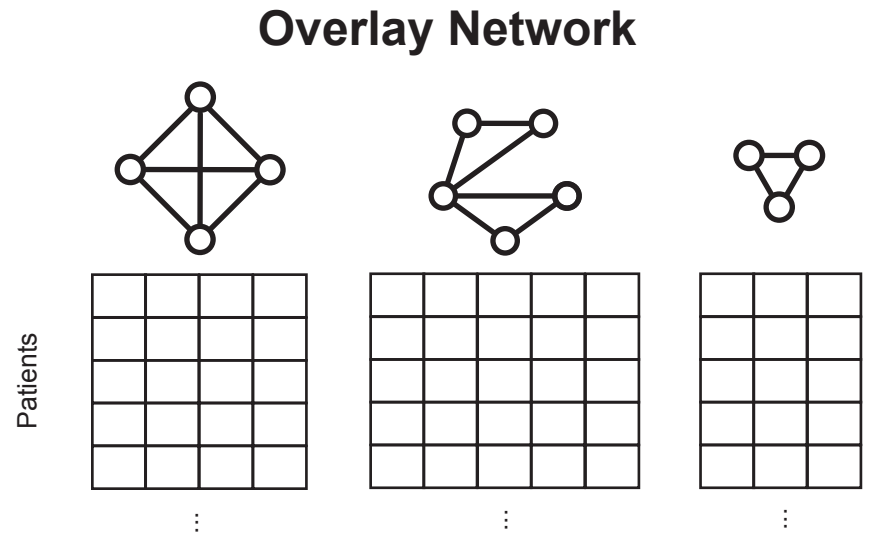
RMA Normalization

1. Background Correction
2. Log2 transformation
3. Quantile normalization
4. Summarization

Normalized Dataset

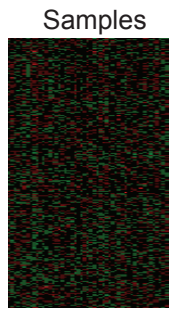


Overlay Network



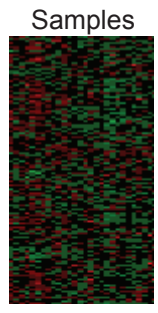
Training Set

Good Outcome Samples



$<E_{kg}>$ and σ_{kg}

Bad Outcome Samples



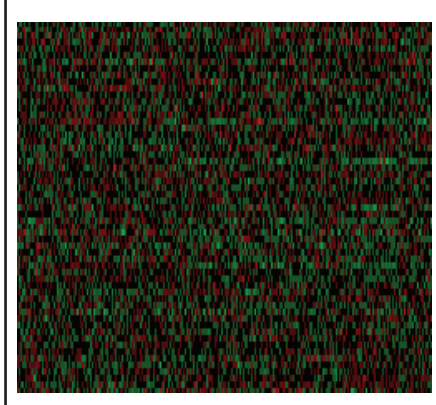
$<E_{kg}>$ and σ_{kb}

For each gene *k*, calculate:

$<E_{kx}>$ = average expression values (E_{ik}) of gene *k* across all patients with prognosis *x*

σ_{kx} = standard deviation of the expression values (E_{ik}) of gene *k* across all patients of prognosis *x*

Test Set



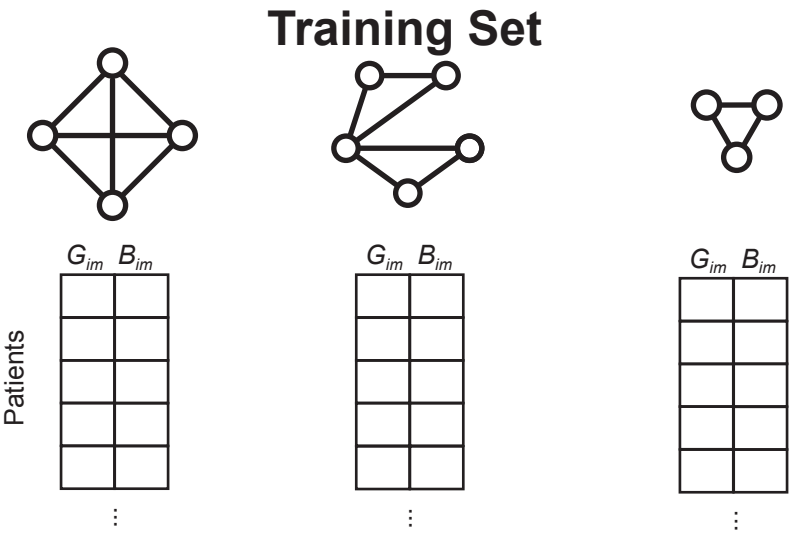
Feature Construction

For each module m and patient *i*, calculate:

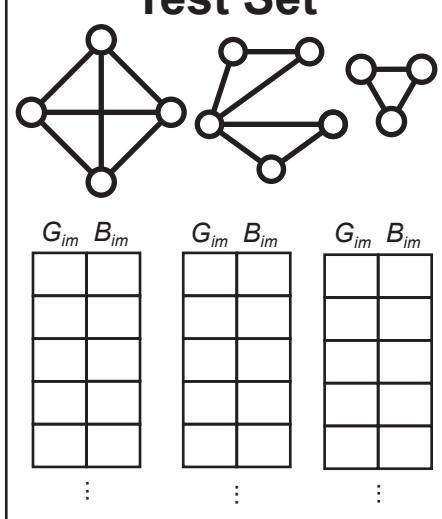
$$G_{im} = \frac{\sum_{k=1}^{s} \frac{E_{ik} - < E_{kg} >}{\sigma_{kg}}}{s} \qquad B_{im} = \frac{\sum_{k=1}^{s} \frac{E_{ik} - < E_{kb} >}{\sigma_{kb}}}{s}$$

where E_{ki} is the expression levels for gene k and patient i, s is the number of genes in module, and σ_{kg} ,σ_{kb} ,$<E_{kg}>$,$<E_{kb}>$ are as defined above.

Training Set



Test Set



The Elastic-Net Based Regression Model

$$\min_{\beta_0,\beta} \left[\sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \beta \right)^2 + \lambda P_\alpha(\beta) \right]$$

where $$P_\alpha(\beta) = \sum_{j=1}^{|x_i|} \left(\frac{1-\alpha}{2} \beta_j^2 + \alpha \mid \beta_j \mid \right)$$

- y_i = prognostic outcome for the *i*th patient (0 or 1 corresponding to good and bad prognosis respectively)
- x_i = vector of features for the *i*th patient
- β's = estimated regression coefficients
- λ = weight of the regularization term (chosen to minimize mean square error)
- α is a number between 0 and 1 with α = 0 corresponding to ridge regression alone and α = 1 corresponding to LASSO alone. (The best α is chosen using cross validation.)

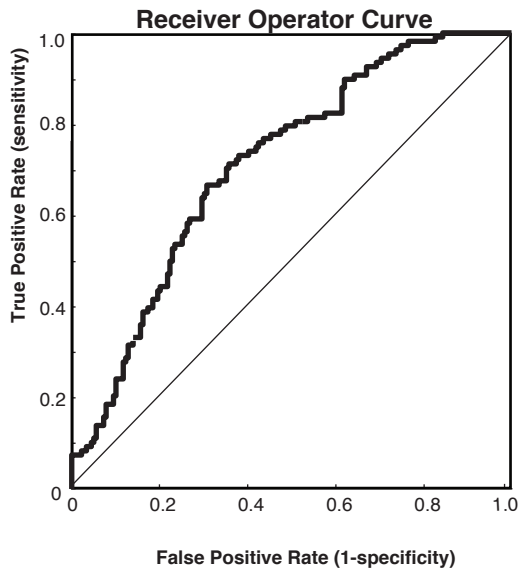
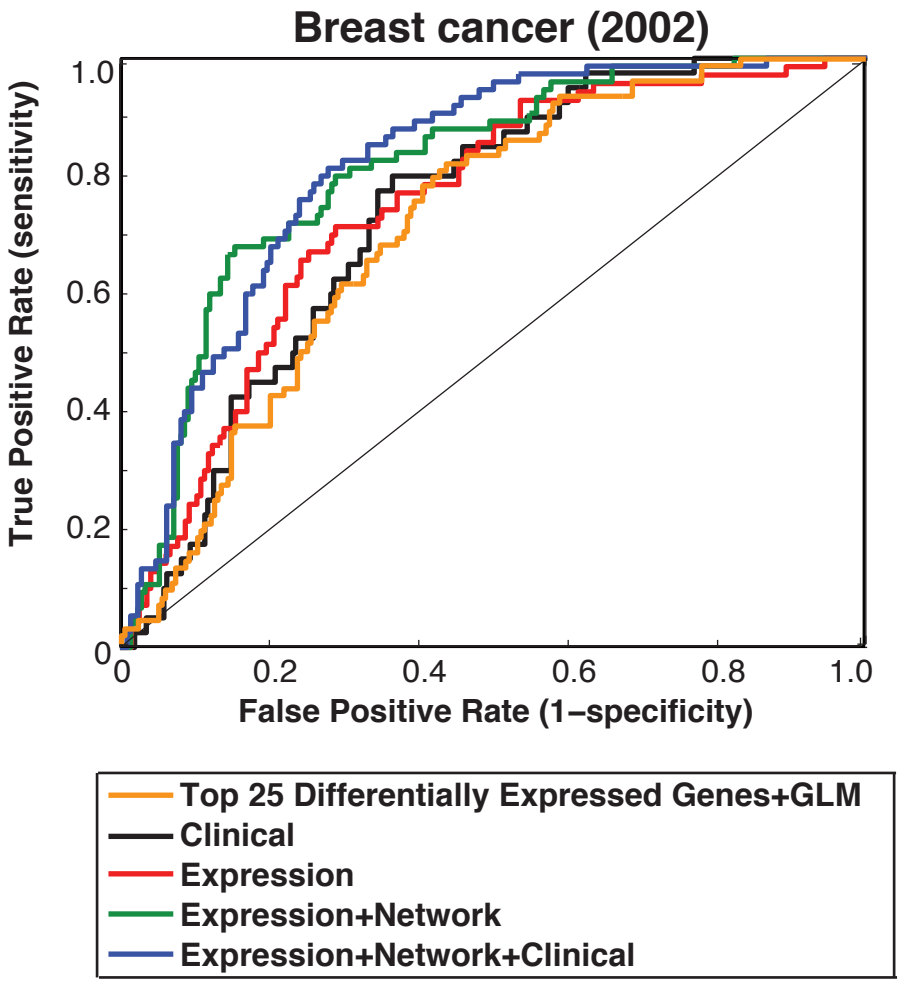
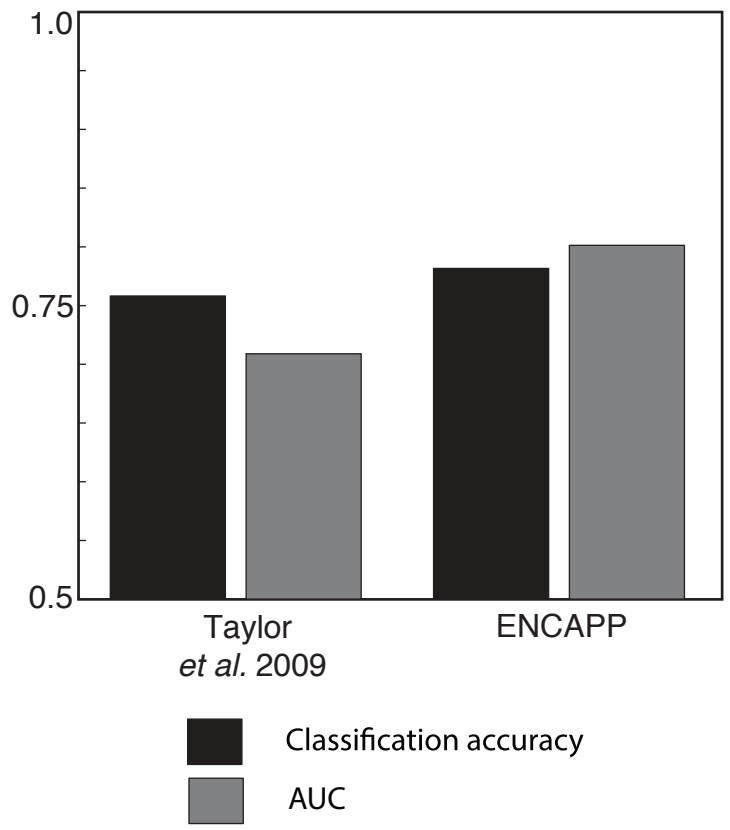


Figure 6.3

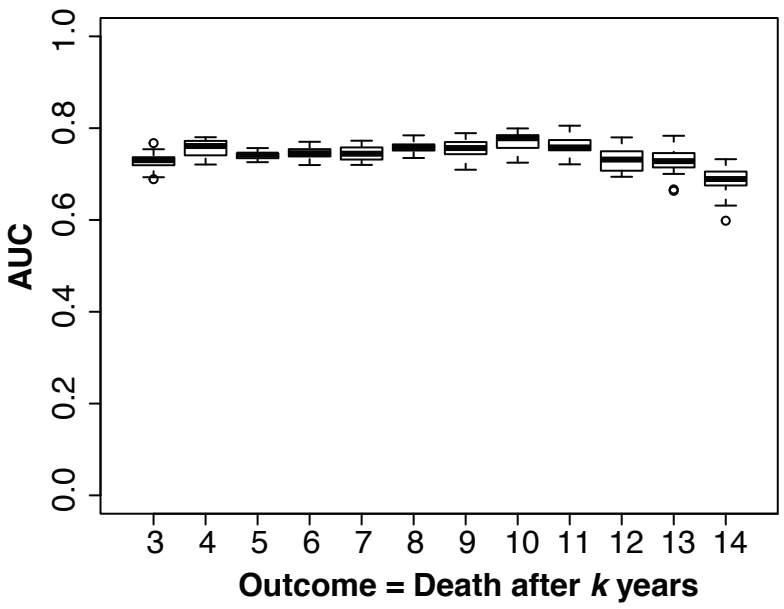
A



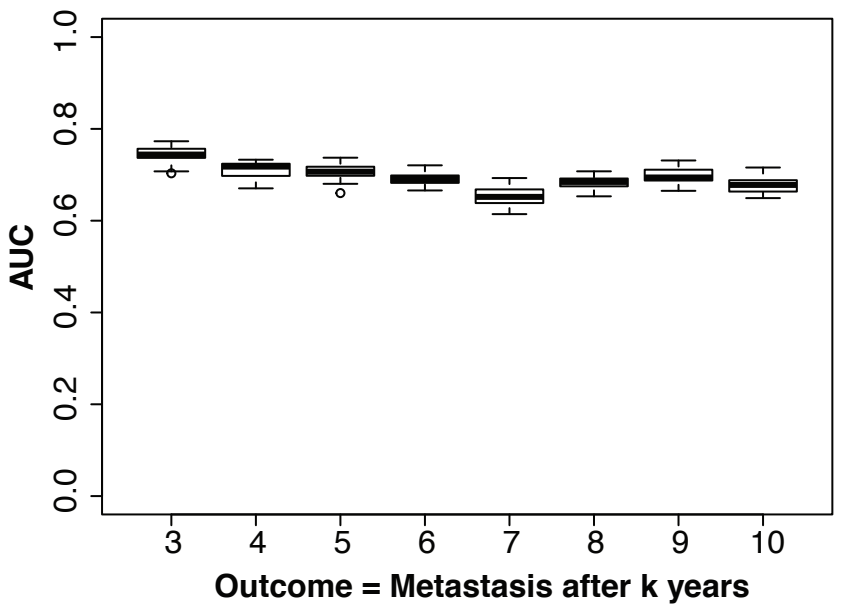
B



C



D



E

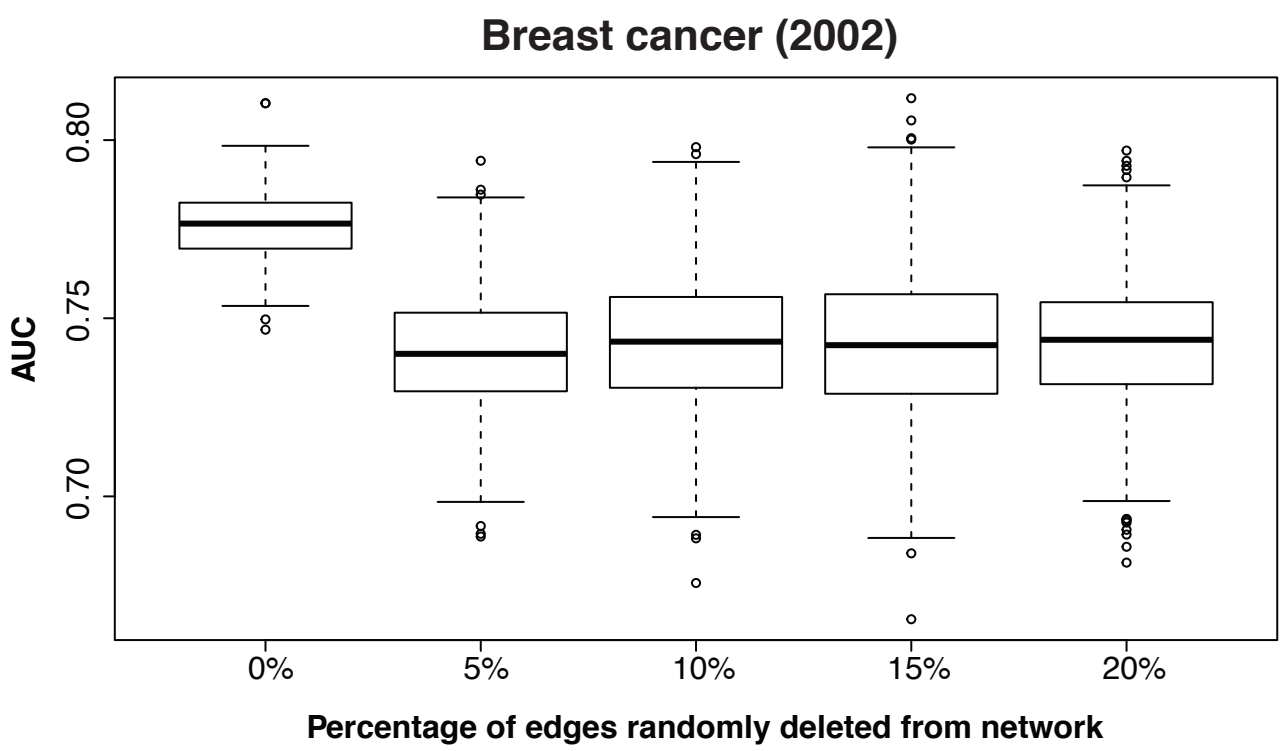
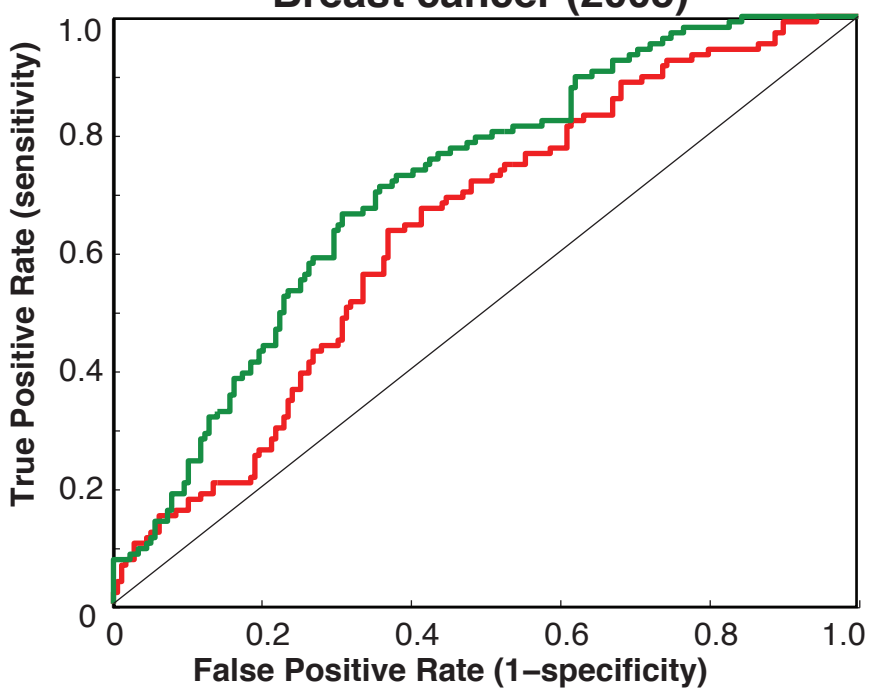
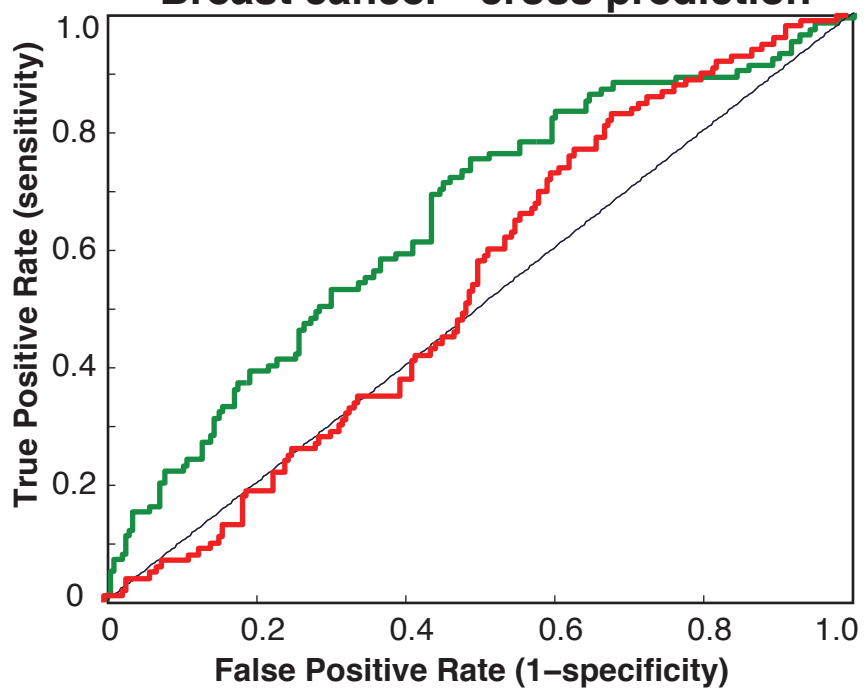


Figure 6.4

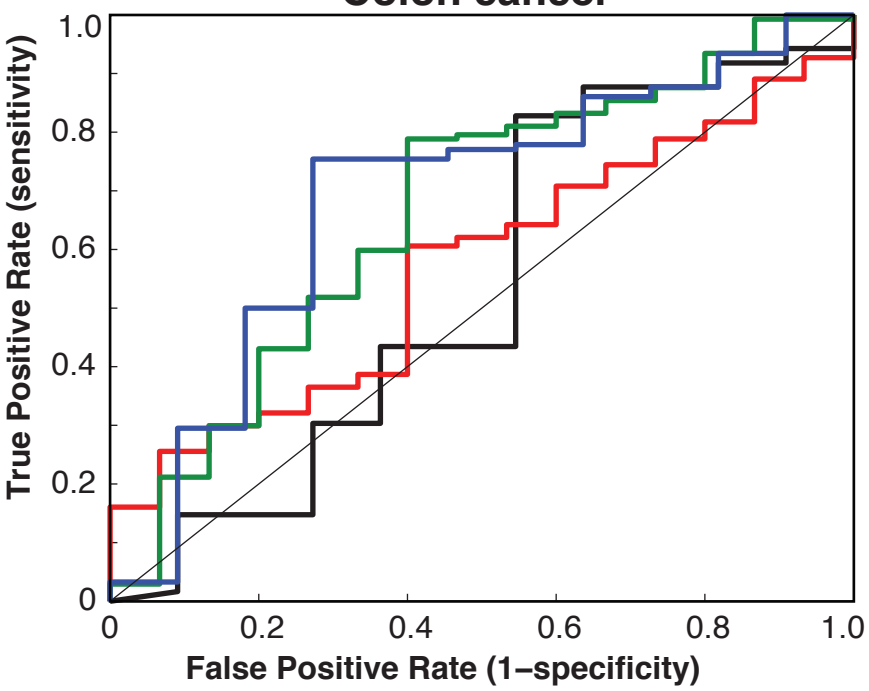
A



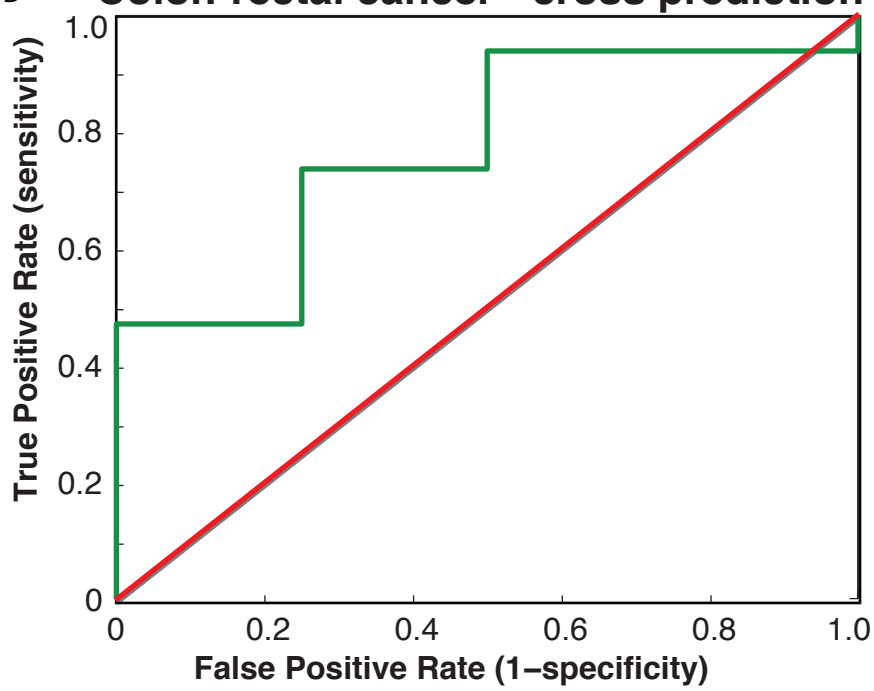
B



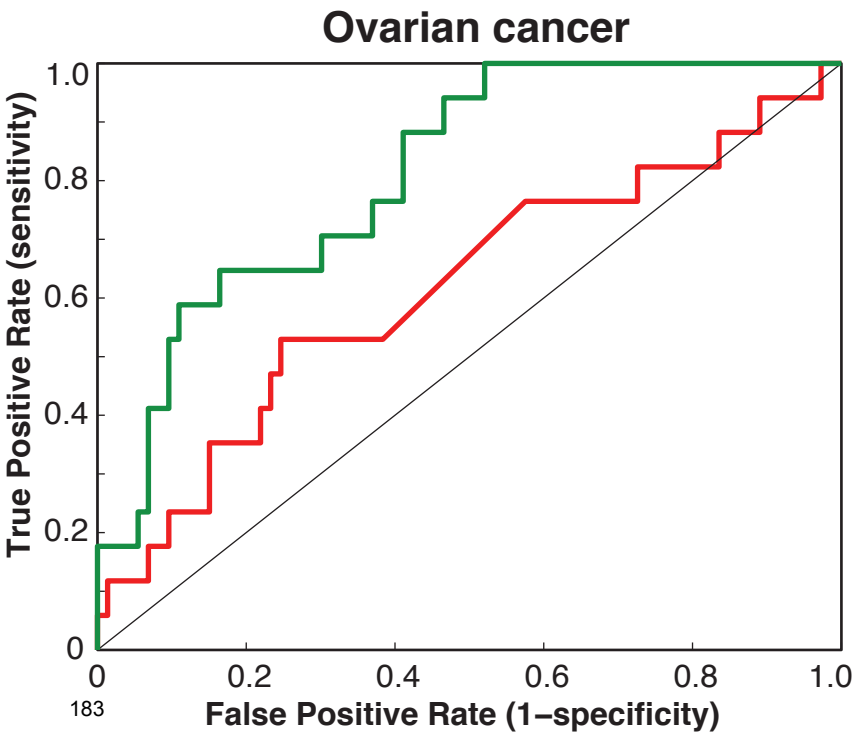
C



D



E



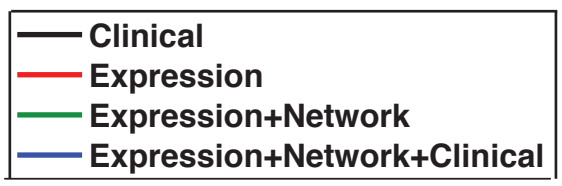


Figure 6.5

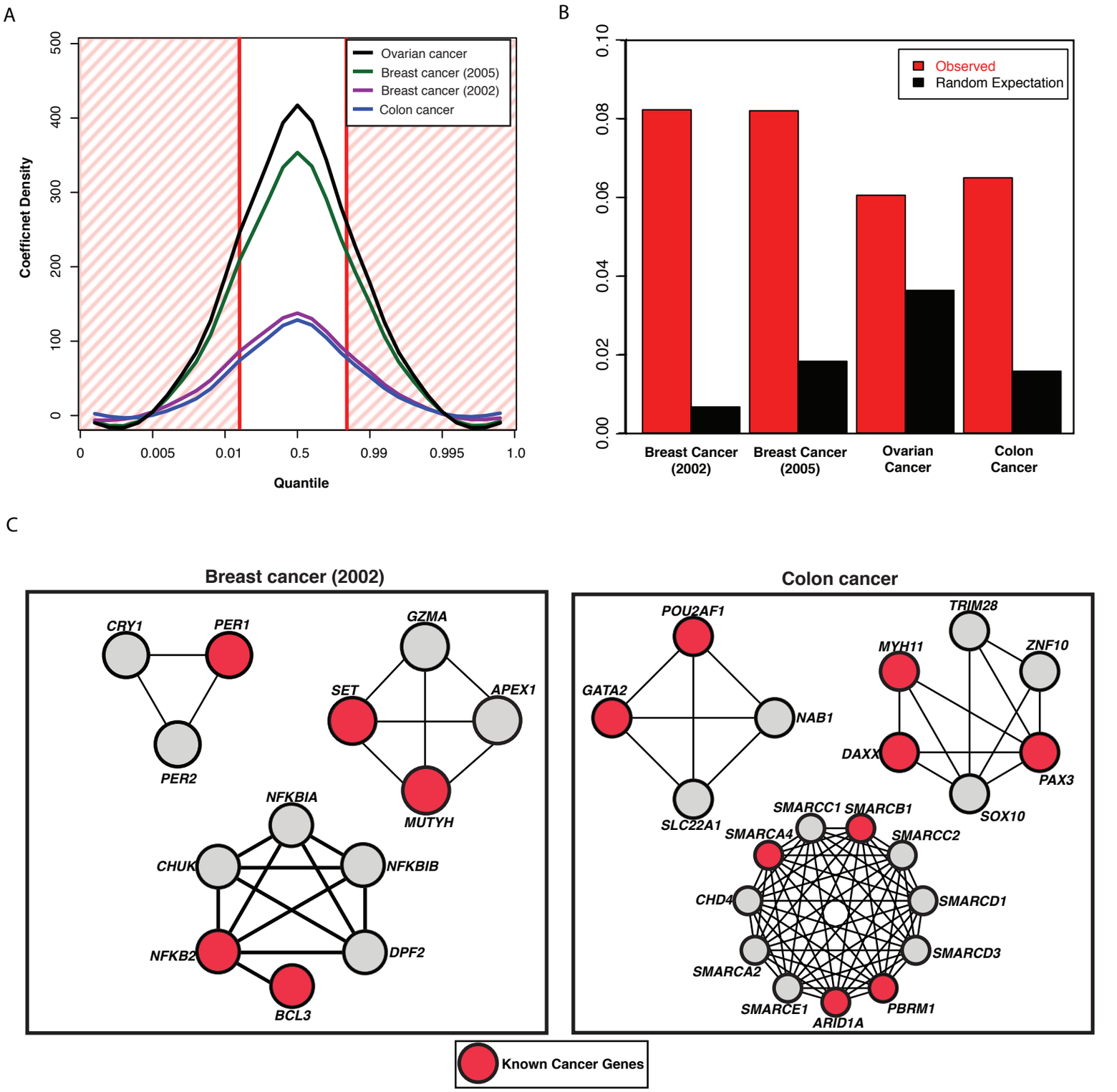

A
Coefficnet Density
Quantile
Ovarian cancer
Breast cancer (2005)
Breast cancer (2002)
Colon cancer
B
Observed
Random Expectation
Breast Cancer (2002)
Breast Cancer (2005)
Ovarian Cancer
Colon Cancer
C
Breast cancer (2002)
CRY1
PER1
PER2
GZMA
SET
APEX1
MUTYH
NFKBIA
CHUK
NFKBIB
NFKB2
DPF2
BCL3
Colon cancer
POU2AF1
GATA2
NAB1
SLC22A1
TRIM28
MYH11
ZNF10
DAXX
PAX3
SOX10
SMARCC1
SMARCB1
SMARCA4
SMARCC2
CHD4
SMARCD1
SMARCD3
SMARCA2
SMARCE1
PBRM1
ARID1A
Known Cancer Genes
D
Breast cancer (2002)
CKS1B
SKP2
DUSP1
Colon cancer
FAM175B
BARD1
UIMC1
CSTF1
BRE

Figure 6.6

Table 6.1	Without clinical data		With clinical data (when available)	
Dataset	**median AUC**	***p*-value (expression vs expression + network)**	**median AUC**	***p*-value (clinical vs clinical + expression + network)**
2002BreastCancer	0.778	< 0.001	0.786	< 0.001
2005BreastCancer	0.690	< 0.001	-	-
Cross-prediction (trained using 2005BreastCancer, tested on 2002BreastCancer)	0.649	0.019	-	-
ColonCancer	0.666	0.053	0.649	< 0.001
Cross-prediction (trained using clon cancer, tested on rectal cancer)	0.803	< 0.001	-	-
OvarianCancer	0.766	0.097	-	-

Table 6.2

2002BreastCancer

	Total number of modules	Modules after filtering out those where the majority of expression values are not available
Affinity propagation clustering	151	57
Hierarchical clustering	200	120
ClusterOne	350	234

2005BreastCancer

	Total number of modules	Modules after filtering out those where the majority of expression values are not available
ClusterOne	584	63

ColonCancer

	Total number of modules	Modules after filtering out those where the majority of expression values are not available
ClusterOne	584	104

OvarianCancer

	Total number of modules	Modules after filtering out those where the majority of expression values are not available
ClusterOne	2221	241

Note: The total number of modules also varies by dataset (for ClusterOne) as we have reported the number of modules corresponding to optimal performance. We did a parameter sweep around the default recommended parameters for ClusterOne and found that the *s* and *d* parameters need to be varied in the neighbourhood of the default recommended parameters for optimal performance.

Table 6.3

Dataset	Sample Size	Positive Outcome (=0) #patients		Negative Outcome (=1) #patients	
2002Breast Cancer	293	215	**GRADE:** **I:** 61 **II:** 76 **III**: 79	78	**GRADE:** **I:** 14 **II:** 25 **III**: 39
2005Breast Cancer	286	179	*Grade/Stage Not Available*	107	*Grade/Stage Not Available*
ColonCancer	152	137	**STAGE:** **I:** 26 **IIA:** 50 **IIB:** 4 **IIIB:** 18 **IIIC:** 13 **IV:** 19 **Unknown:** 2	15	**STAGE:** **I:** 2 **IIA:** 5 **IIB:** 1 **IIIB:** 1 **IIIC:** 3 **IV:** 3
OvarianCancer	90	17	**STAGE:** **IIC**: 0 **IIIB**: 0 **IIIC**: 14 **IV**: 3	73	**STAGE:** **IIC**: 1 **IIIB**: 1 **IIIC**: 61 **IV**: 10
			GRADE: **G2:** 2 **G3:** 15		**GRADE:** **G2:** 6 **G3:** 66 **Unknown:** 1

CHAPTER 7

Cross-Species Protein Interactome Mapping Reveals Species-Specific Wiring of Stress-Response Pathways

In the following chapter, we explore concepts of network evolution and how stress-response pathways have evolved across different yeast species. I am the first author of the paper resulting from this chapter (Das et al Science Signaling 2013) and performed all computational analyses. Experimental analyses described in the chapter were led by graduate student Tommy Vo.

7.1 ABSTRACT

The fission yeast *Schizosaccharomyces pombe* has more metazoan-like features than the budding yeast *Saccharomyces cerevisiae*, yet it has similarly facile genetics. Here, we present a large-scale verified binary protein-protein interactome network, "StressNet", based on high-throughput yeast two-hybrid screens of interacting proteins classified as part of stress-response and signal transduction pathways in *S. pombe*. We performed systematic, cross-species interactome mapping using StressNet and a protein interactome network of orthologous proteins in *S. cerevisiae*. With cross-species comparative network studies, we detected a previously unidentified component (Snr1) of the *S. pombe* mitogen-activated protein kinase Sty1 pathway. Coimmunoprecipitation experiments showed that Snr1 interacted with Sty1 and that deletion of *snr1* increased the sensitivity of *S. pombe* cells to stress. Comparison of StressNet with the interactome network of orthologous proteins in *S. cerevisiae* showed that the majority of interactions among these stress-response and signaling proteins are not conserved between species, but are "rewired;" orthologous proteins have different binding partners in both species. In particular, transient interactions connecting proteins in different functional modules were more likely to be rewired than conserved. By directly testing interactions between proteins in one yeast species and their corresponding binding partners in the other yeast species with yeast two-hybrid assays, we found that about half of the interactions traditionally considered "conserved" form modified interaction interfaces that may potentially accommodate novel functions.

7.2 INTRODUCTION

A crucial step towards understanding properties of cellular systems is to map networks of DNA-protein, RNA-protein, and protein-protein interactions, or the "interactome network," of an organism. Over the last decade, large-scale binary protein-protein interactome datasets have been produced for several eukaryotes – *Saccharomyces cerevisiae* (Ito et al., 2001; Uetz et al., 2000; Yu et al., 2008), *Drosophila melanogaster* (Formstecher et al., 2005; Giot et al., 2003), *Caenorhabditis elegans* (Li et al., 2004; Simonis et al., 2009), *Arabidopsis thaliana* (Consortium, 2011), and human (Rual et al., 2005; Stelzl et al., 2005), among which we produced a high-quality whole-proteome interactome network in *S. cerevisiae* using a high-throughput yeast two-hybrid (HT-Y2H) system (Yu et al., 2008). However, due to large evolutionary distances among these species [the last common ancestor of fungi and human is over 1 billion years ago (Sipiczki, 2000; Wood et al., 2002)] and extremely low coverage (most protein interactions are yet to be detected) of available interactome maps outside of *S. cerevisiae*, the overlap among these networks is sparse (Gandhi et al., 2006). This makes it difficult to extract meaningful information about evolutionary relationships from these interactomes. Thus, to bridge this gap, it is essential to construct a high-coverage interactome network for an intermediate species. The fission yeast, *Schizosaccharomyces pombe*, has an easily manipulatable genome and is estimated to have diverged from the budding yeast, *S. cerevisiae*, approximately 400 million years ago (Sipiczki, 2000; Wood et al., 2002). Furthermore, fission yeast is more similar to metazoans than is budding yeast, especially in its gene regulation by chromatin modification and RNA interference, mechanisms that are differently regulated and absent, respectively, in budding yeast

(Roguev et al., 2008). A high-quality map of the protein-protein interactome network of *S. pombe* will enable analysis of biological properties of many complex pathways common in metazoan species but missing in *S. cerevisiae* (Shevchenko et al., 2008).

The two yeasts live in highly disparate ecological niches and have varied mechanisms of responding to external stimuli. Therefore, in this study, we focus on 658 *S. pombe* genes involved in key regulatory processes of stress response and cellular signaling. Because these pathways control how organisms sense and adapt to their immediate environments, they are likely to have diverged between the two species. Using our HT-Y2H pipeline (Yu et al., 2008), we obtained a binary interactome network among these 658 genes, which we named "StressNet". All interactions were verified with two orthogonal assays to ensure their quality. By comparing with their *S. cerevisiae* counterparts, we measured the conservation rate of these StressNet interactions between fission and budding yeasts using a Bayesian method. We found species-specific wiring of stress-response and signaling pathways beyond what was expected by sequence orthology, indicating that rewiring of protein interactome networks in related species is likely to be a major factor for divergence. We also identified a previously unknown component Snr1 of the Sty1 mitogen-activated protein kinase (MAPK) pathway and experimentally validated that Snr1 has gained functions, through rewiring of its interactions, compared to the orthologous protein in *S. cerevisiae*. Furthermore, to better understand the evolution of proteins and their interactions, we developed a large-scale cross-species interactome mapping approach to directly test interactions between *S. pombe* proteins and the *S. cerevisiae* orthologs of their partners. Such analysis is only possible with the availability of two well-controlled high-coverage interactome maps generated with the same technology. We found that, for many conserved interactions, both partners had co-evolved to accommodate new interactions and

functions, and their interaction interfaces can no longer be recognized by their *S. cerevisiae* counterparts.

7.3 RESULTS

Comparison of known interactions in *S. cerevisiae* and *S. pombe*

The number of known protein-protein interactions in *S. pombe* is disproportionately lower than in other model eukaryotic organisms and human. We estimated the number of all known interactions in *S. cerevisiae* and *S. pombe* by analyzing seven commonly-used databases – BioGRID (Stark et al., 2011), DIP (Salwinski et al., 2004), IntAct (Kerrien et al., 2012), iRefWeb (Turner et al., 2010), MINT (Ceol et al., 2010), MIPS (Mewes et al., 2011), and VisANT (Hu et al., 2007). There identified 110,443 interactions for budding yeast, but only 4,038 for fission yeast, from these databases. Furthermore, only those interactions or interaction sets that have been validated by at least two independent assays are reliable and defined as "high quality" (Cusick et al., 2009; Das and Yu, 2012). Based on this criterion, 519 fission yeast interactions are of high quality, as opposed to 25,335 high-quality interactions known in budding yeast. Of these, only 160 *S. pombe* interactions are binary (a direct biophysical interaction between the two proteins), as opposed to 11,936 in *S. cerevisiae*. These numbers indicate the extent to which the fission yeast interactions are underexplored and necessitate the systematic mapping of its interactome network.

StressNet: A large-scale high-quality protein interactome network for stress response and cellular signaling in *S. pombe*

The subset of 658 genes for this study was selected using Gene Ontology (GO) (Ashburner et al.,

2000) “Biological Process” (BP) functional annotations for fission yeast (Figure 7.1A). To generate a high-quality high-coverage stress-response interactome map for *S. pombe*, we screened all possible protein pairs (>430,000) in this space three times using a high-quality HT-Y2H system, as we had done for *S. cerevisiae* (Yu et al., 2008). The resulting protein interactome network, StressNet (Figure 7.1B), comprises 235 high-quality binary interactions among 200 proteins (Table 7.1). Of these, 218 interactions were previously unknown. To validate our experimental pipeline and the quality of StressNet, from the 160 high-quality binary interactions we selected a set of 54 well-documented protein interactions from the literature and [“positive reference set” (PRS); Table 7.2] and 43 random protein pairs that have never been reported or predicted to interact [“random reference set” (RRS); Table 7.3]. 20 PRS interactions were successfully confirmed in our pipeline, whereas none of the RRS pairs were detected as positives (Figure 7.1C). Therefore, the sensitivity [fraction of detected true positives among all possible true positives (Yu et al., 2008)] of our Y2H assay is 37.0%.

To directly measure the quality of our Y2H-identified interactions (Braun et al., 2009; Yu et al., 2008), we re-tested all 235 interactions detected in our HT-Y2H screen by two orthogonal assays: the protein complementation assay (PCA) (Remy and Michnick, 2006) and the well-based nucleic acid programmable protein array (wNAPPA) (Ramachandran et al., 2004), producing a fully-verified large-scale interactome map. The confirmation rates of our interactions with both orthogonal assays were similar to those of the PRS, further validating the high quality of StressNet (Braun et al., 2009; Yu et al., 2008) (Figure 7.1C). Using the results of the validating assays, we calculated the precision of StressNet as $95.3 \pm 4.7\%$ (Eq. 8 and 9 in Materials and Methods).

To assign a confidence score to each interaction in StressNet, we implemented a random forest algorithm to integrate results from the three orthogonal assays (figs. S1 and S2 and Materials and Methods). Every detected interaction had a confidence score >0.76 (Table 7.1). This value represents a normalized probability on a scale of 0 to 1 and indicated that all the interactions in StressNet were of high quality. Finally, to evaluate the topological properties of our network, we plotted the degree (number of interactions each protein has) distribution of StressNet (Figure 7.1D). Protein interactomes are small-world scale-free networks (Barabasi and Albert, 1999; Jeong et al., 2000) and our stress-response interactome for *S. pombe* exhibited similar topological properties to other large-scale biological networks.

To assess the biological relevance of this network, we investigated overall relationships between protein pairs using expression and genetic interaction profile similarities (Roguev et al., 2008; Rustici et al., 2004), subcellular colocalization (Matsuyama et al., 2006), and GO functional similarities (Ashburner et al., 2000). We found significant enrichment of interactions in StressNet of protein pairs that colocalized or were functionally similar, and that were encoded by coexpressed genes or genes that exhibited similar genetic interaction profiles [calculated using the Pearson Correlation Coefficient (PCC)], relative to random expectation (Figure 7.2A-D). Furthermore, the enrichment of StressNet in all four categories was similar to that of high-quality literature-curated binary interactions. These results confirmed the high quality of StressNet and indicated that these interactions are likely to be functionally relevant.

Evolutionary relationships in StressNet

For biological networks, evolutionary relationships are commonly measured in terms of conservation and rewiring: If a pair of interacting proteins in one species has corresponding

orthologs in another that also interact, then the interaction is considered to be conserved (an interolog); otherwise, the interaction is considered to be rewired (Matthews et al., 2001; Shou et al., 2011; Yu et al., 2004) (Figure 7.3A). To understand key principles governing the evolution of protein-protein interactions, especially for those in stress-response and signaling pathways, we compared the interactions in StressNet to their corresponding ortholog pairs in *S. cerevisiae*. We experimentally tested all corresponding *S. cerevisiae* protein pairs of the 235 interactions in StressNet and found that for 35 interactions, the corresponding budding yeast ortholog pairs were detected as interacting by our Y2H experiments. We developed a Bayesian framework to calculate the percentage of conserved interactions based on three parameters – the proportion of observed conserved interactions (35/235 = 14.9%), the precision ($95.3 \pm 4.7\%$), and the sensitivity ($37.0\% \pm 4.4\%$) of our Y2H assay (see Eq. 12 and 13 in Materials and Methods). Substituting appropriate values, the percentage of conserved interactions between *S. pombe* and *S. cerevisiae* is calculated as $36.3 \pm 2.9\%$ (Figure 7.3B).

Using an orthogonal approach, we supplemented *S. cerevisiae* interactions detected in our Y2H experiments with high-quality known *S. cerevisiae* interactions curated from the literature to obtain 55 more StressNet interactions for which the corresponding budding yeast orthologs were reported to interact in the literature (Das and Yu, 2012). There are 90 (35 + 55) conserved interactions in total and the conservation is $38.3\% \pm 3.2\%$, consistent with the conservation calculated using the Bayesian framework (Figure 7.3B). Furthermore, this agreement shows that after combining our Y2H experimental results with high-quality literature-curated interactions, the number of known interactions in our search space in *S. cerevisiae* is nearly complete, because if there were still a large number of unidentified interactions, the observed proportion of conserved interactions based on literature-curated interactions would have been much lower.

Because it is always difficult to determine a negative interaction (Ben-Hur and Noble, 2006; Yu et al., 2008), to ensure the set of rewired interactions is of high quality, we used a stringent set of criteria to define them as those StressNet interactions without corresponding *S. cerevisiae* ortholog pairs and those interactions whose corresponding *S. cerevisiae* ortholog pairs have other high-quality interactions but have never been reported as interacting in the literature or tested positive in our Y2H experiments, and these ortholog pairs are known to have different cellular localizations (Huh et al., 2003).

Proteins encoded by essential genes, those when deleted cause lethality, tend to have more interacting partners (hubs) and also evolve more slowly than non-essential ones (Fraser et al., 2002; Hirsh and Fraser, 2001). We found that essential and non-essential genes (Kim et al., 2010) in our interactome were equally likely to be involved in conserved interactions (Figure 7.3C), contrary to previous studies (Fraser et al., 2002). Stress-response and signal-transduction pathways play a crucial role in the process of adaptation to distinct ecological environments. As measured by the ratio of nonsynonymous to synonymous substitution rates (*dN/dS*) (Nei and Gojobori, 1986; Rhind et al., 2011), we found that the essential genes in these pathways evolve at the same rate as the non-essential genes in the pathways evolve, although on average all essential genes in the genome evolve significantly slower than non-essential genes. To ensure that this is not an artifact of the calculation method, we also calculated *dN/dS* values for all essential and non-essential genes. Consistent with earlier findings (Das and Yu, 2012), we observed that overall, the essential genes had a significantly lower average *dN/dS*. The average *dN/dS* for all stress-response genes is not significantly different from that for the entire genome. The *dN/dS* distributions for these two species are highly similar. This finding is consistent with analyses that suggest that these species are at comparable evolutionary distances from *S. pombe*

and confirm that there are no inherent biases in our *dN/dS* calculations. Thus, our findings suggest that essential genes in stress-response and signal transduction pathways are under less negative selection such that their interactions are rewired for adaptive advantages through evolution.

To better understand the mechanisms underlying conservation and rewiring of interactions, we examined the relationship between sequence similarity of orthologous pairs and interaction conservation rates. Consistent with expectation (Yu et al., 2004), interactions involving proteins with higher overall sequence similarity or identity were more likely to be conserved (Figure 7.3D). However, proteins interact through specific domains (Finn et al., 2010); therefore, we examined the role of sequence similarity of these interfaces in determining the conservation of corresponding interactions. Previous studies have established a homology modeling approach (Kim et al., 2006; Wang et al., 2012) to locate interaction interfaces using co-crystal structures in PDB (Berman et al., 2000) and have found that analysis of these interfaces provides insights into their evolutionary rate (Kim et al., 2006). The conservation of an interaction depends on the conservation of the interfaces involved (Espadaler et al., 2005). Using a similar approach, we inferred interaction interfaces for proteins involved in 161 interactions in our network (Materials and Methods). We found no significant correlation between the similarity or identity of interaction interfaces and the conservation of the corresponding interactions (Figure 7.3E). Examination of the average *dN/dS* ratios for proteins with different numbers of rewired interactions showed that the selection pressure on the gene did not affect the degree to which the interactions of the corresponding protein were rewired (Figure 7.3F), further indicating that the rewiring of interactome networks and the divergence of related species are not completely dictated by evolution detected at the sequence level.

Functional profile of conserved and rewired interactions

To investigate whether gene pairs encoding proteins involved in conserved and rewired interactions are differently regulated at the transcriptional level, we measured global coexpression between these pairs using the PCC. Global coexpression means that the pattern of gene expression of both genes is the same. Whereas conserved interactions had the highest fraction of coexpressed pairs, gene pairs encoding proteins involved in rewired interactions were also significantly more coexpressed than random in *S. pombe* (Figure 7.4A). We also calculated coexpression relationships for the corresponding budding yeast pairs. By definition, the conserved pairs also interact in budding yeast, but the rewired pairs do not. The enrichment in gene expression is consistent with this distinction: Gene pairs encoding proteins involved in conserved interactions were coexpressed, genes encoding rewired pairs were not significantly enriched than random expectation in *S. cerevisiae* (Figure 7.4A).

PCC captures only global coexpression relationships, but cannot capture local or transient coexpression that occurs only under certain conditions. Furthermore, gene pairs encoding proteins involved in stable interactions tend to be globally coexpressed, whereas those in transient interactions are often only locally coexpressed without significant PCC values (Das et al., 2012). Stable and transient interactions both have important biological functions – the former constitute tightly connected modules, whereas the latter form key links between modules, especially in signal transduction pathways, and are more important than the stable ones or random interactions in maintaining the integrity of cellular networks (Das et al., 2012). To detect transient interactions, we used the Local Expression-correlation Scores (LES) (Das et al., 2012; Qian et al., 2001). Rewired interactions in fission yeast had significantly higher LES values

(Figure 7.4B) than both conserved interactions and random expectation, suggesting that transient interactions are more likely to be rewired through evolution. Rewired pairs in budding yeast had LES values lower than random pairs (Figure 7.4B), indicating that gene regulation for these pairs is also rewired.

Next, we examined GO functional similarities between interacting proteins involved in conserved and rewired interactions. Whereas conserved interactions had higher functional similarity than rewired interactions in fission and budding yeast, interacting protein pairs in both categories were significantly more functionally similar than random (Figure 7.4C). This is in agreement with previous findings that conserved interactions tend to be in modules with specific functions, whereas rewired interactions tend to be inter-modular and have greater diversity in function (Das et al., 2012).

In our analysis of rewired interactions above, we focused on those that are present in fission yeast but lost in budding yeast. Because the *S. pombe* interactome is still considerably underexplored in the literature and the sensitivity of our Y2H assay is 37.0%, it is not yet possible to determine non-interacting pairs in *S. pombe* reliably. Therefore, although it is possible to define lost interactions in *S. cerevisiae* by combining literature-curated interactions with our Y2H-detected ones, the same cannot be done to define lost interactions in *S. pombe*. However, there are 1,638 *S. cerevisiae* interactions where one protein has a corresponding *S. pombe* ortholog in the space of the 658 open reading frames (ORFs) that we explored and another protein has no *S. pombe* ortholog. Thus, there can be no corresponding *S. pombe* interactions and these are rewired interactions in *S. cerevisiae* by definition. We found that these interactions had significantly higher PCC, LES, and functional similarity as compared to random. The trend is comparable to that of rewired interactions in *S. pombe* (Figure 7.4), further

confirming the robustness of our results. We performed PCC and LES analysis of coexpression and functional similarity of conserved and rewired interactions defined at different confidence levels and obtained similar results, indicating that the analysis is robust and reliable.

Modes of rewiring uncovered by cross-species interactome mapping

To further understand the meaning of "conservation" of interactions and experimentally explore the molecular mechanisms through which interaction interfaces evolve, we performed a systematic cross-species interactome mapping by testing all conserved interactions between corresponding *S. cerevisiae* and *S. pombe* proteins. Using orthologous pairs of interacting proteins in the two yeast species, we examined whether a protein in one species interacted with the ortholog of its partner in the other (Figure 7.5A). Because we could detect the original interacting pairs from the same species with our Y2H experiments, we know that all four proteins are correctly expressed, folded, and are amenable to detection by our Y2H approach, thereby avoiding technical false negatives. The traditional definition of "conservation" implies the notion of conserved interfaces across different species. However, there are many examples where proteins with conserved interactions form new interactions and carry out new functions that are not conserved. The interface of a conserved interaction in fission yeast is considered "intact" if the proteins involved could also interact with the corresponding orthologs of their partners in budding yeast; otherwise, the interface is considered "co-evolved" (Figure 7.5A). We found that these conserved interactions were equally likely result from an intact interface or co-evolved interface that formed new interaction interfaces that were unrecognizable by their orthologous counterparts in the other species (Figure 7.5B). Earlier studies have suggested that interacting proteins may co-evolve to maintain structural complementarity and binding

specificity (Goh et al., 2000; Hakes et al., 2007; Kim et al., 2004). In this calculation, we used a lenient definition for an intact interface: We considered the interface intact if one or both of the cross-species interactions was positive, which provides a lower bound estimation of co-evolution between interacting proteins.

Divergence of the Sty1 stress-response pathway through interaction conservation and rewiring

In *S. pombe*, Sty1 is activated in response to various stresses, including oxidative and osmotic stress, starvation, and other conditions (Gasch, 2007; Shiozaki and Russell, 1996). Sty1 has orthologs in *S. cerevisiae* (Hog1, with 89% sequence similarity) and human (p38, with 69% sequence similarity). Both p38 and Sty1 respond to a wide range of stresses and both are different from Hog1 in terms of function (Bone et al., 1998). With our stress-response interactome, we detected key interactions at every step of the MAPK signal transduction pathway and, therefore, completely recapitulated the entire Sty1 pathway. This confirmed the sensitivity and accuracy of our HT-Y2H method, especially for discovering transient interactions in signaling pathways. Among all Sty1 interactions in StressNet, those with its activator (Wis1) and inhibitor (Pyp2) were both conserved between the two yeast species, and the Sty1-Wis1 interaction interface was intact. By contrast, the interaction between Sty1 and its known target in fission yeast, Atf1, represented a rewired interaction (Figure 7.5C). We also identified a previously unknown interactor of Sty1: SPBC2D10.09, a protein that we named Snr1 (Sty1-interacting stress-response protein). To confirm this interaction in vivo, we performed co-immunoprecipitation of tagged proteins expressed in *S. pombe* (Figure 7.5D). The amount of Snr1 pulled down in the presence of Sty1 was greater than that pulled down in the absence of

Sty1, indicating that the interaction with Sty1 stabilizes Snr1 (Figure 7.5D). The corresponding orthologous pair of Hog1 and Ehd3 in *S. cerevisiae* did not interact by Y2H (Figure 7.5E). Cells lacking *snr1* (*snr1*Δ cells) grew slower under stress, similar to *sty*1Δ cells (Figure 7.5F), whereas growth of *ehd3*Δ cells was not compromised. These results suggested that Snr1 is a component of the Sty1 pathway and that its functions diverged from its budding yeast counterpart. Moreover, *snr1* also has a human ortholog, HIBCH, further investigation of which may expand our knowledge of the human p38 MAPK pathway.

7.4 DISCUSSION

We generated StressNet – a high-quality high-coverage binary interactome for stress-response and signal-transduction pathways in the fission yeast, *S. pombe*. All interactions were verified by three orthogonal assays and assigned probabilistic confidence scores. We performed comparative network analysis to study the evolution of protein interactomes between the fission and budding yeast species. Even though 84% of StressNet interactions have corresponding orthologous pairs in *S. cerevisiae*, only about 40% of these interactions are conserved, indicating considerable evolutionary changes beyond simple sequence orthology. Thus, the interolog concept should be used with caution to infer interactions across species, especially if the two are not closely related. Furthermore, our results suggested that rewiring of protein interactome networks in related species is likely a major factor for divergence. Surprisingly, we found no significant correlation between the similarity of interaction interfaces and the conservation of corresponding interactions. This demonstrates that conservation of interactions is more complex than previously expected – domains that are not part of the interaction interface also play some indirect role in

making the interaction possible. Even if the interface is conserved, the corresponding interaction could still be rewired because of steric hindrance due to altered overall structure or loss of nearby structural scaffolds that make the interaction thermodynamically favorable (Kastritis et al., 2011). We also experimentally explored the evolution of interaction interfaces and our analysis indicated that interactions traditionally considered "conserved" are equally likely to have intact interfaces as to have co-evolved ones that are different from their orthologous counterparts. These results suggest a molecular mechanism by which the interactome network is rewired through evolution: Many proteins have co-evolved with their partners to form modified interfaces that can, therefore, accommodate new interactions and functions.

Our results indicated that conserved interactions tended to be stable and rewired ones were more likely to be transient. Therefore, our finding provides a molecular-level mechanistic explanation for previous studies showing that genetic cross talk between functional modules can differ substantially (Frost et al., 2012; Roguev et al., 2008; Ryan et al., 2012). However, our results also suggest that, overall, proteins tend not to rewire all of their interactions; thus, even if they acquire novel interactions, they still generally conserve at least some of the original functions.

Our results indicate that substantial evolutionary changes, both rewiring and co-evolution, of stress-response pathways could be a major mechanism by which different organisms adapt to diverse living environments. Conservation of interactions in other pathways might be different from what we observed here. Therefore, similar cross-species interactome mapping and comparative network analyses of more pathways and species will provide a more comprehensive understanding of underlying principles that help shape distinct characteristics of individual organisms through evolution.

7.5 MATERIALS AND METHODS

Selection of genes for the study

This study focused on stress response and signal transduction proteins (based on GO Biological Process annotations) and their known interactors in *S. pombe*. We also include *S. pombe* orthologs of *S. cerevisiae* proteins that are known to interact with orthologs of fission yeast stress-response and signal transduction proteins. While selecting the 658 ORFs, we also ensured that a set of PRS interactions in *S. pombe* could be constructed with genes from our space, a limiting criterion because there are only 160 binary high-quality *S. pombe* interactions reported in the literature.

Yeast two-hybrid (Y2H)

Y2H experiments were carried out as described (Yu et al., 2011). Briefly, 658 *S. pombe* ORFs in Gateway entry vectors were transferred into AD and DB vectors using Gateway LR reactions. After bacterial transformation, plasmids of all AD-Y and DB-X clones were transformed into yeast two-hybrid strains *MAT***a** Y8800 and *MAT*α Y8930 (genotype: *leu2-3, 112 trp1-901 his3Δ200 ura3-52 gal4Δ gal80Δ GAL2-ADE2 LYS2::GAL1-HIS3 met2::GAL7-lacZ cyh2*^{R}), respectively. The *MAT***a** Y8800 strain was obtained from the *MAT***a** Y550 strain after mutating *CYH2* to introduce cycloheximide resistance. *MAT*α Y8930 was generated by crossing *MAT***a** Y8800 with *MAT*α Y1541 (*3*), followed by sporulation and identification of the *MAT*α cycloheximide-resistant yeast strain by tetrad analysis. After AD-Y and DB-X were transformed

into Y8800 and Y8930, respectively, autoactivators were screened by spotting onto synthetic complete media (SC) lacking histidine and tryptophan (AD-Y) or histidine and leucine (DB-X). These autoactivators were excluded from all further screenings. Each unique DB-X was mated with pools of ~188 unique AD-Y by co-spotting onto yeast extract peptone dextrose (YEPD) plates. Diploids were selected by replica plating onto SC plates without leucine and tryptophan (SC–Leu–Trp). To select for positive interactions, Y2H screening was performed by replica plating the diploids onto SC plates with 1 mM 3-amino-1,2,4-triazole (3-AT) and without leucine, tryptophan, and histidine (SC–Leu–Trp–His+3-AT). SC-Leu-Trp-His plates were used for the HT-Y2H screen in *S. cerevisiae* (Yu et al., 2008). We used 1 mM 3-AT, because this concentration greatly reduces background and improves the quality of the screens (Consortium, 2011; Venkatesan et al., 2009; Yu et al., 2011). Newly occurring autoactivators were determined by concurrently replica plating the diploids onto SC media with cyclohexamide (CHX) and 1 mM 3-AT and lacking leucine and histidine (SC–Leu–His+3-AT+CHX). Screening for these autoactivators relies on CHX to select for cells that do not have the AD plasmid, due to plasmid shuffling. Thus, growth on the latter plate identifies spontaneous autoactivators; these were removed from further analyses. All plates were replica cleaned the following day and scored after three additional days. The space was screened three times.

Y2H positives were grown two to three days at 30°C and then spotted onto four plates for secondary phenotype confirmation (phenotyping II) (SC–Leu–Trp–His+3-AT; SC–Leu–His+3-AT+CHX; SC–Leu–Trp–adenine; SC–Leu–adenine+CHX). Colonies that either grew on SC–Leu–Trp–His+3-AT but not on SC–Leu–His+3-AT+CHX or on SC–Leu–Trp–adenine but not on SC–Leu–adenine+CHX were identified as positives.

For colonies that scored positive in phenotyping II, the identities of DB-X and AD-Y were determined by the Stitch-seq approach (Yu et al., 2011) using Illumina sequencing. All identified interacting pairs were retested by pairwise Y2H.

Construction of PRS and RRS

The PRS and RRS are representatives of true positive interactions and negative pairs, respectively, and we used the PRS and the RRS to optimize the assay performance and they may be interpreted as positive and negative controls. The PRS comprises a set of 54 protein interactions from the literature, each of which is supported by at least two independent assays from two different publications (Table 7.2). RRS pairs were generated from a random selection out of all possible protein pairs within our search space for which no interaction has yet been detected by any method (Table 7.3). Because fission yeast interactions are underexplored, we also required that their corresponding budding yeast ortholog pairs have never been reported to interact.

Another way to construct the RRS is to consider protein pairs with different cellular localizations because these are unlikely to interact. 31 out of the 43 RRS pairs are indeed localized in different cell compartments. Using the whole RRS (Figure 7.1C), we estimate the false positive rate for Y2H, PCA, and wNAPPA are 0/43, 2/43 (4.7% ± 3.2%), and 2/43 (4.7% ± 3.2%), respectively. If we only use the 31 RRS pairs localized in different cell compartments (named "RRS_DiffLocal"), the false positive rates for the three assays are 0/31, 2/31 (6.5% ± 4.4%), and 1/31 (3.2% ± 3.2%). Therefore, the false positive rates for all three assays used in our experiments do not change whether we use the complete RRS or RRS_DiffLocal.

With these controls, we found that 20 of the 54 PRS were detected in our screen and none of the RRS set. We calculated the sensitivity of our assay as 20/54 (37.0% ± 4.4%) .

Protein Complementation Assay (PCA)

S. pombe ORFs available in Gateway entry vectors were transferred by Gateway LR reactions into vectors encoding the two fragments of YFP (Venus variant) fused to the N-terminus of the tested proteins. Baits were fused to the F1 fragment (amino acids 1-158 of YFP) and preys to the F2 fragment (amino acids 159-239 of YFP). After bacterial transformation, plasmid DNA was prepared on a Tecan Freedom Evo bio-robot, and DNA concentrations are determined by OD_{260nm} with a Tecan M1000 in a 96-well format. A 50 ng aliquot of each vector encoding the two proteins was used for transfection into HEK 293T cells in 96-well plates, using Lipofectamine 2000 (Invitrogen) reagent according to the instructions of the manufacturer. At approximately 48 hrs post-transfection, cells were processed with a Tecan M1000. A pair is considered interacting if the YFP fluorescence intensity was ≥2 fold higher over background.

Well-based nucleic-acid programmable protein array (wNAPPA)

ORFs encoding the interacting proteins were cloned into Gateway-compatible pCITE-HA and pCITE-GST vectors by LR reactions. After bacterial transformation, growth, DNA minipreps, and determination of DNA concentration, ~0.5µg of each plasmid were added to Promega TnT coupled transcription-translation mix (catalog number: L4610) and incubated for 90 minutes at 30°C to express proteins. During this time anti-GST antibody-coated 96-well plates (Amersham 96-well GST detection module, catalog number: 27-4592-01) were blocked at room temperature with PBS containing 5% dry milk powder. After protein expression, the expression mix was

diluted in 100µl blocking solution, and added to the emptied pre-blocked 96-well plates. Expression mix was incubated in the 96-well plates for two hours at 15°C with agitation to allow for protein capture. After capture, plates were washed three times and developed by incubation with primary and secondary antibodies. Signal was visualized using chemiluminescence (Amersham ECL Reagents, catalog number: RPN2106) with a Tecan M1000. Wells with ≥3 fold higher intensity over background in either configuration were considered positives.

Measuring the precision of our assay

The precision of the Y2H assay was calculated using PCA and wNAPPA as orthogonal validation assays. Using Bayes' rule we can build relationships between true and false positive rates of Y2H and observed positive interactions by a validating assay as:

$$\Pr(A+ \mid Y+) = \Pr(A+ \mid Y+, T+) \times \Pr(T+ \mid Y+) + \Pr(A+ \mid Y+, T-) \times \Pr(T- \mid Y+) \qquad \text{Eq. 1}$$

where $A+$ corresponds to observing a positive interaction using the validating assay, $Y+$ corresponds to observing a positive interaction using Y2H, and $T+$ ($T-$) corresponds to an interaction being a real positive (negative) interaction. The precision of the Y2H is the term $\Pr(T+ \mid Y+)$ [which is also equal to $1- \Pr(T- \mid Y+)$].

Assuming conditional independence between the validating assay and Y2H based on previously defined reasons (Yu et al., 2008), we can write:

$$\Pr(A+ \mid Y+) = \Pr(A+ \mid T+) \times \Pr(T+ \mid Y+) + \Pr(A+ \mid T-) \times \Pr(T- \mid Y+) \qquad \text{Eq. 2}$$

Solving for the precision of the Y2H assay yields:

$$\Pr(T+ \mid Y+) = \frac{\Pr(A+ \mid Y+) - \Pr(A+ \mid T-)}{\Pr(A+ \mid T+) - \Pr(A+ \mid T-)} \quad \text{Eq. 3}$$

Pr(A+|T+) and Pr(A+|T–) were measured in the PRS and RRS experiments. So, for our Y2H assay we can write precision as:

$$Precision = \frac{F_{StressNet} - F_{RRS}}{F_{PRS} - F_{RRS}} \quad \text{Eq. 4}$$

where $F_{StressNet}$ is the fraction positive by an assay for StressNet, which is the best estimator for Pr(A+|Y+). F_{PRS} is the fraction positive by the assay for the PRS, which is an estimator for Pr(A+|T+). F_{RRS} is the fraction positive by the assay for the RRS, which is an estimator for Pr(A+|T-).

The standard errors of $F_{StressNet}$, F_{PRS}, and F_{RRS} are calculated using the standard error for binomial distributions:

$$StdErr = \sqrt{\frac{F(1-F)}{N}} \quad \text{Eq. 5}$$

where F is the fraction positive by the assay ($F_{StressNet}$, F_{PRS}, or F_{RRS} and N is the total number of pairs tested.

To estimate the standard error for the precision, we used the standard delta method:

$$\sigma_X^2 = (\frac{\partial f}{\partial A}\sigma_A)^2 + (\frac{\partial f}{\partial B}\sigma_B)^2 + (\frac{\partial f}{\partial C}\sigma_C)^2 + ... \quad \text{Eq. 6}$$

where $X = f(A, B, C, \ldots)$. $A, B, C, \ldots$ are independent random variables.

Here, the standard error of the precision is calculated as:

$$\sigma_{precision} = \sqrt{(\frac{1}{F_{PRS} - F_{RRS}})^2 \times \sigma_{StressNet}^2 + \frac{(F_{StressNet} - F_{RRS})^2}{(F_{PRS} - F_{RRS})^4} \times \sigma_{PRS}^2 + \frac{(F_{StressNet} - F_{PRS})^2}{(F_{PRS} - F_{RRS})^4} \times \sigma_{RRS}^2} \quad \text{Eq. 7}$$

We have two validating assays, and we can incorporate the precision rates from these assays by calculating the average precision:

$$Average\ Precision = \frac{Precision_{PCA} + Precision_{wNAPPA}}{2} \quad \text{Eq. 8}$$

The standard error for the average precision is calculated by the delta method as:

$$\sigma_{average\ precision} = \sqrt{\frac{\sigma_{PCA}^2}{4} + \frac{\sigma_{wNAPPA}^2}{4}} \quad \text{Eq. 9}$$

Using this framework, we estimate the precision of our Y2H assay to be 95.3 ± 4.7%.

Calculating confidence scores for interactions

Using the random forest algorithm (Breiman, 2001), we integrate results from Y2H, PCA, and wNAPPA and calculated confidence scores for interactions. Random forest is an ensemble

classifier that constructs multiple decision trees by stochastic discrimination (Kleinberg, 1996) and predicts a final class based on a weighted combination of the output class of each decision tree. It is considered to be a robust and accurate classifier for noisy datasets (Breiman, 2001). We evaluated the performance of our classifier by five-fold cross validation on our reference set (union of PRS and RRS) and obtained moderately good performance (AUC = 0.64).

Determination of orthologs between *S. pombe* and *S. cerevisiae*

We use the list of orthologs provided by PomBase (Wood et al., 2012). The genome of *S. cerevisiae* underwent a duplication event (Kellis et al., 2004). Thus, many *S. pombe* genes have two corresponding *S. cerevisiae* orthologous genes. Moreover, in a number of cases, the same *S. cerevisiae* gene has multiple *S. pombe* orthologs. Thus, the mapping considered for the study is "many-to-many".

Estimation of the conservation of interactions

To estimate the conservation of protein-protein interactions between *S. pombe* and *S. cerevisiae*, we used a Bayesian framework that incorporates the precision and sensitivity of our Y2H assay:

$$\Pr(Det) = \Pr(Det \mid Cons+) \times \Pr(Cons+) + \Pr(Det \mid Cons-) \times \Pr(Cons-) \qquad \text{Eq. 10}$$

where Pr(*Cons*+) corresponds to the conservation of protein-protein interactions between *S. cerevisiae* and *S. pombe*. The best estimator for Pr(*Det*) (the probability of detecting a *S. cerevisiae* interaction among proteins pairs that are orthologous to an interacting protein pair in StressNet) is F_{det}, the fraction of the 235 StressNet interactions in *S. pombe* with corresponding

Y2H-detected interactions in *S. cerevisiae* (35/235). Pr(*Det*|*Cons*+) and Pr(*Det*|*Cons*-) are estimated by F_{PRS} and F_{RRS}, the fractions of PRS and RRS interactions detected by our Y2H assay (20/54 and 0/43, respectively). By definition:

$$\Pr(Cons+) = 1 - \Pr(Cons-) \qquad \text{Eq. 11}$$

We can simplify the earlier equation to obtain an expression for Pr(*Cons*+):

$$\Pr(Cons+) = \frac{\Pr(Det) - \Pr(Det \mid Cons-)}{\Pr(Det \mid Cons+) - \Pr(Det \mid Cons-)} \qquad \text{Eq. 12}$$

To estimate the error for the conservation percentage, we used the standard delta method as described earlier. The standard deviation of *Cons*+ is given by:

$$\sigma_{Cons+} = \sqrt{\frac{(F_{PRS} - F_{RRS})^2 \sigma_{F_{det}}^2 + (F_{det} - F_{RRS})^2 \sigma_{F_{PRS}}^2 + (F_{det} - F_{PRS})^2 \sigma_{F_{RRS}}^2)}{(F_{PRS} - F_{RRS})^4}} \qquad \text{Eq. 13}$$

Using the Y2H data, we calculated a conservation of 36.3 ± 2.9% interactions.

Another approach for measuring the conservation is to calculate fraction of *S. cerevisiae* interactions conserved in *S. pombe*. We mapped all *S. pombe* proteins in our space to their corresponding *S. cerevisiae* orthologs. We calculated the number of interactions in this *S. cerevisiae* space detected by our Y2H assay. We then mapped all the observed *S. cerevisiae* interactions to their corresponding *S. pombe* ortholog pairs and calculate the number of pairs

detected as interacting in StressNet. We find that for 48/386 (12.4%) *S. cerevisiae* interactions, the corresponding *S. pombe* ortholog pairs also interact. Using the Bayesian framework described above, we calculate the conservation between *S. pombe* and *S. cerevisiae* interactions as 34.7 ± 2.0%, which is statistically the same ($P = 0.708$ using a cumulative binomial test) as the conservation calculated using the Y2H results (36.3 ± 2.9%).

Interaction conservation and confidence scores

After supplementing our Y2H experiments with high-quality interactions from the literature, we find that 90/235 (38.3%) interactions are conserved in StressNet. The statistical error associated with this measurement is related to the sample size and is calculated as the standard error [standard error = standard deviation / square root (N), where N is the number of samples]. The standard deviation is calculated based on the underlying probability distribution. The conservation percentage is obtained by a simple division (90/235 = 38.3%) and the underlying probability distribution is binomial (since each interaction can either be conserved or not, it corresponds to a Bernoulli event, the ensemble of which is modeled by a binomial distribution). The standard error is calculated using the appropriate formula for a binomial distribution: square root [$p \times (1-p) / N$] = 3.2%, p = fraction of interactions that are conserved (90/235) and N = sample size (235)].

To test whether interactions with higher confidence scores were more likely to be conserved, we divided all StressNet interactions into two groups. The first group comprises interactions with confidence scores in the lower two quartiles and the second group comprises interactions with confidence scores in the upper two quartiles. We then compared the conservation for these two groups. We find that there is no significant difference ($P = 0.37$ using a two-sided Fisher exact

test) in conservation rate between the two groups. This validates that the observed conservation rate is robust and not correlated with the confidence score associated with each interaction.

Evolutionary rates of genes and protein interactions

The evolutionary rate of genes is commonly measured in terms of the ratio of asynchronous nucleotide substitutions per asynchronous site to synchronous substitutions per synchronous site or *dN/dS*. This quantifies the selective evolutionary pressure on certain protein-coding genes to diverge faster, as opposed to others that may almost remain unchanged across species. To calculate the *dN/dS* values for all *S. pombe* genes, used two sequenced species in the Schizosaccharomyces genus – *S. cryophilus* and *S. octosporus*. To determine orthology relationships, we used BLAST-x with default parameters on all *S. pombe* genes. The top BLAST hit for each *S. pombe* gene against the indexed database of proteins for each of the two species was designated to be an ortholog, provided the E-value of the hit was < 0.05. Although the E-value cutoff is relatively high, it ensures that no potential pairs are missed. For pairs that have been incorrectly estimated to be orthologs, there is a correction step in downstream calculations that will return a *dN/dS* value of NaN (not a number), because of too high divergence. For all orthologous pairs, the Nei-Gojobori algorithm, which uses the Jukes-Cantor substitution model, was used to calculate *dN/dS* values.

Conservation of interactions and sequence similarity

Sequence similarity between *S. pombe* ORFs and their *S. cerevisiae* orthologs was measured by performing pairwise sequence alignment between all known ortholog pairs using the Needle program in the EMBOSS suite (Rice et al., 2000). It uses the Needleman-Wunsch alignment

algorithm (Needleman and Wunsch, 1970) to find the optimum alignment of two sequences along their entire length. The recommended default parameters – an affine gap penalty model (Vingron and Waterman, 1994) with an opening penalty of 10 and an extension penalty of 0.5 and the BLOSUM62 scoring matrix (Henikoff and Henikoff, 1992) – were used for the alignment. Because the lengths of orthologs may be dissimilar, we calculated the overall similarity percentage (OPS) with reference to the length of the *S. pombe* ORFs:

$$OPS = \frac{N_{st}}{L\_Sp_t} \qquad \text{Eq. 14}$$

where, N_{st} is the total number of similar residues and L_Sp_t is the total length of the *S. pombe* ORF.

We then examined the relationship between the similarity percentage and the percentage of conserved interactions. Because the number of interactions varies considerably across different groups corresponding to different similarity percentages, we required each group to have at least 5 interactions. If any group had less than 5 interactions, it was merged with the next (higher) group. This ensured that our results were robust to outlier effects. We found that there was an increase in the degree of conservation with an increase in overall sequence similarity. To examine if the primary cause of this trend is the similarity of conserved domains, we identified domains on ortholog pairs that interact (Finn et al., 2005; Stein et al., 2011). We defined the percentage similarity of interacting domains (PSID) as:

$$PSID = \frac{N_{si}}{L\_Sp_i} \qquad \text{Eq. 15}$$

where N_{si} is the number of similar residues in interacting domains and L_Sp_i is the sum of the lengths of the interacting domains in *S. pombe*.

Inferring interaction interfaces from 3did and iPfam

In this study, we use interacting domains identified by 3did (Stein et al., 2011) and iPfam (Finn et al., 2005) to define interaction interface. To verify the reliability of inferring these domain-domain interactions, we performed three-fold cross-validation for 1,456 interaction pairs that have co-crystal structures. Because there are few co-crystal structures for *S. pombe*, this approach allowed us to obtain a meaningful estimate of the quality of the domain-domain predictions in these two databases. We split the pairs into three subsets such that two subsets were used for training and the third one was the test set. For each interaction pair in the test dataset, we scored a successful structural prediction when the predicted domain-domain interaction(s) had at least one co-crystal structure in support of it. We repeated the procedure thrice with each of the three subsets as the test set. Among the 1,456 PPI pairs, over 90% were correctly predicted with corresponding interacting domains, indicating that the predicted interaction interfaces used for our calculations were accurate (Wang et al., 2012).

Robustness of differences between sets of conserved and rewired interactions

To assess the robustness of the differences between sets of conserved and rewired interactions, we constructed different sets of conserved and rewired interactions corresponding to different confidence levels.

We constructed two sets of conserved interactions at different confidence levels – Conserved_HQ and Conserved_All. Conserved_HQ comprises only those interactions with corresponding *S. cerevisiae* ortholog pairs that tested positive in our Y2H experiments or were confirmed by two or more independent orthogonal assays in the literature. Conserved_All comprises all interactions in Conserved_HQ and those *S. cerevisiae* ortholog pairs that have been reported as interacting in the literature by only one assay.

We constructed five sets of rewired interactions at different confidence levels – Rewired_ByDefn, Rewired_HQ, Rewired_LC, Rewired_All_DiffLocal, and Rewired_All. Rewired_ByDefn comprises only those StressNet interactions for which at least one of the interacting proteins does not have a *S. cerevisiae* ortholog and, therefore, no corresponding interaction can exist in *S. cerevisiae*. Thus, these interactions are rewired by definition. Rewired_HQ comprises all interactions in Rewired_ByDefn and those interactions for which the corresponding *S. cerevisiae* ortholog pairs have other high-quality interactions but have never been reported as interacting in the literature or tested positive in our Y2H experiments, and these ortholog pairs are known to have different cellular localizations. Thus, these correspond to *S. pombe* interactions with corresponding budding yeast ortholog pairs that are in principle non-interacting, because they have different cellular localizations (Jansen et al., 2003; Yu et al., 2008) and they participate in well-validated interactions with other proteins but have never been reported to interact in the literature. Rewired_LC comprises all interactions in Rewired_ByDefn and those interactions with corresponding ortholog pairs that have other high-quality interactions but have never been reported as interacting in the literature or tested positive in our Y2H experiments. Rewired_All_DiffLocal corresponds to all interactions in Rewired_ByDefn and

those interactions with corresponding ortholog pairs that have different cellular localizations. Rewired_All comprises all interactions that are not in Conserved_All.

Construction of myc-sty1 and HA-snr1 expression clones

S. pombe sty1 and *snr1* genes were PCR amplified using the following primers – sty1-pNCH1472-Forward, sty1-pNCH1472-Reverse, snr1-pSGP73-Forward, and snr1-pSGP73-Reverse (Table 7.4). The *sty1* PCR product was cloned into a pNCH1472-myc vector using the NotI and SalI restriction sites. The *snr1* PCR product was cloned into a pSGP73-HA vector using the NotI and BglII restriction sites. pNCH1472-myc-sty1 and pSGP73-HA-snr1 were single or double transformed into *S. pombe* KGY553 (ATCC). Transformed yeast was selected on Edinburgh minimal medium (EMM)–Ura plates for pNCH1472-myc-sty1, EMM–Leu for pSGP73-HA-snr1, and EMM–Ura–Leu for double transformation.

Coimmunoprecipitation and Western blotting

Transformed yeast (KGY553) containing pNCH1472-myc-sty1 or pSGP73-HA-snr1 or both were cultured overnight in 10 mL EMM selection medium. Yeast pellets were washed in 5 mL of cold TE buffer before protein extraction. To lyse cells, 1 mL of lysis buffer (50mM Tris-HCl pH 7.5, 0.2% Tergitol, 150 mM NaCl, 5 mM EDTA, Complete Protease Inhibitor tablet) and 600 μL glass beads were added to each tube and mixed in a beater for two rounds of 10 minutes each. Protein extracts were centrifuged for 10 minutes at 13,200 rpm at 4°C in an Eppendorf 5415R centrifuge.. Then, 500 μL of supernatant was immunoprecipitated overnight using 20 μL of EZview™ Red Anti-c-Myc Affinity Gel (Sigma-Aldrich E6654) or EZview™ Red Anti-HA Affinity Gel (Sigma-Aldrich E6779). The next morning, beads were washed three times using

cold lysis buffer before being subjected to SDS-PAGE and Western blotting analysis. Primary antibodies used in our analysis were anti-c-Myc (Santa Cruz sc-789), anti-HA (Roche 12CA5), and anti-γ-tubulin (Sigma-Aldrich T5192).

Construction of yeast deletion strains

The *snr1*Δ strain was obtained from the Bioneer *Schizosaccharomyces pombe* Genome-wide Deletion Library. The deletion strain was verified by PCR using primers SpEhd3_Up_Fwd and Sp_Dn_Rev spanning the 3' end of *snr1* and the region immediately downstream. Primers specific for KanMX4 (KanMX4-Fwd and KanMX4-Rev) were used to detect the deletion cassette. A PCR-based strategy was used to construct the *sty1*Δ strain. Briefly, in the first round of PCR, primers (PFA6a_Sty1_Fwd and PFA6a_Sty1_Rev) with 20 base pairs (bp) homology to the regions upstream and downstream of *sty1*, respectively, were synthesized for PCR of the pFA6a-KanMX6 cassette. Primers with 20 bp homology to the pFA6a-KanMX6 were synthesized to PCR 290 bp upstream (Sty1Del-Up_Fwd and Sty1Del-Up_Rev) and 290 bp downstream (Sty1Del-Dn_Fwd and Sty1Del-Dn_Rev) of *sty1*, not including the *sty1* gene. The three PCR products were stitched together sequentially with a second round of PCR. Stitch PCR of the upstream region and pFA6a-KanMX6 and of the downstream region and pFA6a-KanMX6 were carried out separately. In the third round of PCR, both upstream and downstream stitched PCR products were further stitched together to produce a final product of pFA6a-KanMX6 flanked on the 5' and 3' ends by 290 bp that are homologous to the upstream and downstream regions of chromosomal sty1 (Sty1Del-Up_Fwd and Sty1Del-Dn_Rev). The final PCR product was transformed into *S. pombe* 972h- canonical wild-type (ATCC). Transformed yeast was selected on yeast extract-sucrose (YES) media plates containing 150 mg/L G418. Insertion of the

pFA6a-KanMX6 cassette by homologous recombination at the *sty1* locus was verified by PCR using primers to target the entire cassette (Sty1Del-Up_Fwd and Sty1Del-Dn_Rev) and to target a sty1 internal region of 401 bp (Sty1_Fwd and Sty1_Rev). In addition, sty1 and snr1 deletion were performed in S. pombe KGY553 (ATCC) wild-type (*h- his3-D1 leu1-32 ura4-D18 ade6-M216*) background using a similar PCR strategy. Sequences of primers used for deletion and verification of strains in this study are listed in Table 7.4.

Stress Sensitivity Assays

S. cerevisiae was grown in YEPD and *S. pombe* was grown in YES medium. All yeast strains were initially grown as a starter culture overnight at 30°C. From the starter culture, yeast cells were diluted into fresh medium to an initial OD_{600nm} = 0.2. The cultures were grown to mid-log phase (OD_{600nm} = 0.7). The *S. cerevisiae* and *S. pombe* strains were serially diluted 4-fold in sterile water and spotted onto YEPD and YES plates, respectively, containing various stressors. Spotted plates were incubated at 30°C and yeast growth was assessed after 3 days.

7.6 FIGURE AND TABLE LEGENDS

Figure 7.1. *S. pombe* stress-response binary interactome network, StressNet. (**A**) Functional classification of the proteins included in our high-quality high-coverage HT-Y2H screen. (**B**) Network view of the stress-response binary interactome network in *S. pombe*. (**C**) Fraction of protein pairs in PRS, RRS, and StressNet that tested positive using Y2H, PCA, and wNAPPA. Data are shown as measurements + statistical error (SE). (**D**) Degree distribution of StressNet. $P(k)$ is the probability that a protein has a degree = k.

Figure 7.2. Biological properties of StressNet interactions. (**A**) Pearson correlation coefficient (PCC) distribution of expression profiles of interacting and random protein pairs (dashed line corresponds to PCC cutoff above which pairs are considered to be significantly coexpressed; inset shows the fraction of significantly coexpressed pairs). (**B**) PCC distribution of genetic interaction profiles of interacting and random protein pairs (dashed line corresponds to PCC cutoff above which pairs are considered to be significantly similar; inset shows the fraction of pairs with significantly similar interaction profiles). (**C**) Enrichment of colocalized protein pairs. (**D**) Enrichment of protein pairs sharing similar functions. For each panel, the random set is constructed by considering all pairwise combinations of genes or proteins in the corresponding space. All *P* values represent comparisons between StressNet interactions and random pairs using a cumulative binomial test. Inset graphs and data in C and D are shown as measurements + SE.

Figure 7.3. Evolutionary analysis of interactions. (**A**) Schematic of conserved and rewired interactions between the two yeast species. *S.p.*, *S. pombe*; *S.c.*, *S. cerevisiae* (**B**) Conservation rate (fraction of conserved interactions) in our interactome calculated in two different ways. Measured represents the value calculated using a Bayesian framework that incorporates the

precision and recall of our assay. Literature represents the value estimated using budding yeast interactions reported in the literature. (**C**) Fraction of conserved interactions involving essential and non-essential proteins. The differences in B and C are not significant based on a cumulative binomial test. (**D**) Distribution of the fraction of conserved interactions as a function of overall sequence similarity. (**E**) Distribution of the fraction of conserved interactions as a function of sequence similarity of interaction interfaces. For D and E, *P* values are used to test whether there is a significant difference (using a cumulative binomial test) in conservation percentage between the groups corresponding to the lowest and highest similarity percentages. R^2 (coefficient of determination) represents the significance of the correlation between conservation and similarity percentages. **(F)** Distribution of *dN/dS* [ratio of the number of non-synonymous substitutions per non-synonymous site (*dN*) to the number of synonymous substitutions per synonymous site (*dS*)] as a function of number of rewired interactions. The differences are not significant based on a two-sided Kolmogorov-Smirnov test. Data are shown as the measurements + SE.

Figure 7.4. Functional analysis of conserved and rewired interactions in *S. pombe* and *S. cerevisiae*. (**A**) Fraction of globally coexpressed pairs (as measured by PCC) among conserved and rewired interactions. (**B**) Fraction of locally coexpressed pairs (as measured by LES) among conserved and rewired interactions (**C**) Fraction of functionally similar pairs among conserved and rewired interactions. For each panel, the random set is constructed by considering all pairwise combinations of genes/proteins in the corresponding space. All *P* values represent comparisons between rewired interactions and random pairs using a cumulative binomial test. Data are shown as measurements + SE.

Figure 7.5. Analysis of intact and co-evolved interactions. (**A**) Schematic of intact and co-evolved interactions. (**B**) Fraction of intact and co-evolved interactions in our interactome. Data

are shown as measurements + SE. No significant difference detected using a cumulative binomial test. (**C**) The MAPK Sty1 stress response pathway. All undirected lines represent interactions detected in our interactome. The black arrow represents transcriptional regulation. (**D**) Sty1-Snr1 interaction validated in *S. pombe* using co-immunoprecipitation ($N = 3$ blots). (**E**) Y2H analysis of the ability of Hog1 and Ehd3 to interact and Sty1 and Snr1 to interact ($N = 3$ experiments). (**F**) Sensitivity assays for different deletion strains of *S. cerevisiae* and *S. pombe* under various stress conditions ($N = 3$ experiments).

Table 7.1. StressNet and interaction confidence scores.

Table 7.2. Positive reference set.

Table 7.3. Random reference set.

Table 7.4 Primers used in the study.

7.7 REFERENCES

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

Barabasi, A.L., and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509-512.

Ben-Hur, A., and Noble, W.S. (2006). Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics *7 Suppl 1*, S2.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Bone, N., Millar, J.B., Toda, T., and Armstrong, J. (1998). Regulated vacuole fusion and fission in Schizosaccharomyces pombe: an osmotic response dependent on MAP kinases. Curr Biol *8*, 135-144.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.S.*, et al.* (2009). An experimentally derived confidence score for binary protein-protein interactions. Nat Methods *6*, 91-97.

Breiman, L. (2001). Random forests. Machine Learning *45*, 5-32.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. Nucleic Acids Res *38*, D532-539.

Consortium, A.I.M. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.R., Simonis, N., Rual, J.F., Borick, H., Braun, P., Dreze, M.*, et al.* (2009). Literature-curated protein interaction datasets. Nat Methods *6*, 39-46.

Das, J., Mohammed, J., and Yu, H. (2012). Genome scale analysis of interaction dynamics reveals organization of biological networks. Bioinformatics.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

Espadaler, J., Romero-Isart, O., Jackson, R.M., and Oliva, B. (2005). Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. Bioinformatics *21*, 3360-3368.

Finn, R.D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics *21*, 410-412.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K.*, et al.* (2010). The Pfam protein families database. Nucleic Acids Res *38*, D211-222.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C.*, et al.* (2005). Protein interaction mapping: a Drosophila case study. Genome Res *15*, 376-384.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., and Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. Science *296*, 750-752.

Frost, A., Elgort, M.G., Brandman, O., Ives, C., Collins, S.R., Miller-Vedam, L., Weibezahn, J., Hein, M.Y., Poser, I., Mann, M.*, et al.* (2012). Functional repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. Cell *149*, 1339-1352.

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

Gasch, A.P. (2007). Comparative genomics of the environmental stress response in ascomycete fungi. Yeast *24*, 961-976.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E.*, et al.* (2003). A protein interaction map of Drosophila melanogaster. Science *302*, 1727-1736.

Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., and Cohen, F.E. (2000). Co-evolution of proteins with their interaction partners. J Mol Biol *299*, 283-293.

Hakes, L., Lovell, S.C., Oliver, S.G., and Robertson, D.L. (2007). Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proc Natl Acad Sci U S A *104*, 7999-8004.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A *89*, 10915-10919.

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

Hu, Z., Ng, D.M., Yamada, T., Chen, C., Kawashima, S., Mellor, J., Linghu, B., Kanehisa, M., Stuart, J.M., and DeLisi, C. (2007). VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. Nucleic Acids Res *35*, W625-632.

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

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A *98*, 4569-4574.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science *302*, 449-453.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L. (2000). The large-scale organization of metabolic networks. Nature *407*, 651-654.

Kastritis, P.L., Moal, I.H., Hwang, H., Weng, Z., Bates, P.A., Bonvin, A.M., and Janin, J. (2011). A structure-based benchmark for protein-protein binding affinity. Protein Sci *20*, 482-491.

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

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U.*, et al.* (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res *40*, D841-846.

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

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

Kim, W.K., Bolser, D.M., and Park, J.H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). Bioinformatics *20*, 1138-1150.

Kleinberg, E.M. (1996). An overtraining-resistant stochastic modeling method for pattern recognition. Ann Stat *24*, 2319-2349.

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

Matsuyama, A., Arai, R., Yashiroda, Y., Shirai, A., Kamata, A., Sekido, S., Kobayashi, Y., Hashimoto, A., Hamamoto, M., Hiraoka, Y.*, et al.* (2006). ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol *24*, 841-847.

Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res *11*, 2120-2126.

Mewes, H.W., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K.F., Stumpflen, V.*, et al.* (2011). MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res *39*, D220-224.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443-453.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol *3*, 418-426.

Qian, J., Dolled-Filhart, M., Lin, J., Yu, H., and Gerstein, M. (2001). Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. J Mol Biol *314*, 1053-1066.

Ramachandran, N., Hainsworth, E., Bhullar, B., Eisenstein, S., Rosen, B., Lau, A.Y., Walter, J.C., and LaBaer, J. (2004). Self-assembling protein microarrays. Science *305*, 86-90.

Remy, I., and Michnick, S.W. (2006). A highly sensitive protein-protein interaction assay based on Gaussia luciferase. Nat Methods *3*, 977-979.

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

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

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

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N.*, et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.

Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C.J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bahler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. Nat Genet *36*, 809-817.

Ryan, C.J., Roguev, A., Patrick, K., Xu, J., Jahari, H., Tong, Z., Beltrao, P., Shales, M., Qu, H., Collins, S.R.*, et al.* (2012). Hierarchical modularity and the evolution of genetic interactomes across species. Mol Cell *46*, 691-704.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res *32*, D449-451.

Shevchenko, A., Roguev, A., Schaft, D., Buchanan, L., Habermann, B., Sakalar, C., Thomas, H., Krogan, N.J., and Stewart, A.F. (2008). Chromatin Central: towards the comparative proteome by accurate mapping of the yeast proteomic environment. Genome Biol *9*, R167.

Shiozaki, K., and Russell, P. (1996). Conjugation, meiosis, and the osmotic stress response are regulated by Spc1 kinase through Atf1 transcription factor in fission yeast. Genes Dev *10*, 2276-2288.

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

Simonis, N., Rual, J.F., Carvunis, A.R., Tasan, M., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Sahalie, J.M., Venkatesan, K., Gebreab, F.*, et al.* (2009). Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. Nat Methods *6*, 47-54.

Sipiczki, M. (2000). Where does fission yeast sit on the tree of life? Genome Biol *1*, REVIEWS1011.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X.*, et al.* (2011). The BioGRID Interaction Database: 2011 update. Nucleic Acids Res *39*, D698-704.

Stein, A., Ceol, A., and Aloy, P. (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res *39*, D718-723.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S.*, et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957-968.

Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M., and Wodak, S.J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford) *2010*, baq023.

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

Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Vingron, M., and Waterman, M.S. (1994). Sequence alignment and penalty choice. Review of concepts, case studies and implications. J Mol Biol *235*, 1-12.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotech *30*, 159-164.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S.*, et al.* (2002). The genome sequence of Schizosaccharomyces pombe. Nature *415*, 871-880.

Wood, V., Harris, M.A., McDowall, M.D., Rutherford, K., Vaughan, B.W., Staines, D.M., Aslett, M., Lock, A., Bahler, J., Kersey, P.J.*, et al.* (2012). PomBase: a comprehensive online resource for fission yeast. Nucleic Acids Res *40*, D695-699.

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

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res *14*, 1107-1118.

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X.*, et al.* (2011). Next-generation sequencing to generate interactome datasets. Nat Methods *8*, 478-480.

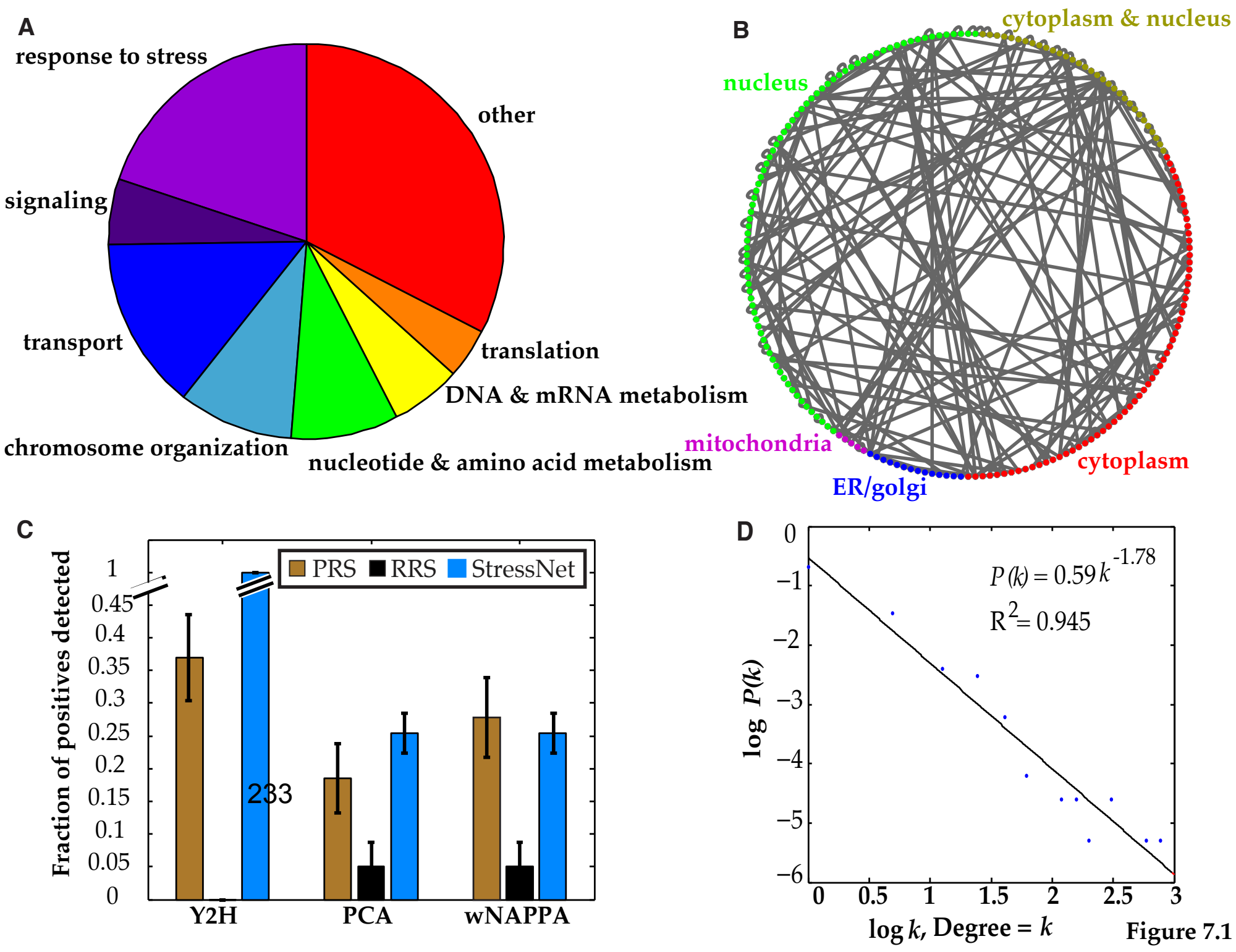

A
response to stress
other
signaling
transport
translation
DNA & mRNA metabolism
chromosome organization
nucleotide & amino acid metabolism
B
cytoplasm & nucleus
nucleus
mitochondria
cytoplasm
ER/golgi
C
Fraction of positives detected
PRS
RRS
StressNet
233
Y2H
PCA
wNAPPA
D
log P(k)
P(k) = 0.59 k^-1.78
R^2 = 0.945
log k, Degree = k

Figure 7.1

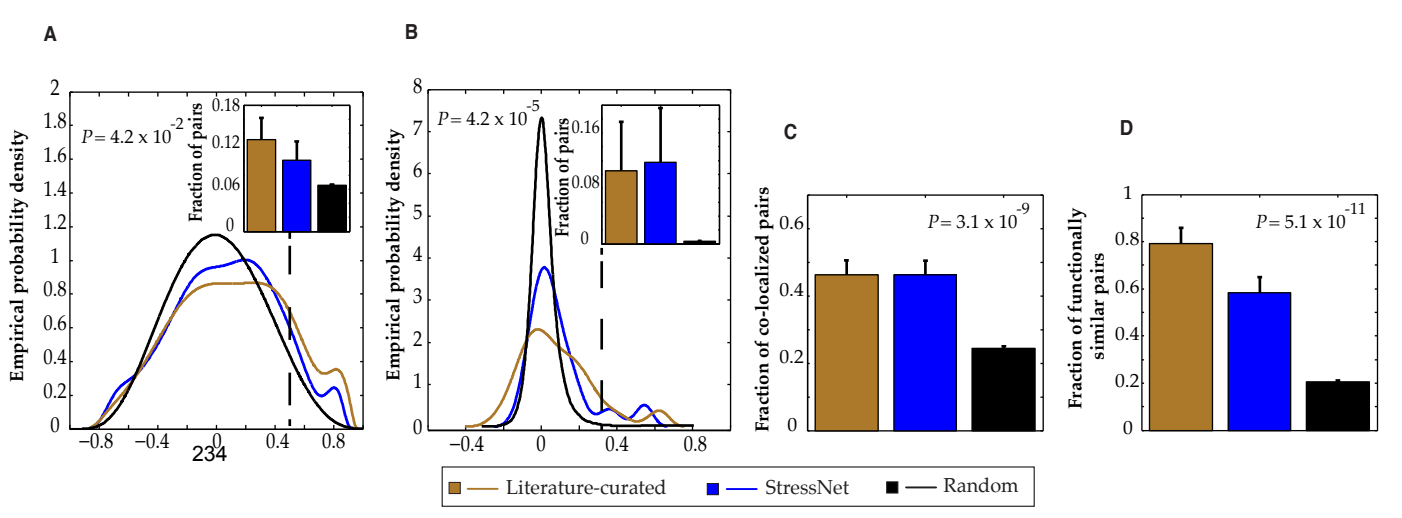

A
B
C
D
Empirical probability density
P = 4.2 x 10^-2
Fraction of pairs
0.18
0.12
0.06
0
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−0.8
−0.4
0
0.4
0.8
234
P = 4.2 x 10^-5
0.16
0.08
7
6
5
4
3
Fraction of co-localized pairs
P = 3.1 x 10^-9
Fraction of functionally similar pairs
P = 5.1 x 10^-11
Literature-curated
StressNet
Random

Figure 7.2

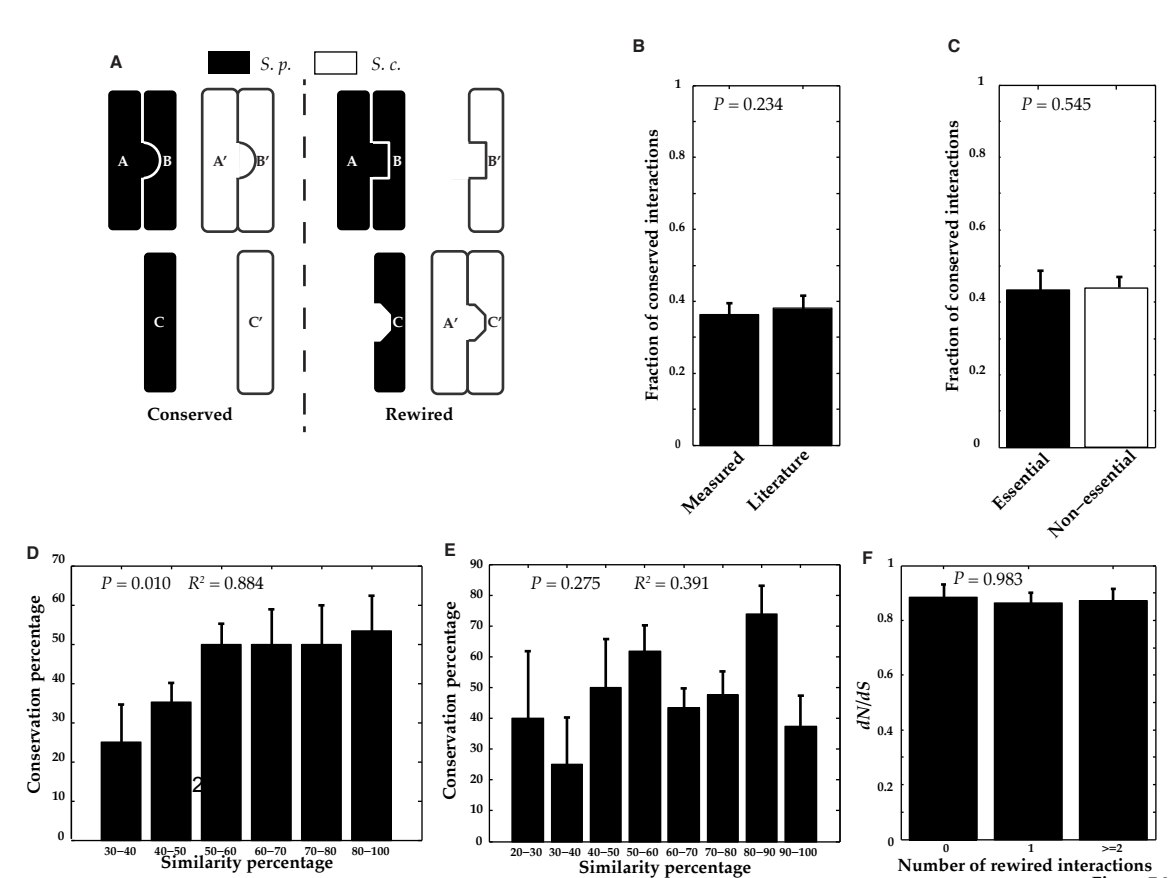

A
S. p.
S. c.
A
B
A′
B′
C
C′
Conserved
Rewired
B
Fraction of conserved interactions
P = 0.234
Measured
Literature
C
P = 0.545
Essential
Non–essential
D
P = 0.010 R² = 0.884
Conservation percentage
30–40
40–50
50–60
60–70
70–80
80–100
Similarity percentage
E
P = 0.275 R² = 0.391
20–30
30–40
40–50
50–60
60–70
70–80
80–90
90–100
F
P = 0.983
dN/dS
0
1
>=2
Number of rewired interactions

Figure 7.3

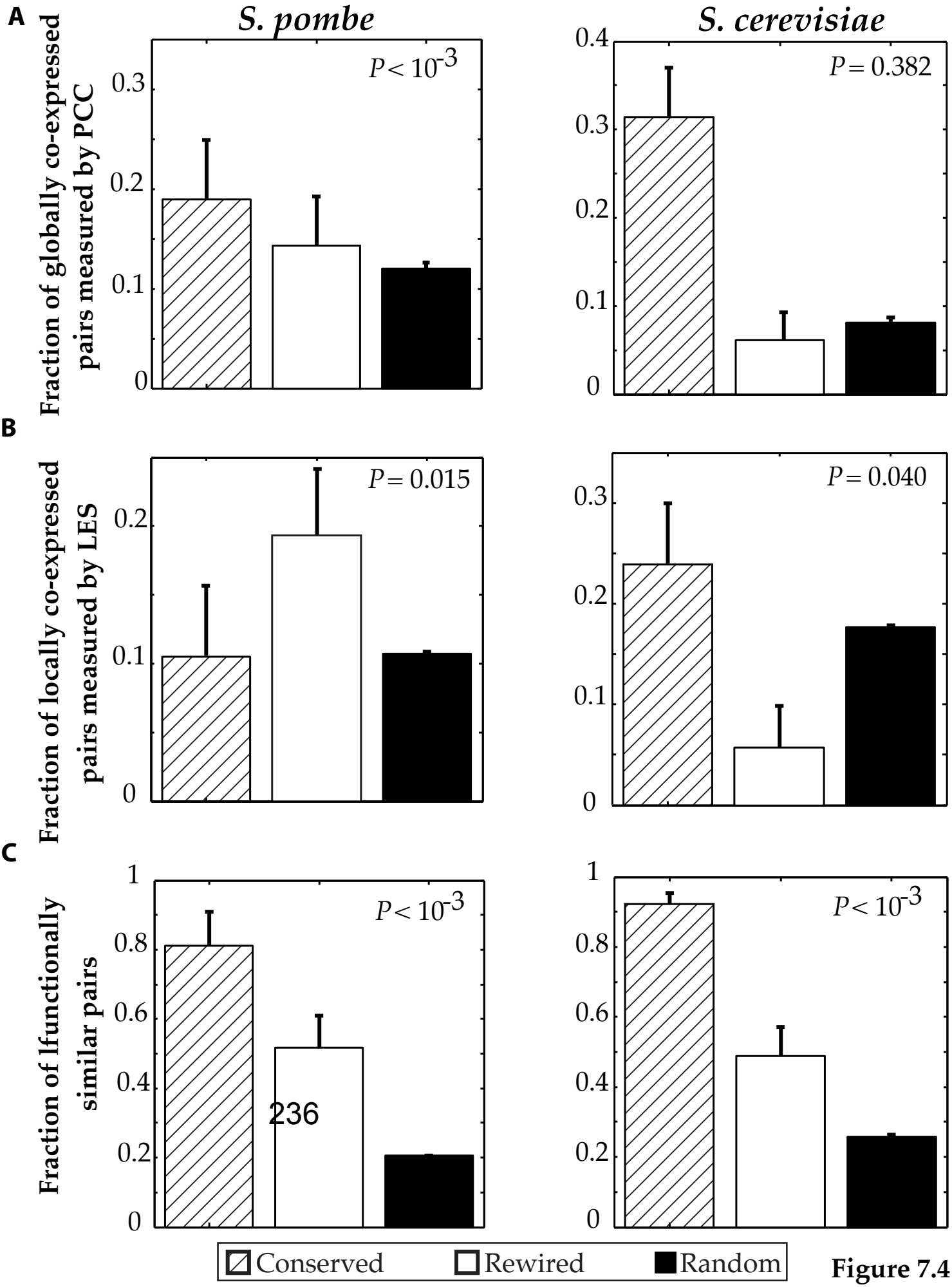

A
B
C
S. pombe
S. cerevisiae
Fraction of globally co-expressed pairs measured by PCC
Fraction of locally co-expressed pairs measured by LES
Fraction of lfunctionally similar pairs
P < 10-3
P = 0.382
P = 0.015
P = 0.040
P < 10-3
P < 10-3
236
0
0.1
0.2
0.3
0.4
0.6
0.8
1
Conserved
Rewired
Random

Figure 7.4

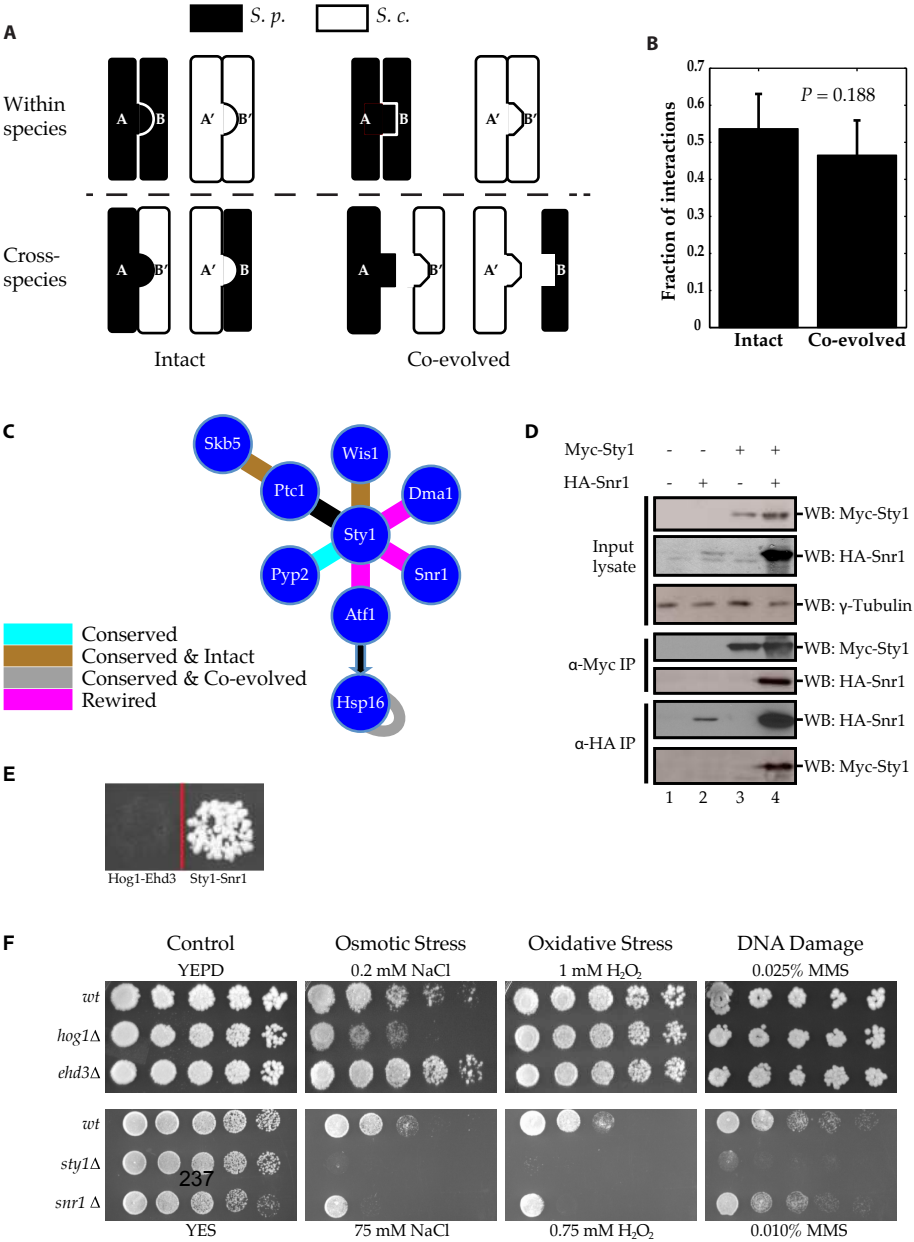

A
S. p.
S. c.
Within species
Cross-species
A
B
A′
B′
Intact
Co-evolved
B
Fraction of interactions
P = 0.188
Intact
Co-evolved
C
Skb5
Wis1
Ptc1
Dma1
Sty1
Pyp2
Snr1
Atf1
Hsp16
Conserved
Conserved & Intact
Conserved & Co-evolved
Rewired
D
Myc-Sty1
HA-Snr1
Input lysate
α-Myc IP
α-HA IP
WB: Myc-Sty1
WB: HA-Snr1
WB: γ-Tubulin
1 2 3 4
E
Hog1-Ehd3
Sty1-Snr1
F
Control
YEPD
Osmotic Stress
0.2 mM NaCl
Oxidative Stress
1 mM H2O2
DNA Damage
0.025% MMS
wt
hog1Δ
ehd3Δ
wt
sty1Δ
snr1 Δ
237
YES
75 mM NaCl
0.75 mM H2O2
0.010% MMS

Figure 7.5

Table 7.1: StressNet and interaction confidence scores

ORF_A	ORF_B	Gene_A	Gene_B	Confidence Score
SPAC1002.17C	SPAC1002.17C	URG2	URG2	0.9132
SPAC10F6.11C	SPAC23H3.06	ATG17	APL6	0.8545
SPAC1142.06	SPAC14C4.05C	GET3	HEH2	0.8958
SPAC1142.06	SPAC19A8.14	GET3	SPAC19A8.14	0.9139
SPAC1142.06	SPBC4C3.06	GET3	SYP1	0.9338
SPAC11E3.08C	SPBC651.10	NSE6	NSE5	0.9501
SPAC12B10.02C	SPAC14C4.05C	SPAC12B10.02C	HEH2	0.8522
SPAC13G7.02C	SPBC3B9.01	SSA1	SPBC3B9.01	0.9322
SPAC14C4.05C	SPAC17G6.09	HEH2	SEC62	0.8958
SPAC14C4.05C	SPAC22F8.01	HEH2	SPAC22F8.01	0.8958
SPAC14C4.05C	SPAC23A1.02C	HEH2	SPAC23A1.02C	0.8804
SPAC14C4.05C	SPAC26A3.16	HEH2	DPH1	0.8516
SPAC14C4.05C	SPAC4G8.10	HEH2	GOS1	0.9176
SPAC14C4.05C	SPAC644.13C	HEH2	SPAC644.13C	0.8096
SPAC14C4.05C	SPAC688.04C	HEH2	GST3	0.8958
SPAC14C4.05C	SPAC6F12.04	HEH2	SPAC6F12.04	0.8666
SPAC14C4.05C	SPAC824.08	HEH2	GDA1	0.8814
SPAC14C4.05C	SPBC119.09C	HEH2	SPBC119.09C	0.9179
SPAC14C4.05C	SPBC12D12.01	HEH2	SAD1	0.8524
SPAC14C4.05C	SPBC146.04	HEH2	SPBC146.04	0.8547
SPAC14C4.05C	SPBC365.12C	HEH2	ISH1	0.8433
SPAC14C4.05C	SPBC428.14	HEH2	SPBC428.14	0.8588
SPAC14C4.05C	SPBC582.03	HEH2	CDC13	0.8270
SPAC14C4.05C	SPBC646.05C	HEH2	ERG9	0.9453
SPAC1565.04C	SPBC1D7.05	STE4	BYR2	0.8529
SPAC15A10.03C	SPAC644.14C	RHP54	RHP51	0.9491
SPAC16A10.05C	SPAC589.08C	DAD1	DAM1	0.8522
SPAC17A5.10	SPAC17A5.10	SPAC17A5.10	SPAC17A5.10	0.8606
SPAC17A5.10	SPAC328.04	SPAC17A5.10	SPAC328.04	0.7981
SPAC17A5.10	SPAC821.07C	SPAC17A5.10	MOC3	0.8522
SPAC17A5.10	SPBC1734.06	SPAC17A5.10	RHP18	0.8606
SPAC17D4.03C	SPBC16E9.14C	CIS4	ZRG17	0.9403
SPAC17G6.09	SPBC12D12.01	SEC62	SAD1	0.9286
SPAC17G6.14C	SPAC17G6.14C	UAP56	UAP56	0.9532
SPAC17G6.14C	SPCC31H12.03C	UAP56	SPCC31H12.03C	0.8791
SPAC17G8.10C	SPAC24B11.06C	DMA1	STY1	0.8132
SPAC17G8.10C	SPAC4H3.11C	DMA1	PPC89	0.7932
SPAC17G8.10C	SPAC644.14C	DMA1	RHP51	0.8958
SPAC17G8.10C	SPAC6F12.04	DMA1	SPAC6F12.04	0.8554
SPAC17G8.10C	SPBC12D12.01	DMA1	SAD1	0.8149
SPAC17G8.10C	SPBC13E7.08C	DMA1	SPBC13E7.08C	0.8155
SPAC17G8.10C	SPBC29A3.16	DMA1	RRS1	0.8958
SPAC17G8.10C	SPBC317.01	DMA1	MBX2	0.8522
SPAC17G8.10C	SPBC56F2.07C	DMA1	SPBC56F2.07C	0.7932
SPAC17H9.04C	SPAC821.07C	SPAC17H9.04C	MOC3	0.9523
SPAC1805.16C	SPAC1805.16C	SPAC1805.16C	SPAC1805.16C	0.8100
SPAC1834.11C	SPAC1834.11C	SEC18	SEC18	0.9306
SPAC1834.11C	SPAC227.13C	SEC18	ISU1	0.9524
SPAC1834.11C	SPAC29B12.06C	SEC18	RCD1	0.8792
SPAC19A8.07C	SPCC757.09C	SPAC19A8.07C	RNC1	0.8316
SPAC19A8.10	SPBC1734.06	RFP1	RHP18	0.9562
SPAC19A8.10	SPBC3D6.11C	RFP1	SLX8	0.9029

SPAC19D5.01	SPAC24B11.06C	PYP2	STY1	0.9328
SPAC19G12.03	SPAC19G12.03	CDA1	CDA1	0.9212
SPAC19G12.04	SPAC19G12.04	SPAC19G12.04	SPAC19G12.04	0.9105
SPAC1D4.11C	SPBC1685.01	LKH1	PMP1	0.8591
SPAC1D4.13	SPAC31G5.09C	BYR1	SPK1	0.8097
SPAC1D4.13	SPAC821.07C	BYR1	MOC3	0.8958
SPAC1D4.13	SPAC9E9.10C	BYR1	CBH1	0.8958
SPAC1D4.13	SPBC14F5.12C	BYR1	CBH2	0.8958
SPAC1D4.13	SPBC1D7.05	BYR1	BYR2	0.8958
SPAC1D4.13	SPBC56F2.07C	BYR1	SPBC56F2.07C	0.8958
SPAC1F3.02C	SPBC543.07	MKH1	PEK1	0.9627
SPAC1F8.07C	SPAC1F8.07C	SPAC1F8.07C	SPAC1F8.07C	0.9464
SPAC20H4.07	SPAC30D11.10	RHP57	RAD22	0.9212
SPAC20H4.08	SPAC2F7.02C	SPAC20H4.08	SPAC2F7.02C	0.8370
SPAC222.08C	SPAC222.08C	SPAC222.08C	SPAC222.08C	0.9542
SPAC222.08C	SPAC29B12.04	SPAC222.08C	SNZ1	0.8376
SPAC222.11	SPAC26H5.09C	HEM13	SPAC26H5.09C	0.8100
SPAC227.06	SPAC644.13C	SPAC227.06	SPAC644.13C	0.8522
SPAC227.13C	SPAC644.14C	ISU1	RHP51	0.9501
SPAC227.13C	SPBC2D10.11C	ISU1	NAP2	0.9406
SPAC227.13C	SPCC162.08C	ISU1	NUP211	0.9464
SPAC227.18	SPBC725.07	LYS3	PEX5	0.9380
SPAC22F8.08	SPBC26H8.01	SEC24	THI2	0.8587
SPAC23A1.14C	SPAC23A1.14C	SPAC23A1.14C	SPAC23A1.14C	0.8919
SPAC23A1.14C	SPBC725.07	SPAC23A1.14C	PEX5	0.8853
SPAC23A1.15C	SPBC691.02C	SEC20	SPBC691.02C	0.9370
SPAC23D3.06C	SPBC31E1.05	NUP146	GLE1	0.9328
SPAC24B11.06C	SPAC24B11.06C	STY1	STY1	0.8027
SPAC24B11.06C	SPBC29B5.01	STY1	ATF1	0.7991
SPAC24B11.06C	SPBC2D10.09	STY1	SPBC2D10.09	0.7998
SPAC24B11.06C	SPBC409.07C	STY1	WIS1	0.7657
SPAC24B11.06C	SPCC4F11.02	STY1	PTC1	0.8225
SPAC24C9.14	SPAC343.09	OTU1	UBX3	0.8082
SPAC24C9.14	SPBC119.05C	OTU1	SPBC119.05C	0.8958
SPAC25G10.08	SPACUNK12.01	SPAC25G10.08	SPACUNK12.01	0.9393
SPAC26A3.16	SPAC26A3.16	DPH1	DPH1	0.9139
SPAC26A3.16	SPAC3C7.12	DPH1	TIP1	0.9475
SPAC26A3.16	SPBC27.01C	DPH1	SPBC27.01C	0.8055
SPAC26H5.05	SPBC32F12.08C	SPAC26H5.05	DUO1	0.8799
SPAC27D7.03C	SPBC19C2.05	MEI2	RAN1	0.9491
SPAC27D7.04	SPAC27D7.04	OMT2	OMT2	0.9328
SPAC27D7.04	SPAC29B12.06C	OMT2	RCD1	0.8351
SPAC27D7.04	SPCC962.03C	OMT2	CUT15	0.8217
SPAC29B12.04	SPAC29B12.04	SNZ1	SNZ1	0.9088
SPAC29B12.04	SPBC119.04	SNZ1	MEI3	0.8958
SPAC2C4.15C	SPAC2C4.15C	UBX2	UBX2	0.8626
SPAC2C4.15C	SPAC343.09	UBX2	UBX3	0.8290
SPAC30D11.10	SPAC30D11.10	RAD22	RAD22	0.8433
SPAC30D11.10	SPAC3C7.03C	RAD22	RHP55	0.9571
SPAC30D11.10	SPAC644.14C	RAD22	RHP51	0.8575
SPAC31A2.11C	SPAC31A2.11C	CUF1	CUF1	0.9380
SPAC328.04	SPAC328.04	SPAC328.04	SPAC328.04	0.9491
SPAC328.04	SPAC3C7.02C	SPAC328.04	SPAC3C7.02C	0.9501
SPAC328.04	SPBC582.03	SPAC328.04	CDC13	0.8958
SPAC343.09	SPBC428.05C	UBX3	ARG12	0.8958

SPAC343.11C	SPAC637.12C	MSC1	MST1	0.9491
SPAC3A12.10	SPBC2D10.11C	RPL2001	NAP2	0.8370
SPAC3A12.12	SPAC3C7.02C	ATP11	SPAC3C7.02C	0.8958
SPAC3A12.12	SPAC644.14C	ATP11	RHP51	0.8155
SPAC3A12.12	SPAC7D4.04	ATP11	TAF1	0.7961
SPAC3A12.12	SPACUNK12.01	ATP11	SPACUNK12.01	0.8606
SPAC3A12.12	SPBC2D10.11C	ATP11	NAP2	0.8958
SPAC3C7.02C	SPBC1347.10	SPAC3C7.02C	CDC23	0.8958
SPAC3C7.02C	SPBC32F12.08C	SPAC3C7.02C	DUO1	0.8958
SPAC3C7.12	SPBC1604.20C	TIP1	TEA2	0.9229
SPAC3C7.12	SPCC1223.06	TIP1	TEA1	0.9447
SPAC3H5.05C	SPCC830.11C	RPS1401	SPCC830.11C	0.9212
SPAC3H5.10	SPBC2D10.11C	RPL3202	NAP2	0.9369
SPAC4H3.11C	SPBC582.03	PPC89	CDC13	0.8958
SPAC589.08C	SPAC8C9.17C	DAM1	SPC34	0.9491
SPAC589.08C	SPBC32F12.08C	DAM1	DUO1	0.9491
SPAC5H10.09C	SPAC5H10.09C	SPAC5H10.09C	SPAC5H10.09C	0.8919
SPAC637.12C	SPCC830.05C	MST1	EPL1	0.9441
SPAC644.14C	SPAC644.14C	RHP51	RHP51	0.8861
SPAC644.14C	SPBC119.04	RHP51	MEI3	0.8958
SPAC644.14C	SPBC1347.10	RHP51	CDC23	0.8341
SPAC644.14C	SPBC1921.03C	RHP51	MEX67	0.9129
SPAC644.14C	SPBC19C2.05	RHP51	RAN1	0.9292
SPAC644.14C	SPBC28F2.07	RHP51	SFR1	0.8245
SPAC644.14C	SPBC2D10.09	RHP51	SPBC2D10.09	0.8958
SPAC644.14C	SPBC317.01	RHP51	MBX2	0.8958
SPAC644.14C	SPBC32F12.08C	RHP51	DUO1	0.8928
SPAC644.14C	SPBC582.03	RHP51	CDC13	0.9501
SPAC688.11	SPAC688.11	END4	END4	0.9399
SPAC7D4.04	SPAC7D4.04	TAF1	TAF1	0.9486
SPAC7D4.04	SPBC1347.10	TAF1	CDC23	0.8556
SPAC7D4.04	SPBC16A3.11	TAF1	ESO1	0.8919
SPAC7D4.04	SPBC1718.07C	TAF1	ZFS1	0.8683
SPAC7D4.04	SPBC32F12.08C	TAF1	DUO1	0.9161
SPAC7D4.04	SPBC3B8.11	TAF1	RRN6	0.8523
SPAC7D4.04	SPBC839.07	TAF1	IBP1	0.9030
SPAC7D4.04	SPCC364.02C	TAF1	BIS1	0.8958
SPAC7D4.04	SPCC4F11.02	TAF1	PTC1	0.9225
SPAC7D4.04	SPCC548.05C	TAF1	SPCC548.05C	0.8064
SPAC806.06C	SPAC806.06C	SPAC806.06C	SPAC806.06C	0.9212
SPAC806.07	SPAC806.07	NDK1	NDK1	0.9238
SPAC821.07C	SPAC821.07C	MOC3	MOC3	0.9362
SPAC821.07C	SPAC9E9.10C	MOC3	CBH1	0.8958
SPAC821.07C	SPBC19C2.05	MOC3	RAN1	0.8341
SPAC821.07C	SPBC21B10.05C	MOC3	POP3	0.8708
SPAC821.07C	SPBC582.03	MOC3	CDC13	0.9478
SPAC890.02C	SPCC895.07	ALP7	ALP14	0.8347
SPAC8C9.03	SPAC8C9.03	CGS1	CGS1	0.8376
SPAC8C9.17C	SPAC8C9.17C	SPC34	SPC34	0.8958
SPAC8E11.11	SPCC16C4.13C	SPAC8E11.11	RPL1201	0.8354
SPAC959.10	SPAC959.10	SEN15	SEN15	0.9226
SPAC9E9.10C	SPAC9E9.10C	CBH1	CBH1	0.8881
SPAC9E9.10C	SPBC2D10.09	CBH1	SPBC2D10.09	0.9338
SPAC9G1.02	SPAC9G1.02	WIS4	WIS4	0.8958
SPAC9G1.02	SPBC725.02	WIS4	MPR1	0.9491

SPAC9G1.02	SPBC887.10	WIS4	MCS4	0.9417
SPACUNK12.01	SPBC19G7.15	SPACUNK12.01	NUP44	0.8958
SPACUNK12.01	SPBC337.13C	SPACUNK12.01	GTR1	0.9362
SPACUNK12.01	SPBC839.07	SPACUNK12.01	IBP1	0.8958
SPAP27G11.09C	SPAP27G11.09C	SPAP27G11.09C	SPAP27G11.09C	0.8559
SPAP8A3.06	SPBC146.07	SPAP8A3.06	PRP2	0.8231
SPAPB1E7.12	SPBC2D10.11C	RPS602	NAP2	0.9485
SPBC1105.04C	SPBC1105.04C	CBP1	CBP1	0.9491
SPBC119.04	SPBC19C2.05	MEI3	RAN1	0.9491
SPBC11B10.10C	SPBC2D10.11C	PHT1	NAP2	0.9476
SPBC11B10.10C	SPCC830.11C	PHT1	SPCC830.11C	0.8958
SPBC12C2.07C	SPBC12C2.07C	SPBC12C2.07C	SPBC12C2.07C	0.9212
SPBC12D12.01	SPBC12D12.01	SAD1	SAD1	0.9416
SPBC12D12.01	SPBC582.03	SAD1	CDC13	0.8881
SPBC1347.10	SPBC211.04C	CDC23	MCM6	0.9478
SPBC1347.10	SPCC162.08C	CDC23	NUP211	0.9523
SPBC13G1.03C	SPCC338.13	PEX14	COG4	0.9107
SPBC14F5.05C	SPBC14F5.05C	SAM1	SAM1	0.9088
SPBC14F5.09C	SPBC14F5.09C	ADE8	ADE8	0.9226
SPBC14F5.12C	SPBC14F5.12C	CBH2	CBH2	0.9559
SPBC14F5.12C	SPBC354.03	CBH2	SWD3	0.8732
SPBC16E9.01C	SPBC26H8.06	PHP4	GRX4	0.9491
SPBC1706.01	SPCC1223.06	TEA4	TEA1	0.8692
SPBC1734.06	SPBC1734.06	RHP18	RHP18	0.9340
SPBC1773.05C	SPBC1773.05C	TMS1	TMS1	0.9427
SPBC211.02C	SPCC364.02C	CWF3	BIS1	0.9429
SPBC215.14C	SPBC651.05C	VPS20	DOT2	0.8522
SPBC215.15	SPBC8D2.20C	SEC13	SEC31	0.7981
SPBC21C3.10C	SPBC21C3.10C	SPBC21C3.10C	SPBC21C3.10C	0.9229
SPBC23E6.10C	SPBC23E6.10C	SPBC23E6.10C	SPBC23E6.10C	0.9572
SPBC25H2.09	SPCC4F11.02	SPBC25H2.09	PTC1	0.8958
SPBC26H8.01	SPBC26H8.01	THI2	THI2	0.9427
SPBC28F2.01C	SPBC28F2.01C	SPBC28F2.01C	SPBC28F2.01C	0.8438
SPBC29A3.16	SPCC4F11.02	RRS1	PTC1	0.8958
SPBC2D10.09	SPBC725.04	SPBC2D10.09	SPBC725.04	0.8648
SPBC2D10.11C	SPBC2D10.11C	NAP2	NAP2	0.9017
SPBC2D10.11C	SPBC32F12.08C	NAP2	DUO1	0.8958
SPBC2D10.11C	SPBC582.03	NAP2	CDC13	0.8867
SPBC2D10.11C	SPCC1672.07	NAP2	SPCC1672.07	0.8649
SPBC2D10.11C	SPCC663.04	NAP2	RPL39	0.9075
SPBC30D10.05C	SPBC30D10.05C	SPBC30D10.05C	SPBC30D10.05C	0.9108
SPBC317.01	SPBC32F12.08C	MBX2	DUO1	0.8958
SPBC317.01	SPBC354.03	MBX2	SWD3	0.9488
SPBC31F10.11C	SPCC1739.11C	CWF4	CDC11	0.9372
SPBC32F12.08C	SPCC162.08C	DUO1	NUP211	0.8370
SPBC32H8.12C	SPBC32H8.12C	ACT1	ACT1	0.9462
SPBC342.05	SPBC582.03	CRB2	CDC13	0.9464
SPBC342.05	SPCC4F11.02	CRB2	PTC1	0.9406
SPBC354.03	SPBC685.09	SWD3	ORC2	0.9399
SPBC3E7.02C	SPBC3E7.02C	HSP16	HSP16	0.9471
SPBC3E7.02C	SPBC543.07	HSP16	PEK1	0.8958
SPBC3F6.03	SPBC3F6.03	TRR1	TRR1	0.9427
SPBC409.07C	SPBC725.02	WIS1	MPR1	0.9427
SPBC428.05C	SPBC428.05C	ARG12	ARG12	0.9212
SPBC56F2.07C	SPBC56F2.07C	SPBC56F2.07C	SPBC56F2.07C	0.8448

SPBC582.03	SPBC685.09	CDC13	ORC2	0.8567
SPBC582.03	SPCC1739.11C	CDC13	CDC11	0.9107
SPBC725.02	SPBC887.10	MPR1	MCS4	0.9491
SPBC725.02	SPCC74.06	MPR1	MAK3	0.9328
SPBC725.13C	SPBP4H10.21C	PSF2	SLD5	0.8426
SPBC800.03	SPBC800.03	CLR3	CLR3	0.9491
SPBC887.10	SPBC887.10	MCS4	MCS4	0.9429
SPCC1223.06	SPCC895.05	TEA1	FOR3	0.9501
SPCC1322.12C	SPCC1322.12C	BUB1	BUB1	0.9464
SPCC1322.12C	SPCC895.02	BUB1	SPCC895.02	0.9429
SPCC162.08C	SPCC364.02C	NUP211	BIS1	0.9417
SPCC1739.11C	SPCC1739.11C	CDC11	CDC11	0.9493
SPCC18.07	SPCC330.13	RPC53	RPC37	0.9244
SPCC24B10.13	SPCC4F11.02	SKB5	PTC1	0.9212
SPCC330.05C	SPCC330.05C	URA4	URA4	0.9088
SPCC4B3.06C	SPCC4B3.06C	SPCC4B3.06C	SPCC4B3.06C	0.9066
SPCC4G3.17	SPCC4G3.17	SPCC4G3.17	SPCC4G3.17	0.9439
SPCC757.09C	SPCC757.09C	RNC1	RNC1	0.9043

Table 7.2: Positive reference set

ORF_A	ORF_B	Gene_A	Gene_B
SPAC926.04C	SPCC613.04C	HSP90	RNG3
SPAC1D4.13	SPAC31G5.09C	BYR1	SPK1
SPAC30D11.10	SPAC30D11.10	RAD22	RAD22
SPAC30D11.10	SPAC644.14C	RAD22	RHP51
SPBC1604.20C	SPBC1604.20C	TEA2	TEA2
SPBC1347.10	SPBC211.04C	CDC23	MCM6
SPAC27F1.09C	SPBC146.07	PRP10	PRP2
SPBC3E7.02C	SPBC3E7.02C	HSP16	HSP16
SPAC3C7.12	SPBC1604.20C	TIP1	TEA2
SPAC3C7.12	SPCC1223.06	TIP1	TEA1
SPBC244.01C	SPCC1739.11C	SID4	CDC11
SPBC244.01C	SPBC244.01C	SID4	SID4
SPBC146.03C	SPBP4H10.06C	CUT3	CUT14
SPAC9G1.02	SPBC409.07C	WIS4	WIS1
SPAC16A10.07C	SPBC1778.02	TAZ1	RAP1
SPAC16E8.09	SPAC22H10.07	SCD1	SCD2
SPAC890.02C	SPCC895.07	ALP7	ALP14
SPAC644.06C	SPCC18B5.03	CDR1	WEE1
SPAC27D7.03C	SPAC8E11.02C	MEI2	RAD24
SPAC27D7.03C	SPBC19C2.05	MEI2	RAN1
SPAC23G3.01	SPBC31F10.09C	RPB2	NUT2
SPBC14F5.08	SPBC31F10.09C	MED7	NUT2
SPBC1105.06	SPBC31F10.09C	PMC4	NUT2
SPBC146.07	SPBC530.14C	PRP2	DSK1
SPAC8E11.03C	SPAC8E11.03C	DMC1	DMC1
SPAC17H9.09C	SPBC1D7.05	RAS1	BYR2
SPAC15A10.03C	SPAC644.14C	RHP54	RHP51
SPAC19A8.12	SPBC3B9.21	DCP2	DCP1
SPAC2E1P5.04C	SPAPB1A10.04C	CWG2	CWP1
SPAC11E3.08C	SPBC651.10	NSE6	NSE5
SPAC24H6.05	SPCC1322.08	CDC25	SRK1
SPBC31F10.09C	SPCC1020.04C	NUT2	RPB6
SPAC637.07	SPBC646.09C	MOE1	INT6
SPAC24B11.06C	SPBC29B5.01	STY1	ATF1
SPAC24B11.06C	SPBC409.07C	STY1	WIS1
SPAP8A3.06	SPBC146.07	SPAP8A3.06	PRP2
SPBC1706.01	SPCC1223.06	TEA4	TEA1
SPAC19D5.01	SPAC24B11.06C	PYP2	STY1
SPAC3C7.03C	SPAC644.14C	RHP55	RHP51
SPAC1F7.04	SPBC12D12.04C	RHO1	PCK2
SPAC19A8.10	SPBC3D6.11C	RFP1	SLX8
SPAC19A8.10	SPBC1921.02	RFP1	RAD60
SPAC1565.04C	SPBC1D7.05	STE4	BYR2
SPAC1142.03C	SPAC664.01C	SWI2	SWI6
SPAC644.14C	SPAC644.14C	RHP51	RHP51
SPBC1105.17	SPBC409.04C	CNP1	MIS12

SPBC725.02	SPBC887.10	MPR1	MCS4
SPAC20H4.07	SPAC3C7.03C	RHP57	RHP55
SPBC1D7.05	SPBC1D7.05	BYR2	BYR2
SPBC216.06C	SPBC30D10.04	SWI1	SWI3
SPAC23C11.16	SPCC4B3.15	PLO1	MID1
SPBC1604.14C	SPBC1604.14C	SHK1	SHK1
SPAC18G6.15	SPAC3C7.12	MAL3	TIP1
SPAC18G6.15	SPAC18G6.15	MAL3	MAL3

Table 7.3: Random reference set

ORF_A	ORF_B	Gene_A	Gene_B
SPCC1020.11C	SPCC24B10.13	SPCC1020.11C	SKB5
SPAC13F5.05	SPBC4C3.06	SPAC13F5.05	SYP1
SPAC227.18	SPAC959.10	LYS3	SEN15
SPAC222.08C	SPAC688.11	SPAC222.08C	END4
SPAC20G8.02	SPAC23H3.05C	SPAC20G8.02	SWD1
SPBC26H8.06	SPCC1223.06	GRX4	TEA1
SPAC3C7.03C	SPBC691.03C	RHP55	APL3
SPAC8C9.03	SPBC3B9.09	CGS1	VPS36
SPAC16E8.17C	SPBC2D10.04	SPAC16E8.17C	SPBC2D10.04
SPAC23C4.06C	SPAC343.09	SPAC23C4.06C	UBX3
SPAC22A12.10	SPAC22F3.13	SPAC22A12.10	TSC1
SPBC15D4.04	SPCC4G3.13C	GPT2	SPCC4G3.13C
SPBC30D10.13C	SPBP4H10.06C	PDB1	CUT14
SPBC27.01C	SPBC31F10.11C	SPBC27.01C	CWF4
SPAC20H4.08	SPBC29A3.06	SPAC20H4.08	SPBC29A3.06
SPCC18.11C	SPCC4B3.06C	SDC1	SPCC4B3.06C
SPAC3G9.08	SPBC21H7.07C	PNG1	HIS5
SPAC1F7.04	SPBC3B9.01	RHO1	SPBC3B9.01
SPAPYUK71.02	SPBC119.16C	SPAPYUK71.02	SPBC119.16C
SPAC23C11.13C	SPAC2F7.02C	HPT1	SPAC2F7.02C
SPBC146.03C	SPBC428.03C	CUT3	PHO4
SPCC4F11.02	SPCC63.05	PTC1	SPCC63.05
SPAC1006.02	SPBC409.20C	ASA1	PSH3
SPAC17A5.17	SPCC74.06	SPAC17A5.17	MAK3
SPAC14C4.02C	SPAC1805.06C	SMC5	HEM2
SPAC23C4.06C	SPBC119.04	SPAC23C4.06C	MEI3
SPBC16A3.11	SPBC28F2.10C	ESO1	NGG1
SPAC1486.03C	SPBC18E5.05C	SPAC1486.03C	SPBC18E5.05C
SPAC23D3.11	SPAC821.10C	AYR1	SOD1
SPAC16E8.03	SPCC1281.07C	GNA1	SPCC1281.07C
SPAC977.15	SPBC543.09	SPAC977.15	SPBC543.09
SPAC31G5.09C	SPCP31B10.06	SPK1	MUG190
SPAC30D11.10	SPCC1235.03	RAD22	SPCC1235.03
SPAC1002.17C	SPCC1393.14	URG2	TEN1
SPAC17A2.09C	SPBC16A3.11	CSX1	ESO1
SPBC13G1.09	SPCC613.10	SPBC13G1.09	QCR2
SPAC3H5.10	SPBC21C3.04C	RPL3202	SPBC21C3.04C
SPAC1486.03C	SPBC2D10.04	SPAC1486.03C	SPBC2D10.04
SPAC6F12.01	SPBC31F10.11C	SPAC6F12.01	CWF4
SPBC1711.08	SPBC1815.01	SPBC1711.08	ENO101
SPAC2F7.11	SPAC823.15	NRD1	PPA1
SPAC22E12.08	SPBC3H7.13	RRN10	SPBC3H7.13
SPAC24H6.12C	SPAC637.14	UBA3	SPAC637.14

Table 7.4: Primers used.

Primer Name	Primer Sequences (5'-3')
Sty1-pNCH1472-Forward	AAGGAAAAAAGCGGCCGCATGGCAGAATTTATTCGTAC
Sty1-pNCH1472-Reverse	GGTGTCGACGGATTGCAGTTCATTATCCATG
snr1-pSGP73-Forward	AAGGAAAAAAGCGGCCGCATGGGATTGAAATTAAATATC
snr1-pSGP73-Reverse	GGAAGATCTCTATAAATAAGGATAAGTC
SpEhd3_Up_Fwd	CTTAAACAGCCTGATTTTGT
SpEhd3_Dn_Rev	AACTATCGTACGCACAGCTA
KanMX4-Fwd	TTAGCTTGCCTCGTCCCC
KanMX4-Rev	TTTCGACACTGGATGGCG
Sty1Del-Up_Fwd	TACAAGCAAACACCACAATC
Sty1Del-Up_Rev	TTAATTAACCCGGGGATCCGTTTATTCAAACTGGTTACAAAAAGGAC
PFA6a_Sty1_Fwd	TTGTAACCAGTTTGAATAAACGGATCCCCGGGTTAATTAA
PFA6a_Sty1_Rev	AGGCTTTATCTACAACTTGTGAATTCGAGCTCGTTTAAAC
Sty1Del-Dn_Fwd	GTTTAAACGAGCTCGAATTCACAAGTTGTAGATAAAGCCTTAAAAGTTGTTC
Sty1Del-Dn_Rev	ACACCACACTTGAAAATCGC
Sty1_Fwd	AATTGAGACGATTTGCAGTAAAAAC
Sty1_Rev	TAATACGCTTACGAGGATCAAAGAC

CHAPTER 8

A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human

In the following chapter, we expand our study of network evolution to the scale of the proteome. Using a proteome-wide fission yeast network, we identify several key mechanisms by which protein interactions have evolved from yeasts to human. I am a co-first author of the paper resulting from this chapter (Vo*, Das* et al Cell 2016, *=Equal contribution) and led all computational analyses. The first author of the paper, Tommy Vo, led the entire experimental effort. Michael Meyer, also a co-first author also made a significant contribution to several analyses in the paper.

8.1 ABSTRACT

Here, we present FissionNet, a proteome-wide binary protein interactome for *S. pombe*, comprising 2,278 high-quality interactions, of which ~50% were previously not reported in any species. FissionNet unravels previously unreported interactions implicated in processes such as gene silencing and pre-mRNA splicing. We developed a rigorous network comparison framework that accounts for assay sensitivity and specificity, revealing extensive species-specific network rewiring between fission yeast, budding yeast, and human. Surprisingly, although genes are better conserved between the yeasts, *S. pombe* interactions are significantly better conserved in human than in *S. cerevisiae*. Our framework also reveals that different modes of gene duplication influence the extent to which paralogous proteins are functionally repurposed. Finally, cross-species interactome mapping demonstrates that coevolution of interacting proteins is remarkably prevalent, a result with important implications for studying human disease in model organisms. Overall, FissionNet is a valuable resource for understanding protein functions and their evolution.

8.2 INTRODUCTION

Proteins function primarily by physically interacting with other proteins. Gain or loss of these interactions within an organism can modulate protein functions and disease states (Sahni et al., 2015; Wei et al., 2014). The importance of protein interactions to our understanding of fundamental biological processes has spurred the mapping of protein interactome networks for several organisms (Arabidopsis Interactome Mapping Consortium, 2011; Giot et al., 2003; Rolland et al., 2014; Stelzl et al., 2005; Yu et al., 2008). However, the budding yeast *Saccharomyces cerevisiae* remains the only eukaryotic organism for which a high-coverage binary protein interactome has been mapped by systematic interrogation of pairwise combinations of all proteins in triplicate (Yu et al., 2008). Here, we present FissionNet, a high-coverage proteome-wide protein interactome network generated for the fission yeast *Schizosaccharomyces pombe*.

We compared FissionNet with the only other proteome-scale eukaryotic interactomes available (>50% of all protein pairs screened), the interactome networks of *S. cerevisiae* and human. Surprisingly, we find that FissionNet is more similar to the human network than it is to that of *S. cerevisiae*. Furthermore, among interactions involving conserved proteins, there is significant species-specific rewiring that is not completely determined by overall sequence similarity of orthologs. Instead, we identify several other determinants of interaction conservation, including local network constraints and conservation of interacting protein domains. Also, by comparing FissionNet with the proteome-wide interactome of *S. cerevisiae*, we are able to ascertain how gene duplication events influence the process by which paralogs acquire novel functions.

S. pombe is an important model organism for studying fundamental biological processes such as RNA splicing, cell cycle regulation, RNA interference (RNAi), and centromeric maintenance, which are conserved in metazoans but divergent in budding yeast (Wood et al., 2002). We use FissionNet to unveil previously unreported protein associations between gene regulatory factors involved in pre-mRNA splicing and silencing of stress-response genes and at pericentromeric regions, illustrating the value of our network as a proteome-scale resource to understand biological processes.

8.3 RESULTS

A proteome-wide high-coverage binary protein interactome map of *S. pombe*

To generate a proteome-wide interactome network for *S. pombe*, which we call FissionNet, we systematically tested all pairwise combinations of proteins encoded by 4,989 *S. pombe* genes (corresponding to >99% of all *S. pombe* coding genes) using our high-quality yeast two-hybrid (Y2H) assay, the same pipeline that we used to generate the budding yeast and human interactome networks (Yu et al., 2008; Yu et al., 2011). Extensive screenings in triplicate (a total of ~75 million protein pairs) yielded 2,278 interactions between 1,305 proteins, of which 2,130 (93.5%) have not been previously reported in *S. pombe* (Figure 8.1A, downloadable from hint.yulab.org under *S. pombe* binary HQ corresponding to PMID: 26771498) (Das and Yu, 2012). Furthermore, FissionNet contains 1,034 interactions that have not been reported between orthologs in any other species before. Of these, 142 interactions involve *S. pombe* proteins that both have human orthologs, but at least one does not have a *S. cerevisiae* ortholog and, hence,

cannot be studied in *S. cerevisiae*. Thus, FissionNet provides a valuable repertoire of biological insights.

To assess the sensitivity and specificity of our Y2H assay (Yu et al., 2008), we constructed a positive reference set (PRS) consisting of 93 well-validated *S. pombe* interactions from the literature and a negative reference set (NRS) of 168 random *S. pombe* protein pairs that are not known to interact in the literature and whose orthologs in other species are also not known to interact (Table 8.1, see Materials and Methods). We performed Y2H and protein complementation assay (PCA) (Das et al., 2013; Yu et al., 2008) to test what fraction of the PRS, NRS, and a random sample of 220 FissionNet interactions can be detected using orthogonal methods (Figure 8.1B). We found that the detection rates of the PRS and FissionNet interactions are indistinguishable from each other and are significantly higher than that of the NRS (Figure 8.1B; >15% difference in detection rates between the PRS and NRS for both assays, $P<10^{-3}$, Z test). The robust validation rates of FissionNet interactions by an orthogonal assay confirm the high quality of the network. Furthermore, although it has been speculated that Y2H interactions involving proteins with many interaction partners (hubs) could be of low quality (Bader et al., 2004), we found that the validation rate by PCA of hub interactions is the same as the overall PCA validation rate for FissionNet (Figure 8.1B; P=0.34, Z test), confirming that FissionNet interactions involving hubs are of high quality.

Biological relationships between interacting proteins in FissionNet were assessed by measuring similarities in protein localization, functional annotations, and expression profiles (see Materials and Methods). We found that FissionNet interactions are significantly enriched for protein pairs that are co-localized, functionally similar, and encoded by coexpressed genes relative to random expectation (Figures 8.1C to 8.1E; P<0.05 in all three cases using a KS test

for coexpression and *Z* test for co-localization and functional similarity). Furthermore, the enrichment of these interactions for all three categories is similar to that of literature-curated binary interactions. These results confirm that FissionNet interactions are functionally relevant *in vivo*. We illustrate this by focusing on two previously unreported interactions: Tas3-Hhp1 and Atf1-Cid12, and their potential roles in gene silencing.

FissionNet provides insights into functions of proteins and interactions

The regulation of centromeric silencing is a well-conserved process in *S. pombe* and metazoans but is divergent from that in *S. cerevisiae* (Holoch and Moazed, 2015). FissionNet revealed a previously unidentified interaction between Tas3 and Hhp1 that we confirmed *in vivo* (Figures 8.1F and 8.1G). Tas3 is a component of the RNA-induced transcriptional silencing (RITS) complex that mediates gene silencing at *S. pombe* centromeres (Verdel et al., 2004). Hhp1 is a conserved mitotic checkpoint kinase (Johnson et al., 2013) not known to be involved in centromeric silencing. In *S. pombe* cells where the *ura4*$^{+}$ reporter gene was inserted at the centromere inner repeats of chromosome 1 (*imr1R*) (Verdel et al., 2004), we find that *hhp1Δ* confers loss of silencing at the centromere, similar to *tas3Δ* cells (Figure 8.1H). Furthermore, levels of endogenous centromeric transcripts are elevated in *hhp1Δ* cells. Moreover, loss of *hhp1* leads to a decrease in the dimethylation of histone 3 lysine 9 (H3K9me2) at the centromere. These results show that Hhp1 is involved in centromeric silencing.

We also identified a previously unreported interaction between the transcription factor Atf1 and the polyadenylation polymerase Cid12 (Figure 8.2A). Atf1 mediates transcriptional responses to stresses such as high temperatures (Shiozaki and Russell, 1996). At *S. pombe* centromeres, Cid12 is a core component of the RNA-directed RNA-polymerase complex

(RDRC) (Motamedi et al., 2004). The RDRC is responsible for generating double-stranded RNAs, a key step for Dcr1-dependent centromeric silencing. Interestingly, it has been reported that Dcr1 transcriptionally represses the Atf1-target genes *hsp16* and *hsp104* under non-stressed conditions (Woolcock et al., 2012).

Pull-down experiments confirm the interaction of Atf1 and Cid12 in *S. pombe* (Figure 8.2B), and *cid12Δ* cells grown under non-stressed conditions show elevated mRNA levels of *hsp16* and *hsp104* as compared to wild-type cells, similar to *dcr1Δ* cells (Figure 8.2C). Additionally, double mutant *cid12Δ dcr1Δ* cells do not exhibit more drastic transcript accumulation than the single deletion mutants, suggesting both genes function in the same pathway (Figure 8.2C). Together, these results suggest that Cid12 may be involved in repressing aberrant gene expression of Atf1-target genes.

Next, we identified two Cid12 mutations, lysine-213 to isoleucine ($Cid12^{K213I}$) and aspartic acid-260 to valine ($Cid12^{D260V}$), that disrupt the interaction of Cid12 with Atf1 while preserving interactions within the RDRC complex (Figure 8.2D; see Materials and Methods). Exogenous expression of wild-type Cid12 in *cid12Δ* cells enables the transcriptional repression of *hsp16* and *hsp104*. In stark contrast, neither mutant can repress gene expression (Figure 8.2E). The mutant phenotype is not due to complete loss of protein caused by destabilization because these Cid12 mutant proteins express in *S. pombe* cells. Furthermore, in *cid12Δ* cells where the $ura4^+$ reporter gene was inserted at the centromeric *imr1R*, we find that exogenous expression of either Cid12 wild-type or mutants equally permit the silencing of the $ura4^+$ reporter (Figure 8.2F). Thus, we show that Cid12 has dual roles in regulating the expression of heat-shock genes and the centromere. Importantly, the roles can be selectively uncoupled via specific disruption of the

Atf1-Cid12 interaction. These examples illustrate the usefulness of FissionNet as a resource to uncover areas of biological inquiry.

Comparative network analyses reveal species-specific conservation of interactions

High-quality protein interactome networks have previously been reported in budding yeast (Yu et al., 2008) and human (Rolland et al., 2014). A fundamental question, which can be addressed with FissionNet and these networks, is how protein-protein interactions have evolved and whether this trend mirrors gene-level evolution. From sequence-based phylogenetic analyses, the two yeasts are less divergent from each other than either yeast is from human (Figure 8.3A) (Sipiczki, 2000). Additionally, the two yeasts share a greater fraction of protein-coding genes than either yeast does with human (Figures 8.4A and 8.4B).

To calculate interaction conservation, we considered only those interactions that have the potential to be conserved, *i.e.*, the two interacting proteins in the reference species have orthologs in the other species. However, directly calculating the overlap between sets of interactions obtained from the literature would be erroneous because currently available interactomes are incomplete and are derived from assays with varied and often unreported false positive and false negative rates (Yu et al., 2008). Therefore, to accurately estimate the underlying interaction conservation fractions, we required interactomes of all species to be derived from the same experimental assay. Since interactomes in budding yeast (Yu et al., 2008) and human (Rolland et al., 2014) have been generated using our version of Y2H (Figure 8.4C and 8.4D), we were able to compare FissionNet to these interactome networks to measure the observed extent of interaction conservation. We developed a rigorous Bayesian framework that incorporates both the false positive and false negative rates of our Y2H assay to estimate the underlying interaction

conservation fraction from the observed fraction for each pair of species (see Materials and Methods). Surprisingly, we find that interaction conservation follows a completely different trend from gene conservation (Figures 8.3B, 8.4E, and 8.4F). While only ~40% of *S. pombe* interactions are conserved in *S. cerevisiae* (of the 1,331 interactions where both proteins have *S. cerevisiae* orthologs and were pairwise retested using our Y2H assay), ~65% of *S. pombe* interactions are conserved in human (of the 652 interactions where both proteins have human orthologs and were pairwise retested using our Y2H assay) (Figure 8.3B; $P=1.4\times10^{-4}$, *Z* test). However, when using budding yeast as the reference species, the fraction of conserved interactions is as high in fission yeast as in human, comparable to the fraction conserved between fission yeast and human (Figure 8.3B). We were able to recapitulate these results using interaction datasets generated by other assays (Figures 8.4G to 8.4I; >1.5 fold difference between fission yeast interactions conserved in budding yeast and human; $P<10^{-3}$ in all cases, *Z* test; see Materials and Methods). Thus, our results suggest that a large fraction of interactions are conserved between human and *S. pombe*, but have been lost specifically in the *S. cerevisi*ae lineage.

One possible explanation for these surprising results is that fission yeast proteins that are conserved in human could have higher overall sequence similarity than those that are conserved in budding yeast. However, we find that proteins in interactions that have the potential to be conserved based on orthology are actually slightly more similar in sequence between the two yeasts than between *S. pombe* and human (Figure 8.3C; $P<10^{-5}$, *U* test; see Materials and Methods).

Another possibility is that the observed difference primarily arises from interactions involving proteins that are conserved between fission yeast and human but lost in budding yeast.

To test this, we first focused on proteins that are conserved in all 3 species. We still find that ~20% more interactions are conserved between *S. pombe* and human as compared to between the two yeasts (Figures 8.3D, 8.4J, and 8.4K; $P<0.05$, *Z* test).

We next explored the conservation of interactions involved in various biological processes as defined by the Gene Ontology (GO) (Ashburner et al., 2000). We find wide variation in species-specific interaction conservation among different processes (Figures 8.3E and 8.4L to 8.4N). We show that *S. pombe* interactions are more conserved in human than in *S. cerevisiae* for 10 out of 13 GO Slim categories containing ≥50 interactions (Figure 8.3E; $P<0.05$, as marked, *Z* test). The same trend is observed with GO Slim categories containing ≥30 or ≥75 interactions (Figures 8.4L and 8.4N). Some of these categories, such as "chromosomal organization", "chromosome segregation", and "cell cycle", are far better conserved in human than in *S. cerevisiae*, and accordingly *S. pombe* has been used as a model organism for studying these processes (Wood et al., 2002). Furthermore, considering GO Slim categories that are well conserved in all three species (using cutoffs of ≥50, 100, and 200 genes annotated per species), we find that the conservation of *S. pombe* interactions in these core biological processes is also higher in human than in *S. cerevisiae* (Figures 8.3F and 8.4O; $P<10^{-3}$, *Z* test). Overall, these results suggest that insights gained from FissionNet may be widely applicable to the study of human biology across many important cellular processes.

We validated three cases of previously unreported functional conservation between fission yeast and human proteins. Uncharacterized *S. pombe* factors Srrm1, SPAC30D11.14C, and SPAC1952.06C interact with known splice factors Srp1, Usp104, and Cwf15, respectively (Figure 8.3G**)**. Although these proteins have no orthologs in *S. cerevisiae*, they are orthologous to human SRRM1, KIAA0907, and CTNNBL1, respectively. Interestingly, all three human

orthologs have been implicated in pre-mRNA splicing or were found to associate with spliceosomal factors in human (Blencowe et al., 1998; Hegele et al., 2012; Rolland et al., 2014). We used DNA microarrays to measure changes in the splicing of every known intron in the *S. pombe* deletion mutants. The loss of *srrm1*, *SPAC30D11.14C*, or *SPAC1952.06C* results in widespread splicing defects, confirming the roles for these proteins in the splicing pathway (Figure 8.3H). Moreover, Srrm1 and Srp1 share many gene targets, suggesting that the interacting proteins are functionally related (Figure 8.4P). Notably, an analysis of the introns whose splicing is affected by *srrm1* deletion shows a strong enrichment for introns with weak splice site signals (Figure 8.4Q). This is consistent with previous findings that human *SRRM1* affects splice site selection by binding to exonic splicing enhancers and facilitating interactions between spliceosomal proteins (Blencowe et al., 1998). These results highlight the utility of FissionNet to reveal proteins that are functionally conserved between *S. pombe* and human.

Determinants of interaction conservation

Previous studies have shown that increased protein sequence similarity facilitates conservation of protein interactions (Matthews et al., 2001). Indeed, we also found a positive correlation between sequence similarity of proteins and the fraction of their associated interactions conserved between *S. pombe* and human or *S. cerevisiae*, demonstrating a proteome-scale dependence of protein sequence and function (Figure 8.5A; $R^2_{S.p\text{-}H.s}$=0.948 and $R^2_{S.p\text{-}S.c}$=0.976). However, protein interaction conservation is not completely dependent on overall sequence similarity, as we find many instances of conserved interactions involving proteins with low overall sequence similarity (<40%) with their orthologs (Figure 8.5A; 40% and 13% of 116 interactions in human and 196 interactions in *S. cerevisiae*, respectively). To investigate whether certain highly

conserved domains in these proteins play an important role in interaction conservation, we inferred protein interaction domains from co-crystal structures of 124 human interactions conserved in *S. pombe* and 293 conserved in *S. cerevisiae*. We find that the sequence similarity within protein interaction domains tends to be higher than in other domains for interactions conserved between fission yeast and human (Figure 8.5B; 7.0% higher, P=0.012, U test). For instance, the human DR1-DRAP1 heterodimer is orthologous to the protein pair Ncb2 and Dpb3 in *S. pombe.* While the overall sequence similarity of the orthologs is quite low (0.58 and 0.51, respectively), the interaction is conserved in fission yeast. Moreover, we also find that the proteins can interact with the orthologs of their native interaction partner (Figure 8.5C). Based on a crystal structure of the human DR1-DRAP1 complex, we were able to determine the interaction domains of these proteins (Figure 8.5D) (Kamada et al., 2001). The sequence similarity within these domains in DR1 and DRAP1 with their fission yeast orthologs is 0.78 and 0.80, respectively, while the conservation outside of these interaction domains is only 0.45 and 0.38. Thus, the basis for this high degree of functional conservation is likely dependent on the interaction domains.

Strikingly, interaction conservation is nearly three times higher between *S. pombe* and human than between the two yeasts at low levels of overall sequence similarity (Figure 8.5A; at <40% similarity, P=0.030, Z test). As sequence similarity approaches 100%, interaction conservation converges. Therefore, for the vast majority of interactions corresponding to proteins with lower sequence similarity to their orthologs, our results strongly suggest that species-specific factors, independent of overall protein sequence similarity, influence conservation of protein-protein interactions.

We then sought to explore other factors that could explain the basis of interaction conservation. First, we used ClusterOne (Nepusz et al., 2012) to detect topological protein clusters in FissionNet (see Materials and Methods). We find that intra-cluster FissionNet interactions are >3 times more likely to be conserved in both budding yeast and human than inter-cluster interactions (Figures 8.5E; $P<0.05$ for both organisms, Z test). Next, we examined biological processes defined by GO (Ashburner et al., 2000) and observed the same trend (Figures 8.5F; $P<10^{-3}$ for both organisms, Z test). Using genetic interactions, it has been earlier hypothesized that while individual functional modules are conserved, inter-modular connectivity could be rewired across evolution (Roguev et al., 2008). In this study, we provide direct molecular level evidence on a proteome scale that while interactions within modules tend to be conserved across evolution, the cross-talk among these modules changes significantly from one species to another.

Gene duplication shapes the functional fate of paralogs

Gene duplication has long been known as a major source of evolutionary novelty (Arabidopsis Interactome Mapping Consortium, 2011). Previous studies have found that a whole-genome duplication (WGD) event leads to more functional redundancy between paralogous proteins than small-scale duplications (SSDs) (Arabidopsis Interactome Mapping Consortium, 2011; Hakes et al., 2007). However, there has been much debate in the literature regarding the relative extents of sub-functionalization and neo-functionalization for diverged paralogs (Gibson and Goldberg, 2009; He and Zhang, 2005). Previous studies on functional evolution of paralogs often used interaction datasets from the literature, which, as mentioned earlier, suffer from detection and completeness biases (Yu et al., 2008). Until now, it has not been possible to measure the extent

of sub-functionalization and neo-functionalization using an unbiased framework because there was only one proteome-wide high-coverage binary protein interactome available, that of *S. cerevisiae*. Here, we compare the unbiased proteome-wide networks of *S. pombe* (FissionNet) and *S. cerevisiae* (CCSB-YI1) (Yu et al., 2008) that we produced using the same Y2H assay to analyze these two types of functional divergence.

We first examined the extent to which interactions in *S. pombe* and *S. cerevisiae* tend to be conserved across species but not shared between within-species paralogs (sub-functionalized) (Figure 8.6A; see Materials and Methods). We find that fission yeast paralog pairs tend to undergo more sub-functionalization than budding yeast paralog pairs (Figure 8.6B; difference in log odds ratio=2.8 using 1,762 fission yeast paralog pairs and 2,068 budding yeast paralog pairs, $P<10^{-5}$, *Z* test). Since *S. pombe* paralogs arose via SSDs, while many *S. cerevisiae* paralogs arose via a WGD event (Kellis et al., 2004), this result suggests that duplication modes could impact paralog divergence differently. To test this, we compared paralog pairs generated via the WGD event with those generated via SSDs in *S. cerevisiae*. We find that SSD pairs are more sub-functionalized than WGD pairs (Figures 8.6C and 8.7A to 8.7D; $P<0.05$, *Z* test; see Materials and Methods).

Next, we compared the extent of neo-functionalization (rewiring) (Figure 8.6A) for the two species and found that fission yeast paralog pairs tend to undergo more neo-functionalization than budding yeast pairs (Figures 8.6D and 8.7E; difference in log odds ratio=0.5 using 1,158 fission yeast paralog pairs and 1,175 budding yeast paralog pairs, $P=0.015$, *Z* test). Furthermore, within *S. cerevisiae*, SSD pairs are significantly more neo-functionalized than WGD pairs (Figures 8.6E and 8.7F to 8.7I; $P<0.05$, *Z* test).

In a WGD, the entire genome is duplicated almost at once. Soon afterward, a vast majority of the duplicates are purged while only a few are retained (Kellis et al., 2004). However, the duplicates that remain are under strong evolutionary pressure to maintain stoichiometric ratios with their interaction partners and, thus, evolve more slowly (Fares et al., 2013). On the other hand, SSDs arise sporadically and are under less pressure to maintain stoichiometric ratios (Fares et al., 2013), which explains why they can undergo greater functional divergence. This increased pressure on WGD genes to maintain stoichiometry is illustrated by their propensity to be enriched in protein complexes compared to SSD genes (Hakes et al., 2007). Using 408 high-quality literature-curated complexes from CYC2008 (Pu et al., 2009), we observed the same enrichment. Moreover, we find that the enrichment increases with the size of the complex, further supporting the notion that stoichiometric constraint influences the fate of WGD genes (Figure 8.7J).

Since WGD pairs are more functionally redundant than SSD pairs, these genes tend to be non-essential (Guan et al., 2007). It has also been shown that double deletions of these WGD pairs lead to a higher synthetic lethality rate than SSD pairs (Guan et al., 2007). Using a genome-scale genetic interaction map (Costanzo et al., 2010), we confirmed that deletion of WGD pairs is more likely to lead to synthetic lethality (Figure 8.6F; >6 fold difference in the fraction of synthetically lethal pairs, $P<10^{-10}$, Z test). Moreover, when we further stratify both groups of paralogs into pairs that are known to share interactors and pairs that have not been reported to share interactors, double deletions of the former are more likely to cause synthetic lethality than double deletions of the latter (Figure 8.6F; ~1.5 fold difference between the 2 sets for both SSD and WGD pairs, $P<0.05$ for both SSD and WGD pairs, Z test). This shows that paralog pairs that

share interactors are more likely to be functionally redundant, regardless of whether they arose via SSD or WGD.

There have been conflicting reports in the literature regarding coexpression patterns of SSD and WGD pairs (Conant and Wolfe, 2006; Guan et al., 2007). Using a compendium of genome-wide expression datasets for *S. cerevisiae* genes (Yu et al., 2008), we found no significant difference in coexpression patterns of these pairs (Figure 8.7K). However, we find that SSD and WGD paralog pairs that share interactors are significantly more likely to be coexpressed than pairs that are not known to share interactors (Figures 8.6G, 8.7L, and 8.7M; >10% more coexpressed for paralogs that are known to share interactors, $P<0.02$ in both cases, Z test; see Materials and Methods). The tendency to be coexpressed among SSD pairs and WGD pairs that share interactors is the same. Furthermore, among pairs that are not known to share interactors, WGD pairs tend to be more coexpressed than SSD pairs (Figures 8.6G, 8.7L, and 8.7M; >10% more coexpressed for WGD paralogs compared to SSD paralogs, $P<0.02$ in all cases, Z test). These results show that for both duplication modes, because paralog pairs that are known to share interactors tend to be functionally redundant, the regulation of their gene expression also tends to be retained. Only for paralog pairs that are not known to share interactors is there a significant difference in coexpression between SSD and WGD paralogs, suggesting that even these WGD pairs might still be more functionally redundant than SSD pairs. It should be noted that, due to the incompleteness of current interactomes, paralog pairs could share interactors that are currently unreported.

The availability of proteome-wide interactomes helps dissect functional redundancy and divergence, and to some degree the regulation of expression, between paralogs. Overall, our results show that a WGD leads to greater functional redundancy while SSDs lead to greater

functional diversification by sub-functionalization and neo-functionalization. Moreover, while there has been debate in the literature regarding the ubiquity of neo-functionalization (Gibson and Goldberg, 2009; He and Zhang, 2005), our results provide accurate measurements of the extent of neo-functionalization in the two yeasts.

Coevolution of conserved interactions revealed by cross-species interactome mapping

To further dissect the nature of conserved interactions, we implemented a cross-species interactome mapping approach to determine the prevalence of coevolution. We consider an interaction to be coevolved when its proteins have evolved in a coordinated manner to maintain the interaction in different species, but have developed incompatible binding interfaces with orthologs of their partners. To determine whether conserved interactions are intact or coevolved, we test by Y2H whether a protein in one species can interact with the ortholog of its interacting partner in another species. If the cross-species interaction can occur, the interaction is intact (Figure 8.5C), otherwise it is coevolved between the two species (Figure 8.8A). For example, through our cross-species mapping, we discovered that interactions of farnesyltransferase subunit Cwp1 with other subunits Cpp1 and Cwg2 have coevolved between *S. pombe* and *S. cerevisiae*; Cwp1 cannot interact with either Ram1 or Cdc43, *S. cerevisiae* orthologs of Cpp1 and Cwg2, respectively (Figure 8.8B). A previous study showed that expression of Cwp1 cannot complement a non-functional mutant of its *S. cerevisiae* ortholog, Ram2 (Arellano et al., 1998). This suggests that Cwp1, although conserved between *S. pombe* and *S. cerevisiae* at the gene level, has evolved incompatible interaction interfaces with other farnesyltransferase subunits in *S. cerevisiae* and is thus unable to reconstitute an active enzyme complex.

It is known that evolution in protein folds is essentially the result of many random mutation events (Lockless and Ranganathan, 1999). However, since only a small fraction of changes that occur via random drift will satisfy the pairwise constraints necessary for interaction conservation, coevolution at the residue level only occurs at a few specific sites and is relatively rare (Talavera et al., 2015). Surprisingly we find that coevolution at the interaction level is not uncommon: ~33% and 50% of conserved interactions between *S. pombe* and *S. cerevisiae* or human are coevolved, respectively (Figure 8.8C). This shows that even among conserved interactions, only a few key alterations at important binding sites can make the cross-species interactions incompatible and the interactions coevolved. Thus, these sites are critical to protein binding and subsequent function, and changes at these sites alter protein interactions in a manner analogous to a single amino acid change disrupting protein interactions in human disease (Wang et al., 2012; Wei et al., 2014).

Among interactions for which we were able to determine coevolution status, we found that the likelihood for an interaction to be intact between *S. pombe*-*S. cerevisiae* and *S. pombe*-human is significantly higher than random expectation, while the likelihood for an interaction to be intact for one species pair and coevolved for the other species pair is significantly lower (Figure 8.8D; difference in log odds ratio=1.7, P=0.022, Z test; see Materials and Methods). Thus, these intact interactions are likely involved in functions that have remained unchanged among yeasts and human throughout evolution.

We then investigated potential factors that could determine whether an interaction is intact or coevolved with respect to another species. We find that overall sequences of proteins involved in intact interactions tend to be better conserved across species than sequences of proteins in coevolved interactions (Figure 8.8E; 18.0% higher, P=2.1×10^{-4} for *S. pombe*-human, U test).

High sequence conservation may indicate higher levels of evolutionary constraint existing within the local network neighborhood of a given interaction. In fact, we find that proteins involved only in intact interactions have twice the number of interactors as compared to proteins involved in only coevolved interactions (Figure 8.8F; P=1.1×10^{-3}, U test), suggesting that the added evolutionary constraint of maintaining many interacting partners may prevent the coevolution of two interacting proteins. Finally, we find that the most highly evolutionarily correlated inter-protein residue pairs in coevolved interactions are significantly more correlated than top residue pairs in intact interactions, suggesting that the maintenance of coevolved interactions involves compensatory changes at the amino acid residue level.

Implications of FissionNet for the study of human disease

We explored the relevance of FissionNet to human disease by considering the context of known human disease mutations from HGMD (Stenson et al., 2014) within proteins of the human interactome conserved in *S. pombe*. We find that among human interactions conserved in either *S. pombe*, *S. cerevisiae*, or both, ~40% of inter-protein pairs of disease mutations cause the same disease (Figure 8.9A). This is significantly higher than in human interactions that are not reported to be conserved in either yeast or cannot be conserved in either due to lack of protein orthologs (Figure 8.9A; P<10^{-10} for all pairwise comparisons, Z test). Based on these results, mutations that break specific protein-protein interactions to cause diseases may be overrepresented among interactions conserved in model organisms. From a global network view, FissionNet may be highly relevant to the study of human disease based on the large portion of *S. pombe* interactions in which both proteins have human orthologs with known germline disease or

somatic cancer-associated mutations (Figure 8.9B; 902 interactions) (Forbes et al., 2015; Stenson et al., 2014).

To demonstrate the plausibility of studying specific human disease mutations using FissionNet, we explored whether human disease mutations that disrupt human interactions intact in *S. pombe* also disrupt the corresponding interactions of the fission yeast orthologs. We focused on three examples: two Mendelian disease variants (Stenson et al., 2014) that disrupt the human NMNAT1-NMNAT1 and PCBD1-PCBD1 interactions and one population variant from the Exome Sequencing project (Fu et al., 2013) that disrupts the human SNW1-PPIL1 interaction. We find that introducing these human protein residue changes into their *S. pombe* orthologs also disrupts the fission yeast interactions (Figure 8.9C). These results indicate that cross-species interactome mapping enables investigation of whether interaction interfaces are altered at the molecular level between model organisms and human, a finding with potentially far-reaching implications for the study of protein function and human disease.

Our results regarding gene duplication modes may also be relevant to the study of human disease. We find that human WGD paralog pairs have a significantly higher likelihood to be involved in the same disease compared to human SSD paralog pairs, in agreement with our observation that WGD paralog pairs tend to be functionally redundant (Figure 8.9D; 7 fold difference in the fraction of WGD and SSD pairs that cause the same disease, $P<10^{-10}$, Z test; see Materials and Methods). Thus, our findings have direct implications for understanding the functional roles of paralogous genes, from yeasts to human.

8.4 DISCUSSION

FissionNet provides a wealth of functional information. For example, we find that the Atf1-Cid12 interaction mediates silencing at Atf1-target genes *hsp16* and *hsp104*. It has been shown that the RNAi pathway is involved in silencing of these genes (Woolcock et al., 2012). Hence, it is possible that the Atf1-Cid12 interaction is part of an RNAi-dependent regulatory pathway.

By comparing FissionNet to protein networks in budding yeast and human, we have shown that the molecular bases for interaction conservation among orthologous proteins are complex and different from those that underlie gene conservation. This is highly relevant to the use of the two yeasts as model organisms as there are functions that can be better studied using fission yeast. We find that divergence across species is not completely dictated by sequence level changes, suggesting that rewiring of interactomes plays an important role in species evolution. Additionally, our finding that proteins in a significant fraction of conserved interactions have undergone coevolution to maintain interactions has major implications for studies reliant on the expression of human proteins in model organisms to identify functional mechanisms (Tardiff et al., 2013).

Gene duplications introduce evolutionary innovation and robustness

Gene duplication is a key process shaping evolution (Figure 8.9E). Our results show that paralogs arising via WGD are under strong constraints to maintain stoichiometric ratios with their interaction partners and, hence, tend to maintain functional redundancy; on the other hand, duplicates arising via SSDs are not under such strong constraints and are more likely to gain novel functions (Figure 8.9F). For example, it has previously been reported that duplicate copies

of the *SRGAP2* gene that arose via segmental duplications (SSD-like events) have gained new functions related to brain development specifically in the human lineage (Dennis et al., 2012).

Gene duplications play an important role in the evolutionary mechanism governing speciation as well as the evolution of developmental and morphological complexity in vertebrates (Rensing, 2014; Ting et al., 2004). For example, two rounds of WGD have been predicted in the origin of jawed vertebrates (Figure 8.9E) (Kasahara, 2007). During speciation, while certain key functions need to be evolutionarily preserved, new functions are necessary for differential adaptation between species (Ting et al., 2004). Previous studies have identified how duplication events can lead to functional changes through gene dosage alterations (Papp et al., 2003). Our results help establish on a proteomic scale that paralogs arising via WGD are more likely to preserve functions and provide robustness for important cellular functions, while paralogs arising via SSDs are more likely to contribute to novel functions gained by specific species. These findings further our understanding of human biology and disease.

Future directions

Our analyses focus on budding yeast, fission yeast, and human, as they are the only three eukaryotic organisms for which we have proteome-scale interactome networks using our version of the Y2H assay (>50% of all protein pairs screened). Once more interactome networks are systematically generated in other species, using assays with measured sensitivity and specificity, the comparative network analysis framework established in this study can be readily applied to further elucidate the extent and nature of the evolution of protein functions across many species.

8.5 MATERIALS AND METHODS

Generation of the binary protein-protein interactome map of *S. pombe*

FissionNet was generated by triplicate independent screening of ~4,900 *S. pombe* ORFs. The network was validated by testing a representative 220 interacting ORF pairs using PCA assays and by determining its functional properties with respect to random pairs and to a literature-curated network.

Conservation of interactions in *S. pombe*, *S. cerevisiae*, and human

We focused only on interactions that can be conserved, *i.e.*, both proteins involved in the interaction have orthologs in the other species We mapped interactions in the reference species to their corresponding ortholog pairs in the other species and tested these pairs using our Y2H assay in a pairwise fashion. Overall, results from these pairwise retests for all three species (a total of ~20,000 individual Y2H experiments) are used to obtain the observed conservation fraction. To accurately estimate the true conservation fraction, we developed a rigorous Bayesian framework that takes into account both the false positive and false negative rates of our Y2H assay, and computes the true conservation fraction from the observed fraction.

Positive and negative reference sets

The PRS and NRS constitute sets of positive and negative controls, respectively. Our PRS comprises 93 *S. pombe* interactions that have been previously reported in 2 or more publications. To construct the NRS we choose 168 random protein pairs that have not been reported to interact in *S. pombe* and whose orthologs have not been reported to interact in any species. In a set of

random protein pairs, the expected fraction of interactions is ~10^{-3}-10^{-4} (Riley et al., 2005; Yu et al., 2008), the expected number of interactions in a random set of 168 pairs is <10^{-3}×168 (≈0.2). Since we exclude pairs that are known to interact, the expected number of interactions in our NRS is even lower.

Identification of Cid12 mutants

In order to select residues integral to the Cid12-Atf1 interaction (*i.e.*, at the interface), but not to Cid12-Hrr1 or Cid12-Rdp1, we used Direct Coupling Analysis (Morcos et al., 2011) to determine evolutionarily correlated residues across interfaces of these interactions in 28 yeast species. Cid12 residues exhibiting the strongest evolutionary couplings with Atf1 residues were considered likely to facilitate the Cid12-Atf1 interaction. In order to increase the chances of selecting Cid12 residues that are not at the interaction interface of other Cid12 interactions, we did not consider any Cid12 residues with strong evolutionary couplings with Hrr1 or Rdp1. Once Cid12 residues were chosen, we introduced amino acid mutations designed to strongly alter the hydrophobicity of the wild-type amino acid.

Detection rates of the positive (PRS) and negative (NRS) reference sets

Of the 168 NRS pairs, 78 are between proteins with different sub-cellular localizations and 90 have the same sub-cellular localization (Matsuyama et al., 2006). However, there is no significant difference in the fraction of random pairs detected by either Y2H (0/78 and 0/90 for the two sets respectively, P=0.92 using a Z test) or PCA (8/78 and 9/90 for the 2 sets respectively, P=0.96 using a Z test). To examine if there are any species-specific biases of our Y2H assay, we computed the fractions of PRS and PRS-nonY2H (subset of PRS interactions that

have been detected using an assay other than Y2H) interactions in the three different species that are recapitulated by our Y2H assay. We find that there is no significant difference between the detection rates across species (P>0.35 for all pairs, Z test). Furthermore, we find that there is no significant difference in interaction density (*i.e.*, number of interactions detected divided by total number of protein pairs screened) for FissionNet and previously reported Y2H interactomes in *S. cerevisiae* (Yu et al., 2008) and human (Rolland et al., 2014) (all interaction densities differ by <2 fold). These results confirm that our Y2H assay has no species-specific detection biases.

Calculating the coexpression of genes

To measure the coexpression of transcripts corresponding to proteins involved in FissionNet interactions, we calculated the Pearson Correlation Coefficient (*PCC*) between their expression profiles: expression values measured at different time-points in the cell cycle (Rustici et al., 2004). We also calculated the *PCC* between expression profiles of transcripts corresponding to proteins involved in high-quality *S. pombe* interactions from literature curation. Finally, we defined two different sets of random pairs: (1) all random pairs, (2) random pairs by permuting edges between proteins in the network. We first compared the different distributions using a KS test. Next, we calculated the fractions of significantly co-expressed interactions, as well as the fraction of significantly co-expressed random pairs. We defined significant coexpression as *PCC* ≥ a threshold value. When comparing the fractions of interactions or pairs that are significantly co-expressed, P-values were calculated using a Z test.

Other functional properties of FissionNet

For other calculations, since small-scale studies could focus on proteins with more complete annotations in GO, we restricted our analyses to a set of proteins found in both high-quality literature-curated *S. pombe* interactions (Das and Yu, 2012) and interactions in FissionNet. We then defined 3 sets of protein pairs such that both proteins are from the previously defined set: (1) high-quality *S. pombe* interactions from literature curation, (2) *S. pombe* interactions from FissionNet, and (3) all pairs of proteins for which the two proteins have never been reported to interact. We performed the following calculations on these 3 sets:

Calculating functional similarity

We calculated functional similarity using a total ancestry method that computes all pairwise functional similarities in a set of proteins by determining for each given pair of proteins, the number of other protein pairs sharing the same set of parent GO terms (Yu et al., 2007). In this framework, a pair of proteins that are very dissimilar will share their GO ancestry with a large number of other protein pairs. Conversely, a pair of proteins that are very similar will share their GO ancestry with only a few or none of the other pairwise combinations of proteins in the same set. Each similarity score for a pair of proteins was computed as a percentile ranking of their total ancestry score among all such scores calculated for all pairwise combinations of proteins in the set. We considered the top 1% of protein pairs in this ranking to be functionally similar. *P*-values were calculated using a *Z* test.

Calculating co-localization

To calculate the co-localization of proteins involved in FissionNet interactions, we calculated the fraction of protein pairs that have the same sub-cellular localization (Matsuyama et al., 2006). *P*-values were calculated using a *Z* test.

Conservation of genes

To analyze the extent to which genes are conserved, we calculated the fraction of genes in the reference species *i* that also have orthologs in the other species *j*:

$$Gene\_cons_{ij} = \frac{G_i^{j'}}{G_i}$$

where G_i denotes the total number of genes in species *i* and $G_i^{j'}$ the number of genes in species *i* that have corresponding orthologs in species *j*. Using ortholog annotations from PomBase and the Saccharomyces Genome Database, we computed the extent of gene conservation between different species pairs for all coding genes (Cherry et al., 2012; McDowall et al., 2015). We also used orthologs from InParanoid to compute the extent of gene conservation between different species pairs for all coding genes (Sonnhammer and Ostlund, 2015). We observe the same gene conservation trends regardless of which database is used for determining orthology, confirming the robustness of our result.

Estimating true interaction conservation fractions

To calculate the extent to which interactions are conserved, we focused only on those interactions that can be conserved, *i.e.*, both proteins involved in the interaction have orthologs in the other species. For each pair of organisms, we used both organisms as the reference (six comparisons for 3 species). We mapped interactions in the reference species to their corresponding ortholog pairs in the other species and tested these pairs using our Y2H assay in a

pairwise fashion. We performed pairwise retests because we have shown earlier that not all interactions detected by Y2H in a pairwise fashion will be detected in a high-throughput screen where individual baits are tested against minipools of ~188 preys (Yu et al., 2008). Overall, results from these pairwise retests for all three species (a total of ~20,000 individual Y2H experiments) are used to obtain the observed conservation fraction. To accurately estimate the true conservation fraction, we used a rigorous Bayesian framework that takes into account both the false positive and false negative rates of our Y2H assay, and computes the true conservation fraction from the observed fraction.

Using the law of total probability, we can write:

$$P(D \mid I') = P(D \mid I,\ I') \times P(I \mid I') + P(D \mid \bar{I},\ I') \times P(\bar{I} \mid I') \quad (1)$$

Here, I' denotes the event that an interaction occurs in the reference species, I the event that the interaction occurs in another species, $\bar{I}$ the event that the interaction does not occur in the other species and D the event that it is detected in the other species using our Y2H pipeline. The observed conservation rate is $P(D|I')$. The true conservation rate is $P(I|I')$. As an interaction in the reference species can only be either conserved or rewired in the other species:

$$P(I \mid I') + P(\bar{I} \mid I') = 1 \quad (2)$$

Finally, we can assume conditional independence between D and I' given I. In other words, given that an interaction occurs in the other species, whether it is Y2H detectable in that species and whether its ortholog pair interacts in the reference species are independent of each other. Using this:

$$P(D \mid I, I') = \frac{P(D,I,I')}{P(I,I')} = \frac{P(D, I' \mid I) \times P(I)}{P(I,I')} = P(D \mid I) \times \left(\frac{P(I' \mid I) \times P(I)}{P(I,I')} \right) = P(D \mid I) \quad (3)$$

Using similar arguments,

$$P(D \mid \bar{I}, I') = P(D \mid \bar{I}) \quad (4)$$

Substituting equations (2), (3) and (4) in equation (1), we obtain:

$$P(I \mid I') = \frac{P(D \mid I') - P(D \mid \bar{I})}{P(D \mid I) - P(D \mid \bar{I})} \quad (5)$$

$P(D| I')$ is estimated using the fraction of interactions in the reference species that are detected by Y2H to interact in the other species (f_d). $P(D|I)$ is estimated using the fraction of a set of true interactions (PRS) that we can detect using our Y2H assay (f_{prs}). Finally, $P(D|\bar{I})$ is estimated using the fraction of a set of random pairs that are unlikely to interact (NRS) that we can detect using our Y2H assay (f_{nrs}). So, for any species pairs:

$$P(I \mid I') = \frac{f_d - f_{nrs}}{f_{prs} - f_{nrs}} (6)$$

We can estimate the error using the delta method:

$$SE_{P(I|I')} = \sqrt{\frac{\left(f_{prs} - f_{nrs}\right)^2 \times \left(SE_{f_d}\right)^2 + \left(f_{prs} - f_d\right)^2 \times \left(SE_{f_{nrs}}\right)^2 + \left(f_d - f_{nrs}\right)^2 \times \left(SE_{f_{prs}}\right)^2}{\left(f_{prs} - f_{nrs}\right)^4}} \quad (7)$$

Interaction conservation using assays other than Y2H

We examined the observed conservation as detected by individual assays rather than using overall interactome networks from the literature as these are derived from assays with varied and unknown false positive and false negative rates. However, for a single assay with unknown false positive and false negative rates, while we will be unable to calculate the true underlying conservation fraction, we can still compute the observed conservation fraction. We first calculated the fraction of FissionNet interactions whose corresponding *S. cerevisiae* and human ortholog pairs have been shown to interact in co-crystal structures (Das and Yu, 2012). We find that that fission yeast interactions are better conserved in human than in budding yeast (>2 fold difference in observed conservation, $P<10^{-3}$). Next, we calculated the fraction of FissionNet

interactions whose corresponding *S. cerevisiae* and human ortholog pairs have been detected as interacting by proteome-scale affinity purification/mass spectrometry experiments (Gavin et al., 2006; Huttlin et al., 2015; Krogan et al., 2006). Here, we also find that fission yeast interactions are better conserved in human than in budding yeast (>1.5 fold difference in observed conservation, $P<10^{-3}$ in both cases).

Identifying proteins conserved in eukaryotes

To identify proteins that are conserved across eukaryotes, we used clusters of conserved eukaryotic orthologous groups of genes (KOGs) as defined by Koonin *et al.* (Koonin et al., 2004). These conserved KOGs often comprise genes essential for survival and could be considered to approximate "a minimal set of essential eukaryotic genes" (Koonin et al., 2004). Each KOG consists of orthologous genes in up to 7 representative eukaryotic species studied by the authors. We defined proteins conserved in eukaryotes as those proteins from these KOGs that are conserved in ≥5 species.

Interaction conservation in different biological processes

We used the Gene Ontology (GO) (Ashburner et al., 2000) to categorize interactions based on the annotations of the proteins involved. We computed interaction conservation in GO Slim Biological Process (BP) categories, a set of 70 terms representative of diverse biological processes not specific to any one organism. For all analyses, we considered only genes annotated with experimental evidence codes (Ashburner et al., 2000). We considered an interaction to be within a category if either of its interacting proteins is annotated in that category or one of its children.

Sequence conservation of proteins and interactions

To determine the sequence conservation between two proteins, alignments were produced using the pairwise2.align.global function of the BioPython Python module, an implementation of the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch, 1970). We used the BLOSUM62 scoring matrix, a gap-open penalty of -10 and a gap-extend penalty of -0.5. Two amino acids are considered similar if the BLOSUM62 score associated with a substitution between the two residues is >0. Unless otherwise specified, sequence similarity is measured with sequences of *S. pombe* proteins serving as the reference. Sequence similarity between an *S. pombe* protein and an ortholog in another species is measured as the fraction of *S. pombe* residues similar to their aligned residues in either *S. cerevisiae* or human. To calculate the sequence similarity of pairs of proteins with orthologous pairs, the individual sequence similarities of each protein with their orthologs are averaged. *P*-values were calculated using a *Z* test.

Interface domain conservation based on co-crystal structures

We compiled a set of co-crystal structures from the PDB representing human protein-protein interactions. For each structure, we calculated interface residues using NACCESS to determine surface residues whose solvent accessible surface area was altered by ≥1Å^2 between bound and unbound states (Hubbard, 1996). To determine interface residues of protein interactions, we took the union of interface residues determined from each representative PDB chain pair for which at least 5 interface residues were calculated in each chain. In the human interactions, we identified Pfam domains at the interaction interface as those domains containing at least 5 interface

residues. All domains not meeting this criterion are considered 'Other' as we don't know if they facilitate the interaction or not. We then aligned the full human protein sequences in each interaction to their orthologs in *S. pombe* and *S. cerevisiae* using the alignment method mentioned previously. Here, we used the human sequences as the reference and only calculated sequence similarity within the portions of the alignment in the human domain regions. *P*-values were calculated using a *U* test.

ClusterOne

We performed clustering with ClusterONE (Nepusz et al., 2012). ClusterONE finds overlapping functional modules and is specifically tuned for clustering biological networks. We used ClusterONE with parameters s=3 (minimum cluster size) and d=0.5 (minimum cluster density) and found 193 clusters in our network. Since proteins can belong to multiple clusters, we defined an intra-cluster interaction as any interaction for which there is a cluster that contains both proteins and an inter-cluster interaction as any interaction for which both proteins belong to clusters, but there is no cluster that contains both proteins. Intra-cluster and inter-cluster conservations were calculated using the fraction of interactions within and across clusters that are detected as conserved using our Y2H assay, transformed via the Bayesian framework described above to obtain the true conservation fractions. *P*-values were calculated using a *Z* test.

Distribution of intact and coevolved interactions across species

We computed the log-odds ratios for 3 scenarios: an interaction is intact in both species pairs (*S. pombe*-*S. cerevisiae* and *S. pombe*-human), an interaction is coevolved in both species pairs, an interaction is intact in one species pair but coevolved in the other:

$$LOR = \log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right)$$

where, p_1 is the observed fraction of interactions in each category and p_2 the expected fraction of interactions in each category. The expected fraction is calculated assuming independence between the events of being intact/coevolved in each species pair. Standard error was calculated using the delta method:

$$SE_{LOR} = \sqrt{\left(\frac{SE_{p1}{}^2}{p_1{}^2 \times (1-p_1)^2} + \frac{SE_{p2}{}^2}{p_2{}^2 \times (1-p_2)^2}\right)}$$

P-values were calculated using a *Z* test.

Sub-functionalization and neo-functionalization

We obtained a set of 3,853 fission yeast paralog pairs and 6,846 budding yeast paralog pairs from Ensembl Biomart (Kinsella et al., 2011). For budding yeast, WGD paralogs were defined based on annotations from Kellis *et al.* (Kellis et al., 2004), and the rest were considered to be SSD paralogs.

To measure the extent of sub-functionalization, we calculated the fraction of interactions that are conserved but not shared among paralog pairs. We normalized this by the fraction of conserved but not shared interactions among all pairs of proteins that do not have a paralog. The fraction of conserved and not shared interactions is equal to 1 – the fraction of conserved and shared interactions. To calculate the fraction of conserved and shared interactions, we first constructed a set of high-quality interactions from the literature that are conserved. If we use the literature to ascertain how many of these interactions are shared, the fraction will be inaccurate as literature-curated interactomes are incomplete and suffer from detection rate biases. To

circumvent this, we first calculated the fraction of conserved interactions that were detected as shared using our Y2H assay. We then used our previously developed framework to calculate the actual number of shared interactions:

$$f_{shared} = \frac{f_{obs} \times precision}{completeness \times \text{assay\_sensitivity} \times sampling\_sensitivity}$$

where f_{obs} is the detected fraction of shared pairs using our Y2H assay and f_{shared} is the actual fraction of shared pairs (Yu et al., 2008). Precision, completeness, assay-sensitivity, and sampling-sensitivity for FissionNet are calculated as previously described (Yu et al., 2008). For the CCSB-YI1 network, they have been previously reported (Yu et al., 2008). With the calculated fractions of conserved and not shared interactions, we computed the following log odds ratio for *S. pombe* and *S. cerevisiae*:

$$LOR = \log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right)$$

where p_1 is the fraction of interactions that are conserved with its ortholog but not shared among paralog pairs and p_2 by the fraction of conserved but not shared interactions among all pairs of proteins that do not have a paralog. Standard error was calculated using the delta method as described earlier. *P*-values were calculated using a *Z* test.

To measure the extent of neo-functionalization, we calculated the log odds ratio of the fractions of rewired interactions involving proteins that have and do not have paralogs. We computed the same log odds ratio for *S. pombe* and *S. cerevisiae*:

$$LOR = \log\left(\frac{\frac{p_3}{1-p_3}}{\frac{p_4}{1-p_4}}\right)$$

where, p_3 is the fraction of rewired interactions involving proteins where at least one has a paralog and p_4 the fraction of rewired interactions between proteins that do not have paralogs. The fraction of rewired interactions is defined as 1 – the fraction of conserved interactions. Since we are able to calculate the true fraction of conserved interactions using a Bayesian framework that accounts for assay false positive and negative rates (please refer to '**Conservation of interactions in *S. pombe*, *S. cerevisiae*, and human**'), the fraction of rewired interactions used for this calculation is also accurate and has taken into account for assay detection rates. Standard error was calculated using the delta method as described earlier. *P*-values were calculated using a *Z* test.

Our definition of rewiring is based on the interactions in the orthologous species. However, in cases where a paralog pair in the reference species shares an interaction that is rewired in the orthologous species, it is possible that the common ancestor may have this interaction. It could be argued that if the common ancestor does have the interaction, it is not truly neo-functionalized. To account for this (and since the interactome for the common ancestor is unknown), we constructed a set of interactions involving at least one protein that has a paralog, and the interactor of the paralog has only degree one (only one interaction), *i.e.*, by definition that interactor cannot be shared between paralogs in the reference species. Even for this set, we find that *S. pombe* paralog pairs are significantly more neo-functionalized than *S. cerevisiae* paralog pairs, confirming the robustness of our results.

Correcting for divergence times, sequence evolution rates and sequence identities

We obtained JTT-corrected divergence times for paralog pairs from Fares *et al*. (Fares et al., 2013). To ensure that the observed differences between SSD and WGD paralog pairs are not

due to differences in divergence times, we selected only those SSD and WGD pairs whose divergence times are between the 10th and the 90th percentile of the WGD divergence time distribution. The rationale here is to use the WGD distribution as a reference (we remove the top and the bottom 10 percentiles to eliminate outliers) and sample SSD paralog pairs that are only from this divergence time window.

We used K_a to calculate sequence evolution rates. K_a (not K_s or K_a / K_s) is an appropriate choice to correct for sequence evolution rate because synonymous substitutions between WGD pairs are essentially saturated (Byrne and Wolfe, 2007). As mentioned earlier, to ensure that the observed differences between SSD and WGD paralog pairs are not due to differences in sequence evolution rates, we selected only those SSD and WGD pairs whose sequence evolution rate are between the 10th and the 90th percentile of the WGD K_a distribution.

We obtained paralog sequence identities from Ensembl BioMart. Here too, as earlier, to ensure that the observed differences between SSD and WGD paralog pairs are not due to differences in sequence identity, we selected only those SSD and WGD pairs whose identities are between the 10th and the 90th percentile of the WGD sequence identity distribution. Since sequence identity depends both on divergence time and sequence evolution rates, correcting for sequence identity simultaneously corrects for both covariates.

Functional properties of *S. cerevisiae* SSD and WGD pairs

To calculate the fraction of SSD and WGD pairs in complexes, we used high-quality literature curated complexes from CYC2008 (Pu et al., 2009). We computed the fractions of proteins from SGD and WGD pairs that are in all CYC2008 complexes, complexes with ≥10 proteins, and complexes with ≥20 proteins. *P*-values were calculated using a *Z* test.

To calculate the fraction of SSD and WGD pairs that involve non-essential genes but lead to synthetic lethality when both genes are deleted, we used genome-scale double knockout phenotype data (Costanzo et al., 2010). We considered a double deletion to lead to synthetic lethality if the genetic interaction score (ε) is strongly negative, *i.e.*, passes a stringent cutoff as defined by the authors at http://drygin.ccbr.utoronto.ca/~costanzo2009/ where ε<-0.12 and P<0.05. We considered a paralog pair to "share interactors" if both proteins had at least 2 interactors and they shared >50% of their interactors. "Other" paralog pairs are defined as those pairs that are not known to have any shared interactors based on the literature. *P*-values were calculated using a *Z* test.

To calculate the fraction of SSD and WGD pairs that are coexpressed, we used a normalized expression dataset constructed as described in Yu *et al*. (Yu et al., 2008). Paralog pairs that "share interactors" and "other" paralog pairs are defined as described above. We defined significant coexpression as *PCC* ≥a threshold value. To ensure that our conclusions are robust to the choice of this threshold, we used three different thresholds: 0.3, 0.4 and 0.5. When comparing the fractions of significantly co-expressed pairs, *P*-values were calculated using a *Z* test.

Calculation involving human SSD and WGD pairs

A set of human WGD (ohnolog) pairs was obtained from Makino and McLysaght (Makino and McLysaght, 2010). A set of human SSD pairs was identified as described in Singh *et al.* (Singh et al., 2014). This study also used the previous set of human WGD pairs (Makino and McLysaght, 2010) for their analyses. We calculated the fractions of SSD and WGD pairs containing genes that are known to cause the same disease based on HGMD (Stenson et al.,

2014). Two genes are said to cause the same disease if at least one HGMD mutation on each of the two genes is associated with the same disease.

8.6 FIGURE LEGENDS

Figure 8.1.

A Proteome-wide Binary Protein Interactome Map of *S. pombe*

(A) Network representation of FissionNet. Proteins are color-grouped based on PomBase GO slim categories. The number of FissionNet interactions per group is indicated. (B) Y2H and PCA detection rates of the PRS, NRS, FissionNet, and FissionNet hub interactions. (C) Pearson correlation coefficient (*PCC*) distribution of gene expression profiles of interacting and all random protein pairs. (D) Enrichment of co-localized protein pairs. (E) Enrichment of protein pairs sharing similar functions. (F) Subnetwork of Tas3 and Hhp1 in FissionNet. (G) Coimmunoprecipitation of Tas3-myc and Hhp1-HA *in vivo*. (H) Centromeric silencing assay of *tas3Δ* and *hhp1Δ* cells. A schematic of the *imr1R* region with the *ura4*$^{+}$ reporter gene is shown. WT denotes wild-type. Data are shown as measurements + standard error (SE). * denotes significant (P<0.05); n.s. denotes not significant.

Figure 8.2.

Atf1-Cid12 Interaction Mediates Silencing at Heat-shock Genes

(A) Subnetwork of Atf1 and Cid12 in FissionNet. (B) Coimmunoprecipitation of Atf1-myc and Cid12-HA *in vivo*. (C) Semi-quantitative real-time PCR (semi qRT-PCR) shows *hsp16* and *hsp104* transcript levels in deletion strains. (D) Y2H confirms Cid12 mutants cannot interact

with Atf1, but maintain interactions with Hrr1 and Rdp1. (E) Semi qRT-PCR shows that the Cid12 mutants in *cid12Δ* cells do not restore the repression of *hsp16* or *hsp104*. (F) Centromeric silencing assay shows that Cid12 mutants retain centromeric silencing function. -RT, no reverse transcriptase. +RT, with reverse transcriptase. $Act1^{+}$ serves as loading control. WT denotes wild-type.

Figure 8.3.

S. pombe Protein Interactions are More Conserved in Human than in *S. cerevisiae*

(A) Sequence-based phylogeny dendrogram of *S. pombe* (*S.p.*), *S. cerevisiae* (*S.c.*), and human (*H.s.*). (B) Interaction conservation between reference-query species. (C) Sequence conservation for ortholog pairs that could be conserved between *S.p.*-*S.c.* and *S.p.*-*H.s.* (D) Interaction conservation between reference-query species for proteins that are conserved in all three species. (E) Interaction conservation in GO Slim categories with at least 50 interactions. (F) Interaction conservation among GO Slim categories that are conserved in all three species. (G) FissionNet subnetworks of Srrm1, SPAC30D11.14C, and SPAC1952.06C. (H) Global splicing profiles of deletion strains relative to wild-type. Columns represent total mRNA (T), pre-mRNA (P), and mature mRNA (M). Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant.

Figure 8.4.

S. pombe Protein Interactions Are More Conserved in Human than S. cerevisiae

(A and B) Pairwise comparisons of the conservation of all coding genes between reference-query species pairs. (A) Orthologs are defined by PomBase (McDowall et al., 2015) and

Saccharomyces Genome Database (SGD) (Cherry et al., 2012). (B) Orthologs are defined by InParanoid (Sonnhammer and Ostlund, 2015). (C) Y2H detection rates of PRS and PRS_nonY2H (subset of PRS interactions that have been detected using an assay other than Y2H) interactions in fission yeast, budding yeast, and human. (D) Interaction density, i.e., interactions detected out of the total number of proteins pairs screened (log scale) in different organisms. (E) Observed interaction conservation between reference-query species pairs. (F) Observed interaction conservation between reference-query species pairs for proteins that have 1:1 orthologs between reference and query species. (G) Observed interaction conservation between reference-query species pairs using co-crystal structures for S. cerevisiae and human.
(H and I) Observed interaction conservation between reference-query species pairs using large-scale AP/MS datasets for S. cerevisiae and human. For both panels, the human AP/MS dataset used is from (Huttlin et al., 2015). (H) The S. cerevisiae AP/MS dataset is from Gavin et al. (2006). (I) The S. cerevisiae AP/MS dataset is from Krogan et al. (2006). (J) Observed interaction conservation between reference-query species pairs for proteins that have 1:1:1 orthologs between fission yeast, budding yeast, and human. (K) Observed interaction conservation between reference-query species pairs for proteins that are conserved in all eukaryotes. (L–N) Observed conservation fractions of S. pombe interactions in S. cerevisiae and human in different GO Slim biological process categories with at least (L) 30, (M) 50, and (N) 75 interactions. (O) Observed interaction conservation among GO Slim categories that are conserved in all three species. (P) Overlap of genes whose intron splicing is affected by deletion of either Srrm1 or its interaction partner Srp1. Indicated within the diagrams are the number of genes affected. (Q) Distribution of log odds scores of affected (intron accumulation of log2 0.5 or greater in srrm1D versus wild type cells) versus unaffected (intron accumulation of less than

log2 0.5 in srrm1D versus wild-type cells). The log odds score for each annotated 50 splice site measures the sequence similarity of that site relative to the consensus 50 splice site of each intron. Data are shown as measurements + SE. * denotes significant ($p < 0.05$); n.s. denotes not significant.

Figure 8.5.

Determinants of Interaction Conservation

(A) Interaction conservation as a function of overall protein sequence similarity. (B) Sequence similarity within protein interaction domains and other domains for interactions conserved between yeasts and human. (C) Y2H confirms the interactions of human (*H.s.*) DRAP1-DR1, the orthologous *S. pombe* (*S.p.*) Dpb3-Ncb2, and the cross-species interactions. (D) Crystal structure of human DR1-DRAP1. Boxed region highlights interaction domains. Gray shaded regions denote aligned interaction domain sequences. (E) Interaction conservation within and across topological clusters. (F) Interaction conservation within and across GO categories. Data are shown as measurements + SE. * denotes significant ($P<0.05$); n.s. denotes not significant. Abbreviations are *S. pombe* (*S.p.*), *S. cerevisiae* (*S.c.*), and human (*H.s.*).

Figure 8.6.

Functional Divergence of Interactions Involving Paralogous Proteins

(A) Schematic representation of sub-functionalization and neo-functionalization. (B-C) Log odds ratios of sub-functionalization (B) for *S. pombe* and *S. cerevisiae* paralog pairs and (C) for *S. cerevisiae* SSD and WGD paralog pairs after correcting for divergence times. (D-E) Log odds ratios of neo-functionalization (D) for *S. pombe* and *S. cerevisiae* paralog pairs and (E) for *S.*

cerevisiae SSD and WGD paralog pairs after correcting for divergence times. (F) Fraction of synthetic lethal pairs among SSD and WGD paralogs known or not known to share interactors. (G) Fraction of coexpressed pairs (*PCC*>0.4) among SSD and WGD paralogs known or not known to share interactors. Data are shown as measurements + SE. * denotes significant (*P*<0.05); n.s. denotes not significant.

Figure 8.7.

Functional Divergence of Interactions Involving Paralogous Proteins

(A–D) Log odds ratio of sub-functionalization for S. cerevisiae SSD and WGD paralog pairs: (A) correcting for sequence evolution rates, (B) correcting for sequence identities, (C) without correcting for any covariates, and (D) where SSD and WGD pairs are defined using an independent dataset (Fares et al., 2013). (E) Log odds ratio of neo-functionalization for S. pombe and S. cerevisiae paralog pairs that do not share interactions. (F–I) Log odds ratio of neo-functionalization for S. cerevisiae SSD and WGD paralog pairs: (F) correcting for sequence evolution rates, (G) correcting for sequence identities, (H) without correcting for any covariates, and (I) where SSD and WGD pairs are defined using an independent dataset (Fares et al., 2013). (J) Enrichment of proteins from WGD paralog pairs compared to proteins from SSD paralog pairs in protein complexes of different sizes. (K) Fraction of SSD and WGD paralog pairs whose proteins are coexpressed (PCC > 0.4), without separating pairs that are known to share and not known to share interactions. (L and M) Fraction of coexpressed pairs at other PCC cutoffs of (L) > 0.3 or (M) > 0.5 among SSD and WGD paralogs that are known to share and not known to share interactions. Data are shown as measurements + SE. * denotes significant ($p < 0.05$); n.s. denotes not significant.

Figure 8.8.

Intact and Coevolved Interactions

(A) Schematic representation of conserved protein interactions that are either intact or coevolved. (B) Within- and cross-species Y2H detects coevolved interactions. (C) Fraction of *S.p.* interactions that are coevolved with respect to *S.c.* or human (*H.s.*). (D) Log odds ratio of co-occurrence of intact and coevolved interactions between *S.p.*-*S.c.* and *S.p.*-*H.s.* (E) Overall protein sequence similarity of *S.p.* proteins involved in intact or coevolved interactions. (F) Number of interactors for proteins involved in intact or coevolved interactions. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant.

Figure 8.9.

FissionNet as a Resource for Studying Human Disease

(A) Fraction of inter-protein HGMD mutation pairs that cause the same disease in human interactions with regard to their conservation status in *S. pombe* and *S. cerevisiae*. (B) Largest connected subcomponent of FissionNet wherein all proteins have human orthologs with known germline disease or somatic cancer-associated mutations. (C) Impact of human disease mutations and a population variant on intact interactions between human and fission yeast. (D) Fraction of human SSD and WGD paralogs that cause the same disease. (E) The 2R hypothesis predicts two recent WGD events leading to the vertebrate lineage. (F) WGD can lead to more functional redundancy through targeted gene loss that maintains stoichiometric ratios of protein products. SSD leads to more neo-functionalization and sub-functionalization through alterations to initially redundant paralogs. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant.

Table 8.1 Positive and negative reference sets

8.7 REFERENCES

Arabidopsis Interactome Mapping Consortium (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Arellano, M., Coll, P.M., Yang, W., Duran, A., Tamanoi, F., and Perez, P. (1998). Characterization of the geranylgeranyl transferase type I from Schizosaccharomyces pombe. Mol Microbiol *29*, 1357-1367.

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

Bader, J.S., Chaudhuri, A., Rothberg, J.M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol *22*, 78-85.

Blencowe, B.J., Issner, R., Nickerson, J.A., and Sharp, P.A. (1998). A coactivator of pre-mRNA splicing. Genes Dev *12*, 996-1009.

Conant, G.C., and Wolfe, K.H. (2006). Functional partitioning of yeast co-expression networks after genome duplication. PLoS Biol *4*, e109.

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

Das, J., Vo, T.V., Wei, X., Mellor, J.C., Tong, V., Degatano, A.G., Wang, X., Wang, L., Cordero, N.A., Kruer-Zerhusen, N.*, et al.* (2013). Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. Sci Signal *6*, ra38.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

Dennis, M.Y., Nuttle, X., Sudmant, P.H., Antonacci, F., Graves, T.A., Nefedov, M., Rosenfeld, J.A., Sajjadian, S., Malig, M., Kotkiewicz, H.*, et al.* (2012). Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. Cell *149*, 912-922.

Fares, M.A., Keane, O.M., Toft, C., Carretero-Paulet, L., and Jones, G.W. (2013). The roles of whole-genome and small-scale duplications in the functional specialization of Saccharomyces cerevisiae genes. PLoS Genet *9*, e1003176.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S.*, et al.* (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res *43*, D805-811.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Gibson, T.A., and Goldberg, D.S. (2009). Questioning the ubiquity of neofunctionalization. PLoS Comput Biol *5*, e1000252.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E.*, et al.* (2003). A protein interaction map of Drosophila melanogaster. Science *302*, 1727-1736.

Guan, Y., Dunham, M.J., and Troyanskaya, O.G. (2007). Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics *175*, 933-943.

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

He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics *169*, 1157-1164.

Hegele, A., Kamburov, A., Grossmann, A., Sourlis, C., Wowro, S., Weimann, M., Will, C.L., Pena, V., Luhrmann, R., and Stelzl, U. (2012). Dynamic protein-protein interaction wiring of the human spliceosome. Mol Cell *45*, 567-580.

Holoch, D., and Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet *16*, 71-84.

Johnson, A.E., Chen, J.S., and Gould, K.L. (2013). CK1 is required for a mitotic checkpoint that delays cytokinesis. Curr Biol *23*, 1920-1926.

Kamada, K., Shu, F., Chen, H., Malik, S., Stelzer, G., Roeder, R.G., Meisterernst, M., and Burley, S.K. (2001). Crystal structure of negative cofactor 2 recognizing the TBP-DNA transcription complex. Cell *106*, 71-81.

Kasahara, M. (2007). The 2R hypothesis: an update. Curr Opin Immunol *19*, 547-552.

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

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science *286*, 295-299.

Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res *11*, 2120-2126.

Motamedi, M.R., Verdel, A., Colmenares, S.U., Gerber, S.A., Gygi, S.P., and Moazed, D. (2004). Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. Cell *119*, 789-802.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods *9*, 471-472.

Papp, B., Pal, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. Nature *424*, 194-197.

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S.J. (2009). Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res *37*, 825-831.

Rensing, S.A. (2014). Gene duplication as a driver of plant morphogenetic evolution. Curr Opin Plant Biol *17*, 43-48.

Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. (2005). Inferring protein domain interactions from databases of interacting proteins. Genome Biol *6*, R89.

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

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

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y.*, et al.* (2015). Widespread macromolecular interaction perturbations in human genetic disorders. Cell *161*, 647-660.

Shiozaki, K., and Russell, P. (1996). Conjugation, meiosis, and the osmotic stress response are regulated by Spc1 kinase through Atf1 transcription factor in fission yeast. Genes Dev *10*, 2276-2288.

Sipiczki, M. (2000). Where does fission yeast sit on the tree of life? Genome Biol *1*, REVIEWS1011.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S.*, et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957-968.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet *133*, 1-9.

Talavera, D., Lovell, S.C., and Whelan, S. (2015). Covariation Is a Poor Measure of Molecular Coevolution. Mol Biol Evol *32*, 2456-2468.

Tardiff, D.F., Jui, N.T., Khurana, V., Tambe, M.A., Thompson, M.L., Chung, C.Y., Kamadurai, H.B., Kim, H.T., Lancaster, A.K., Caldwell, K.A.*, et al.* (2013). Yeast reveal a "druggable" Rsp5/Nedd4 network that ameliorates alpha-synuclein toxicity in neurons. Science *342*, 979-983.

Ting, C.T., Tsaur, S.C., Sun, S., Browne, W.E., Chen, Y.C., Patel, N.H., and Wu, C.I. (2004). Gene duplication and speciation in Drosophila: evidence from the Odysseus locus. Proc Natl Acad Sci U S A *101*, 12232-12235.

Verdel, A., Jia, S., Gerber, S., Sugiyama, T., Gygi, S., Grewal, S.I., and Moazed, D. (2004). RNAi-mediated targeting of heterochromatin by the RITS complex. Science *303*, 672-676.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol *30*, 159-164.

Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet *10*, e1004819.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S.*, et al.* (2002). The genome sequence of Schizosaccharomyces pombe. Nature *415*, 871-880.

Woolcock, K.J., Stunnenberg, R., Gaidatzis, D., Hotz, H.R., Emmerth, S., Barraud, P., and Buhler, M. (2012). RNAi keeps Atf1-bound stress response genes in check at nuclear pores. Genes Dev *26*, 683-692.

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

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X.*, et al.* (2011). Next-generation sequencing to generate interactome datasets. Nat Methods *8*, 478-480.

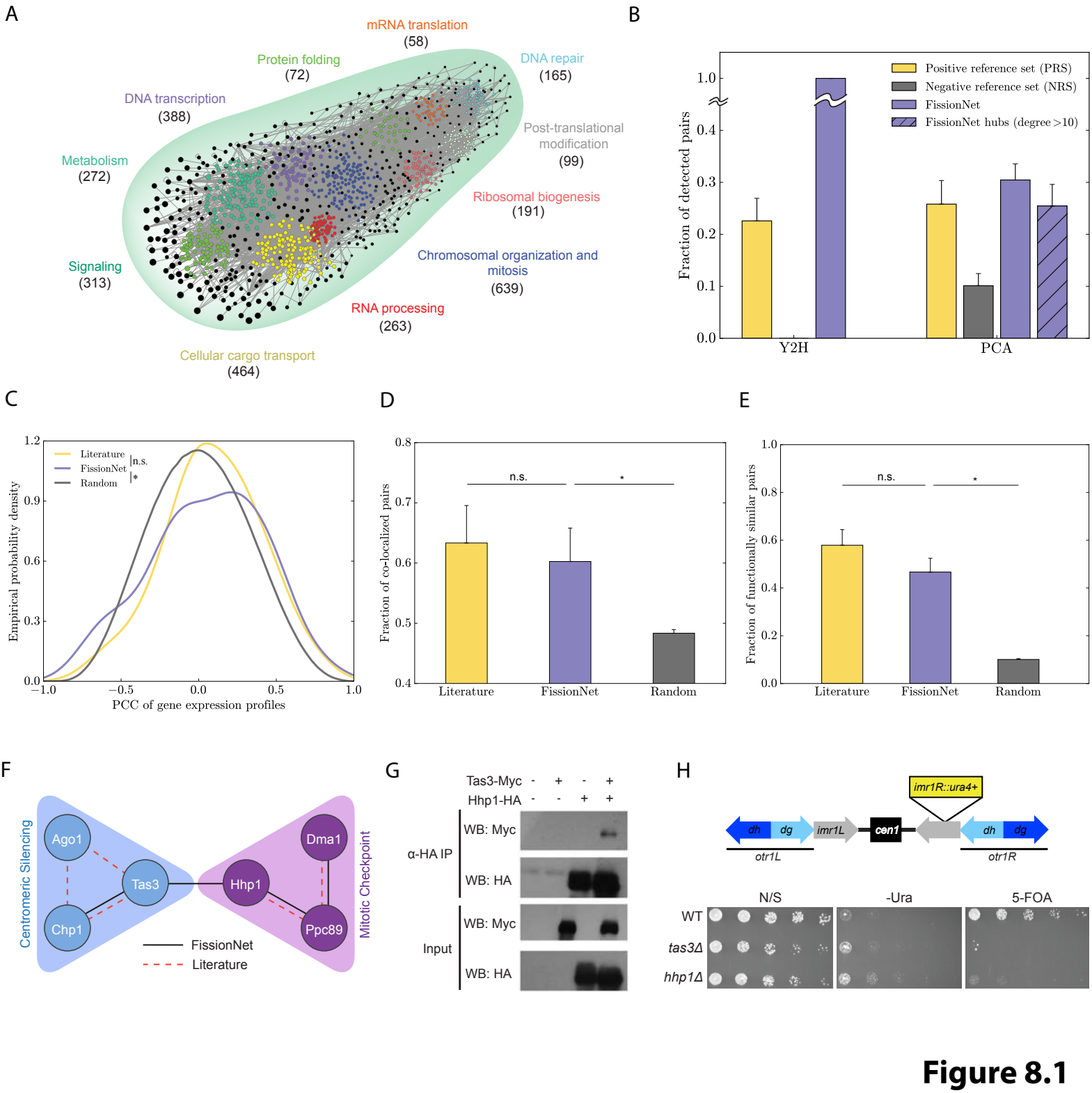

A
mRNA translation
(58)
Protein folding
(72)
DNA repair
(165)
DNA transcription
(388)
Post-translational modification
(99)
Metabolism
(272)
Ribosomal biogenesis
(191)
Signaling
(313)
Chromosomal organization and mitosis
(639)
RNA processing
(263)
Cellular cargo transport
(464)
B
Positive reference set (PRS)
Negative reference set (NRS)
FissionNet
FissionNet hubs (degree >10)
Fraction of detected pairs
Y2H
PCA
C
Literature
FissionNet
Random
n.s.
*
Empirical probability density
PCC of gene expression profiles
D
Fraction of co-localized pairs
n.s.
*
Literature
FissionNet
Random
E
Fraction of functionally similar pairs
n.s.
*
Literature
FissionNet
Random
F
Centromeric Silencing
Ago1
Tas3
Chp1
Hhp1
Dma1
Ppc89
Mitotic Checkpoint
FissionNet
Literature
G
Tas3-Myc
Hhp1-HA
α-HA IP
WB: Myc
WB: HA
Input
WB: Myc
WB: HA
H
imr1R::ura4+
dh
dg
imr1L
cen1
otr1L
otr1R
N/S
-Ura
5-FOA
WT
tas3Δ
hhp1Δ

Figure 8.1

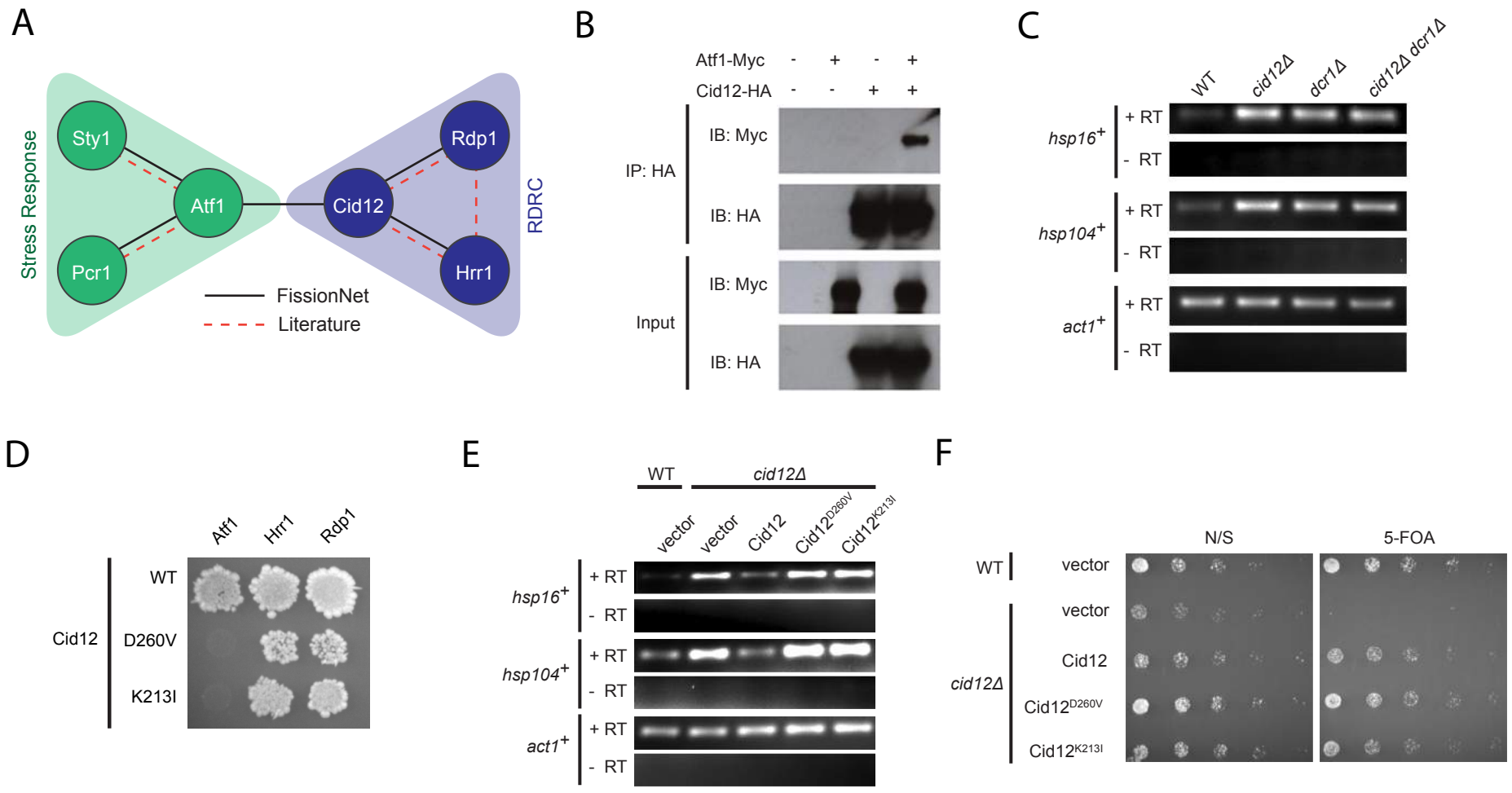

A
Stress Response
Sty1
Atf1
Pcr1
Cid12
Rdp1
Hrr1
RDRC
FissionNet
Literature
B
Atf1-Myc - + - +
Cid12-HA - - + +
IP: HA
IB: Myc
IB: HA
Input
IB: Myc
IB: HA
C
WT
cid12Δ
dcr1Δ
cid12Δ dcr1Δ
hsp16+
hsp104+
act1+
+ RT
- RT
D
Atf1
Hrr1
Rdp1
Cid12
WT
D260V
K213I
E
WT
cid12Δ
vector
vector
Cid12
Cid12D260V
Cid12K213I
hsp16+
hsp104+
act1+
+ RT
- RT
F
N/S
5-FOA
WT
vector
cid12Δ
vector
Cid12
Cid12D260V
Cid12K213I

Figure 8.2

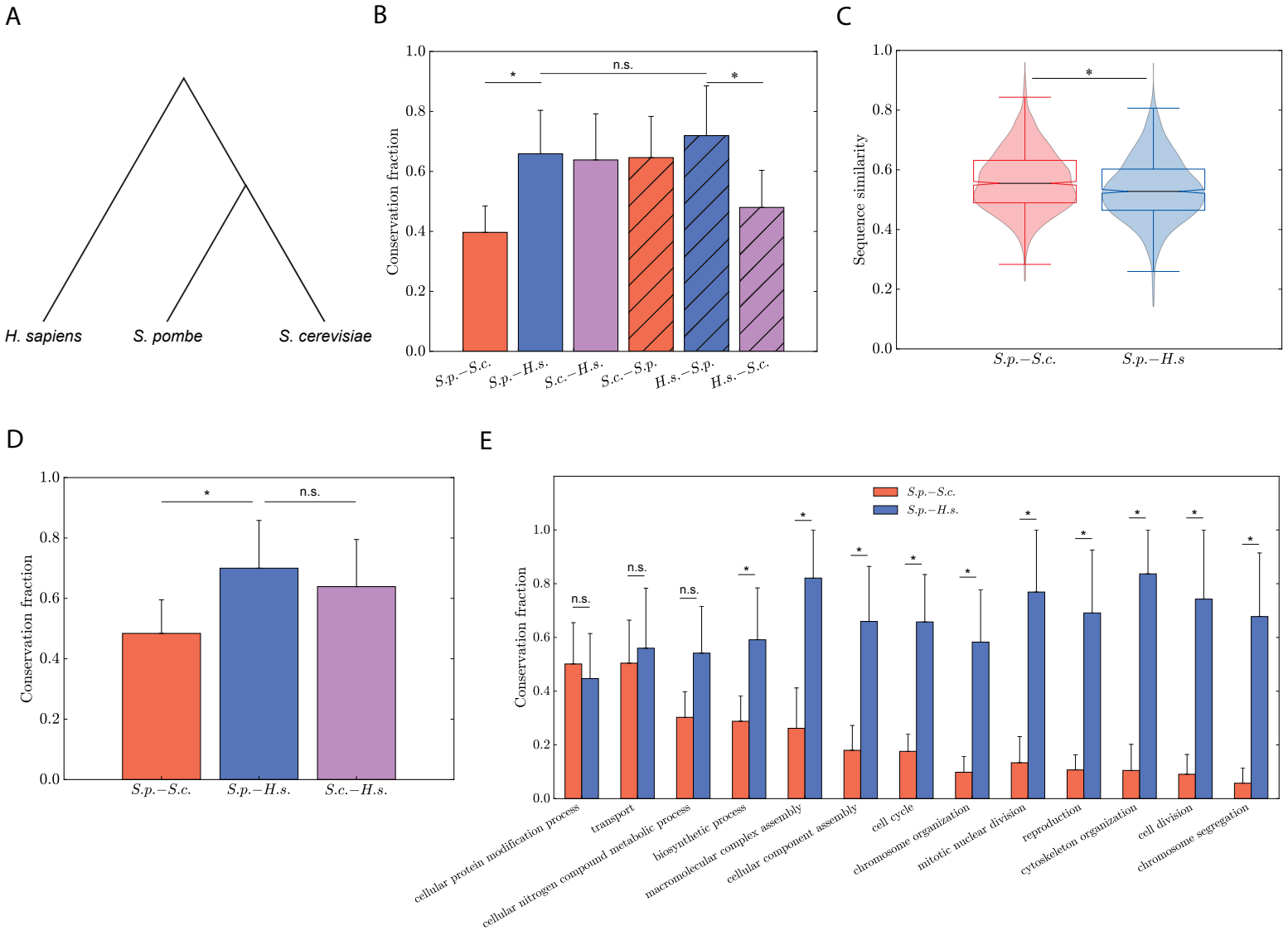

A
H. sapiens
S. pombe
S. cerevisiae
B
Conservation fraction
n.s.
S.p.–S.c.
S.p.–H.s.
S.c.–H.s.
S.c.–S.p.
H.s.–S.p.
H.s.–S.c.
C
Sequence similarity
S.p.–S.c.
S.p.–H.s
D
Conservation fraction
n.s.
S.p.–S.c.
S.p.–H.s.
S.c.–H.s.
E
Conservation fraction
S.p.–S.c.
S.p.–H.s.
n.s.
cellular protein modification process
transport
cellular nitrogen compound metabolic process
biosynthetic process
macromolecular complex assembly
cellular component assembly
cell cycle
chromosome organization
mitotic nuclear division
reproduction
cytoskeleton organization
cell division
chromosome segregation
F
Conservation fraction
S.p.–S.c.
S.p.–H.s.
n = 50
n = 100
n = 200
GO Slim categories with ≥ n genes in each species
G
Srp2
Srp1
Srrm1
Esc1
SPAC 30D11 .14C
Usp104
Cwf15
SPAC 1952 .06C
Proteins:
Interactor-of-interest
Known splice factor
Other
H
Pre-mRNA
T
P
Mature mRNA
T
M
Exon
Intron
srrm1Δ
SPAC30D11.14CΔ
SPAC1952.06CΔ
T P M
> 2,000 S. pombe splice sites
-1.5
1.5
Fold-change (log2)

Figure 8.3

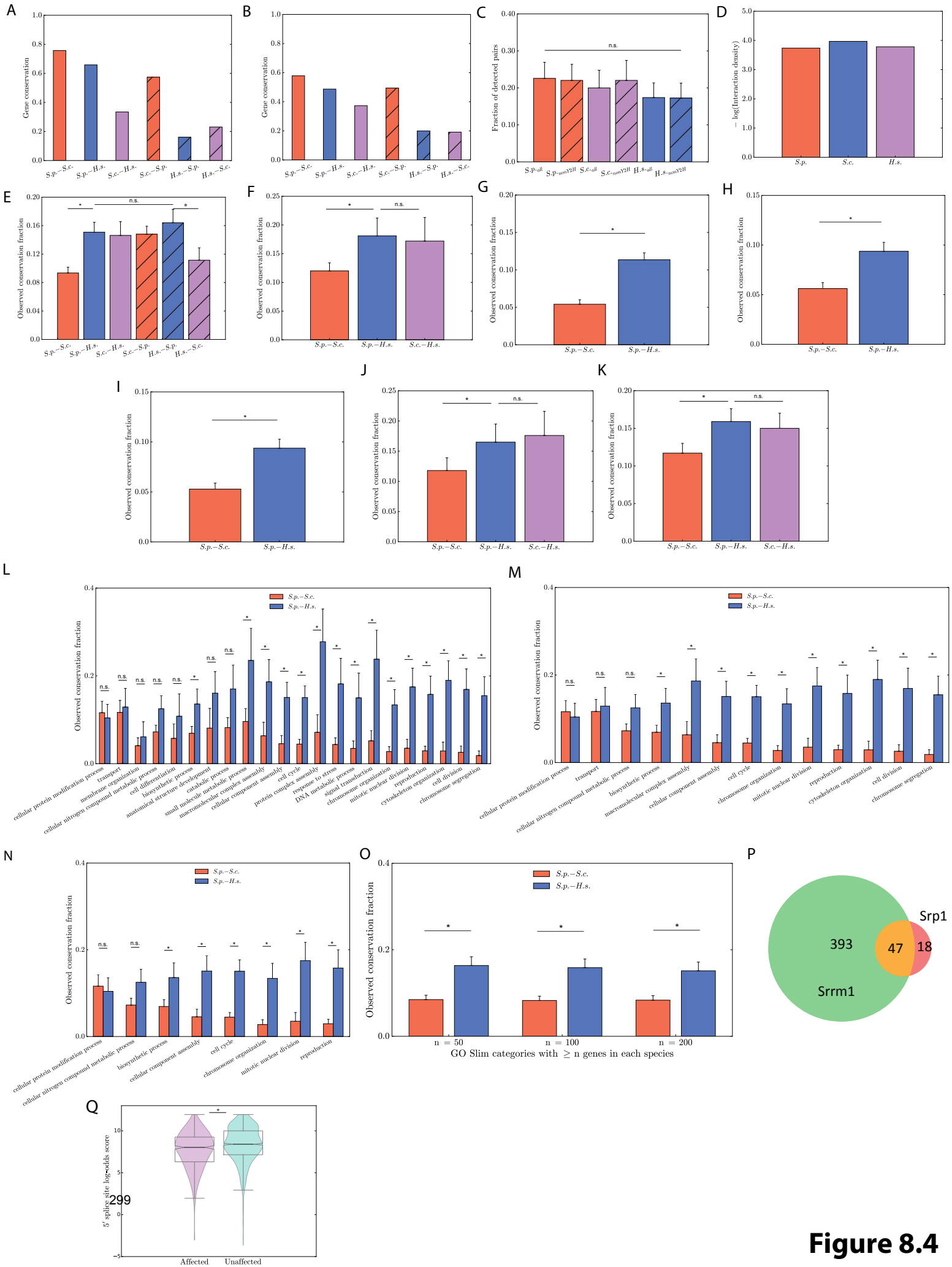
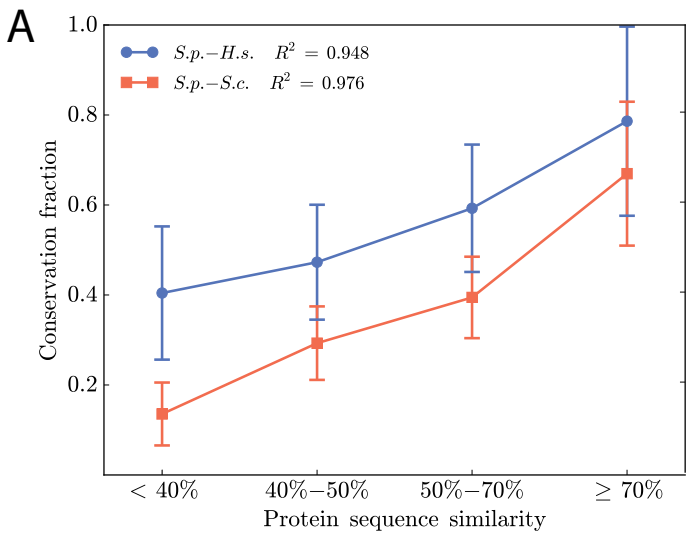
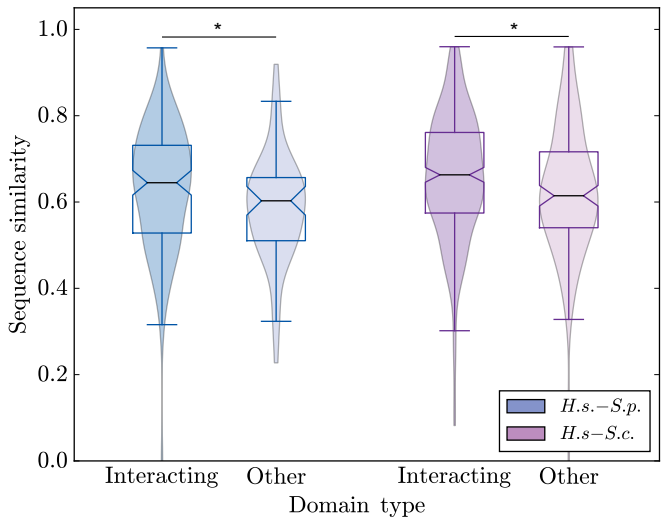
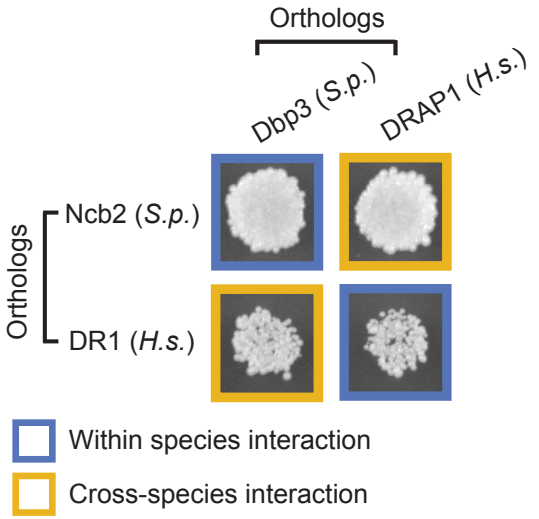
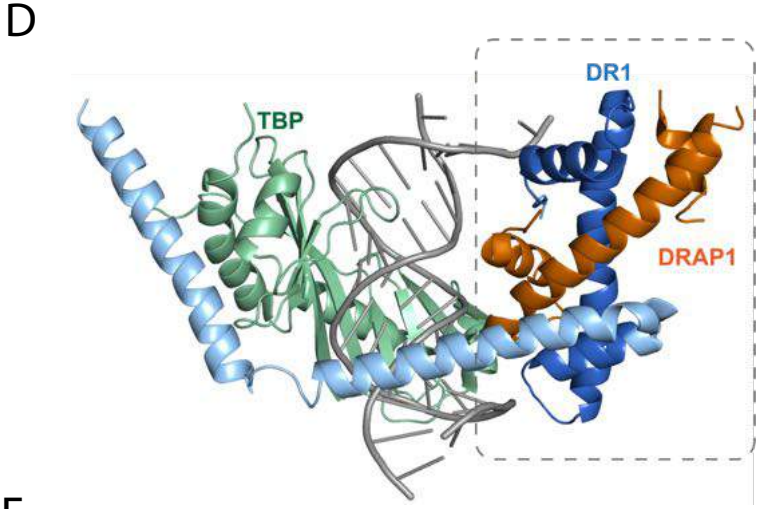

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
Gene conservation
Fraction of detected pairs
− log(Interaction density)
Observed conservation fraction
S.p.−S.c.
S.p.−H.s.
S.c.−H.s.
S.c.−S.p.
H.s.−S.p.
H.s.−S.c.
S.p.-all
S.p.-nonγ2H
S.c.-all
S.c.-nonγ2H
H.s.-all
H.s.-nonγ2H
S.p.
S.c.
H.s.
n.s.
cellular protein modification process
transport
membrane organization
cellular nitrogen compound metabolic process
cell differentiation
biosynthetic process
anatomical structure development
catabolic process
small molecule metabolic process
macromolecular complex assembly
cellular component assembly
protein complex assembly
cell cycle
response to stress
DNA metabolic process
signal transduction
chromosome organization
mitotic nuclear division
reproduction
cytoskeleton organization
cell division
chromosome segregation
GO Slim categories with ≥ n genes in each species
n = 50
n = 100
n = 200
Srp1
Srrm1
393
47
18
5′ splice site log-odds score
299
Affected
Unaffected

Figure 8.4









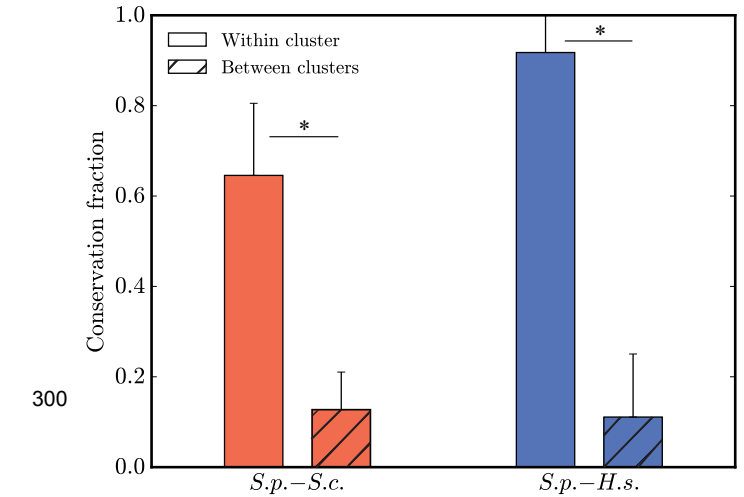

```
                                10        20        30        40        50        60        70        80        90
DR1 (H.s.)      MASSSGNDDDLTIPRAAINKMIKETLP-NVRVANDARELVVNCCTEFIHLISSEANEICNKSEKKTISPEHVIQALESLGFGSYISEVKEVLQEC
                |      ||:|::|:| : ||: : || ::    :||:|:: || |||||:|||||||| |  ||||: ||:|:|||:| |  ||:|  ||  |
Ncb2 (S.p.)     MNDGFA-DDELSLPKATVQKMVSDILPVDLTFTKEARDLLIECCVEFIHLVSSEANEICEKEAKKTIAAEHIIKALENLEFKEYIAEALEVAAEH

                     100       110       120       130       140       150       160       170
DR1 (H.s.)      KTVALKRRKASSRLENLGIPEEELLRQQQELFAKARQQQAELAQQEWLQMQQAAQQAQLAAA----SASASNQAGSSQDEEDDDDI
                |     | | ||: |  |:  :||||||:|| ::||::          : |  |       |    :||          | :  :
Ncb2 (S.p.)     KEQQKNREKKSSKFEQSGVSRDELLRQQEELLSRARER---------FKNQNIAHDNHTTTAIPVPTAS----------ETETKEN

                                10        20        30        40        50        60        70        80
DRAP1 (H.s.)    M-----------PSKKKKYNARFPPARIKKIMQTDEEIGKVAAAVPVIISRALELFLESLLKKACQVTQSRNAKTMTTSHLKQCIELEQQFDFL
                |           |:    : :||| |||||||| |:::|||||   |||:|:|||||::|:::::|: |:   || :| ||||  ::  :|||||
Dpb3 (S.p.)     MGDPTNNNDSETKPNPATYWKSRFPVARIKKIMQADQDVGKVAQVTPVIMSKALELFMQSIIQESCKQTRLHQAKRVTVSHLKHAVQSVEQFDFL

                     90        100       110       120       130       140       150       160
DRAP1 (H.s.)    KDLVASVPDMQGDGEDNHMDGDKGARRGRKPGSGGRKNGGMGTKS------KDKKLSGTDSEQEDESEDTDTDGEEETSQPPPQASHPS------
                :|:|  |||             |  |: ::| :    | |   :|      |   :   : |::||| |:|  :|| :    :| |
Dpb3 (S.p.)     QDIVEKVPDAPPI---------KAERKTKRPRARRAANEGEHNESVPAKKVKKNTVKEEEIEKDDESATTETPVKEEPTGEETEADHEEKASVDH

                   170       180       190       200
DRAP1 (H.s.)    AHFQSPPTPFLPFASTLPLPPAPPGPSAPDEEDEEDYDS
                 :|    |                    |      |
Dpb3 (S.p.)     GNFSDKTT--------------------SEASSASGDE
```



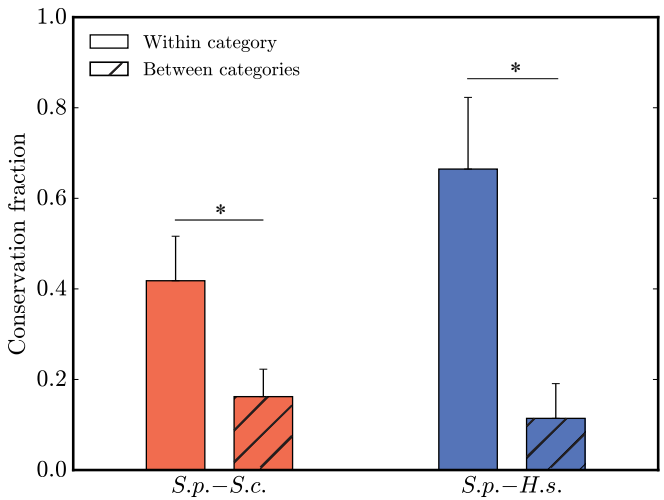


Figure 8.5

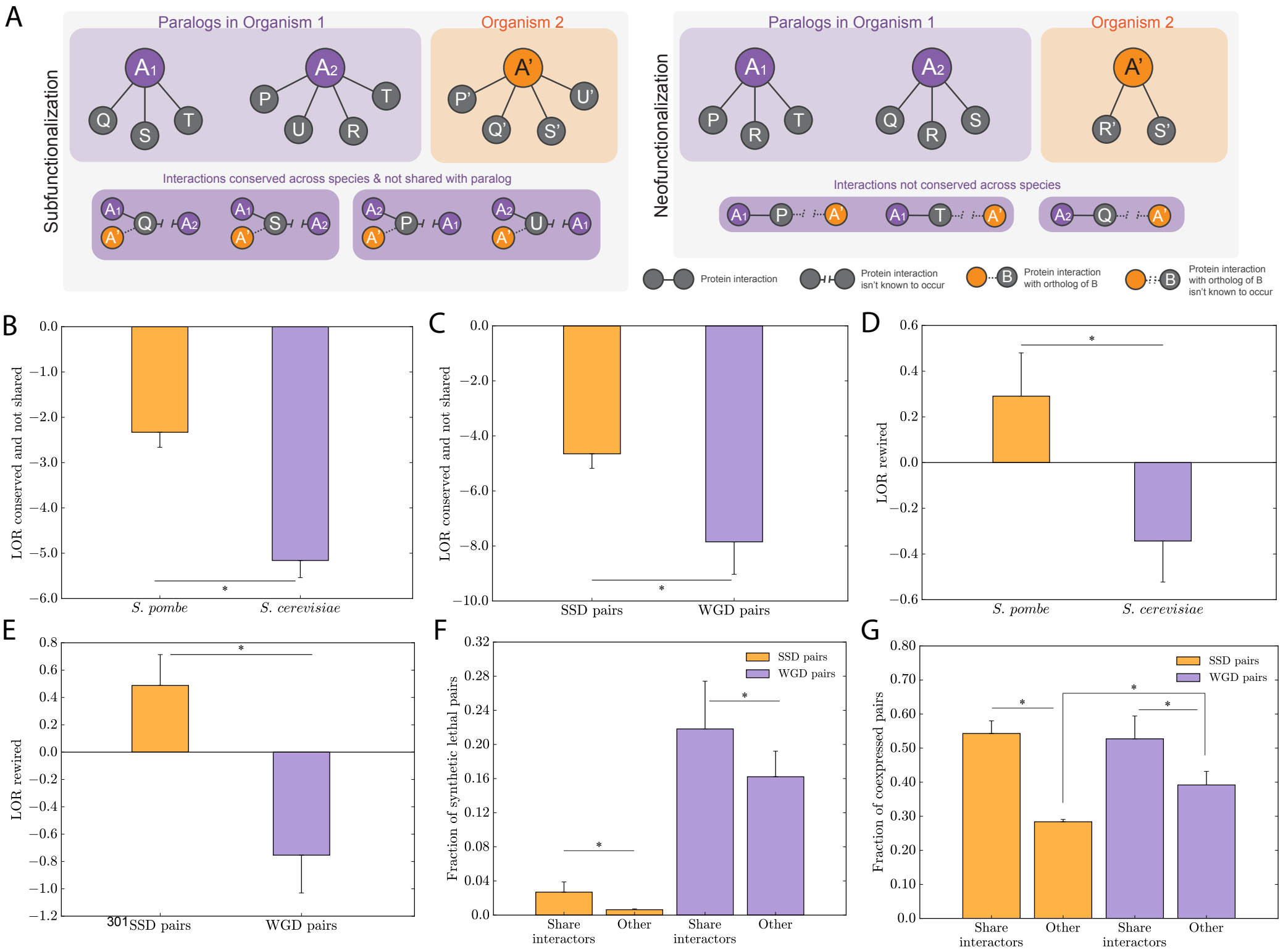

A
Subfunctionalization
Paralogs in Organism 1
Organism 2
Interactions conserved across species & not shared with paralog
Neofunctionalization
Paralogs in Organism 1
Organism 2
Interactions not conserved across species
Protein interaction
Protein interaction isn't known to occur
Protein interaction with ortholog of B
Protein interaction with ortholog of B isn't known to occur
B
LOR conserved and not shared
S. pombe
S. cerevisiae
C
LOR conserved and not shared
SSD pairs
WGD pairs
D
LOR rewired
S. pombe
S. cerevisiae
E
LOR rewired
SSD pairs
WGD pairs
F
Fraction of synthetic lethal pairs
SSD pairs
WGD pairs
Share interactors
Other
Share interactors
Other
G
Fraction of coexpressed pairs
SSD pairs
WGD pairs
Share interactors
Other
Share interactors
Other

Figure 8.6

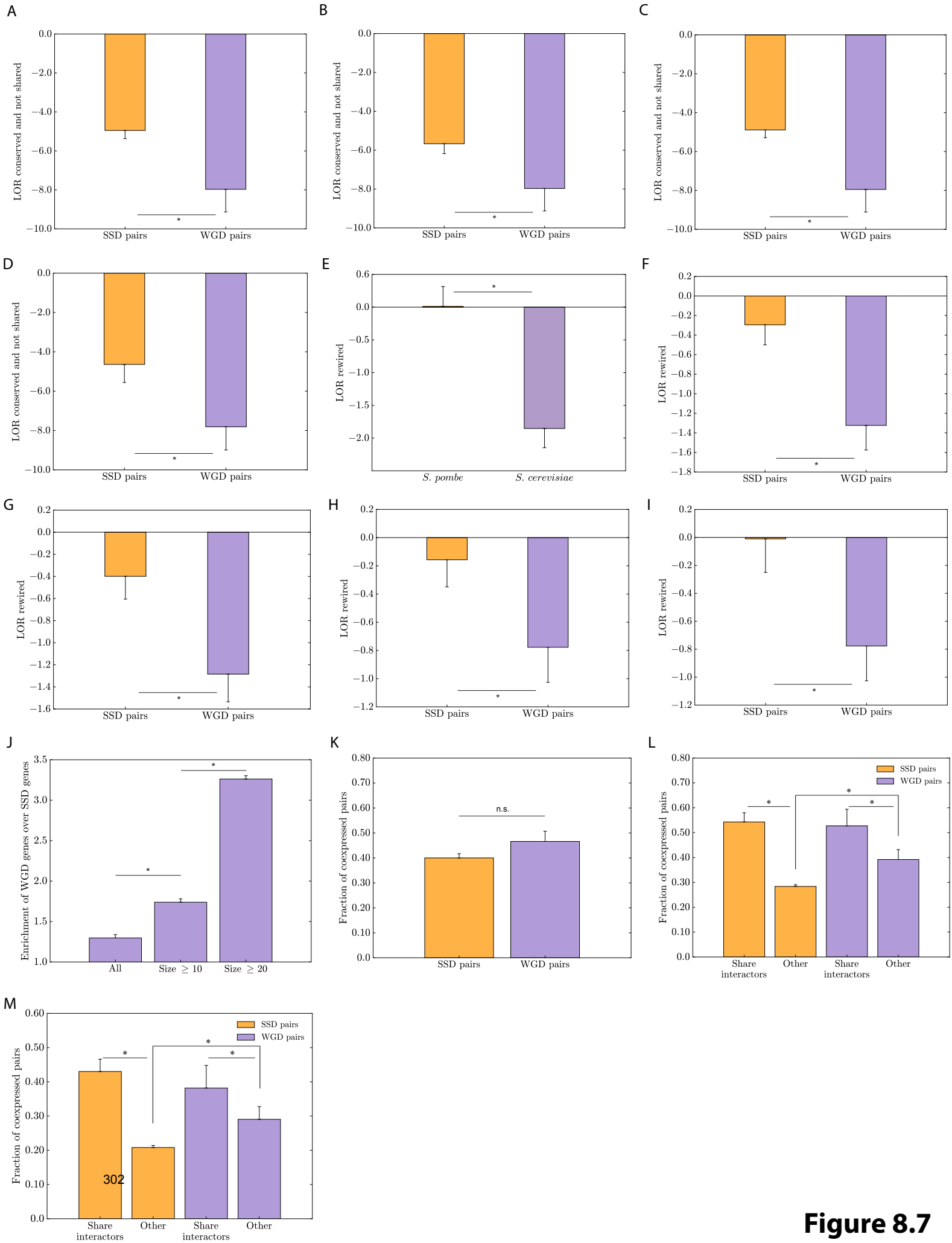

A
B
C
D
E
F
G
H
I
J
K
L
M
LOR conserved and not shared
LOR rewired
SSD pairs
WGD pairs
S. pombe
S. cerevisiae
Enrichment of WGD genes over SSD genes
All
Size ≥ 10
Size ≥ 20
Fraction of coexpressed pairs
n.s.
Share interactors
Other
302

Figure 8.7

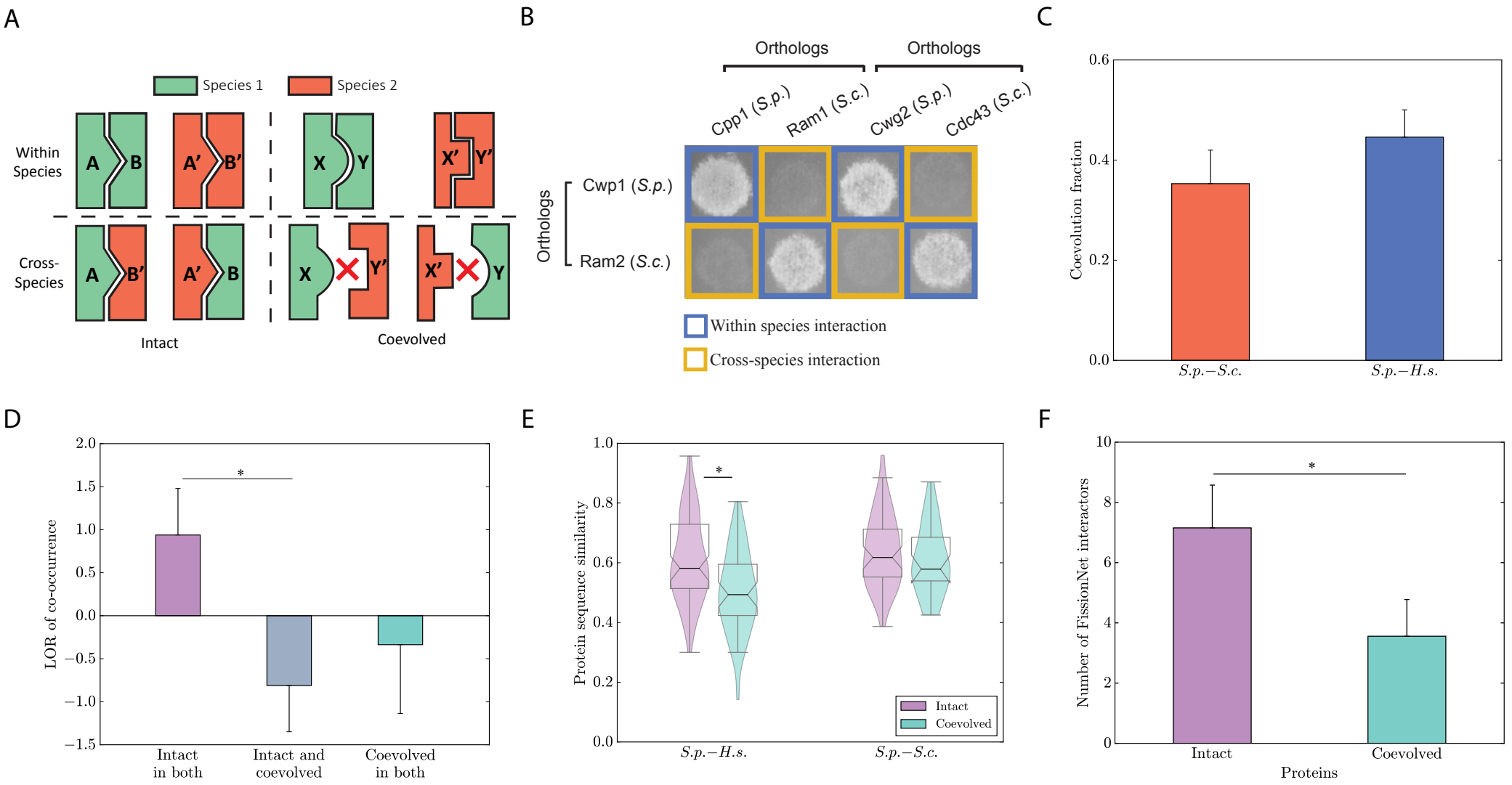

A
Species 1
Species 2
Within Species
Cross-Species
A
B
A'
B'
X
Y
X'
Y'
Intact
Coevolved
B
Orthologs
Orthologs
Cpp1 (S.p.)
Ram1 (S.c.)
Cwg2 (S.p.)
Cdc43 (S.c.)
Orthologs
Cwp1 (S.p.)
Ram2 (S.c.)
Within species interaction
Cross-species interaction
C
Coevolution fraction
0.0
0.2
0.4
0.6
S.p.−S.c.
S.p.−H.s.
D
LOR of co-occurrence
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
*
Intact in both
Intact and coevolved
Coevolved in both
E
Protein sequence similarity
1.0
0.8
0.6
0.4
0.2
0.0
*
Intact
Coevolved
S.p.−H.s.
S.p.−S.c.
F
Number of FissionNet interactors
10
8
6
4
2
0
*
Intact
Coevolved
Proteins

Figure 8.8

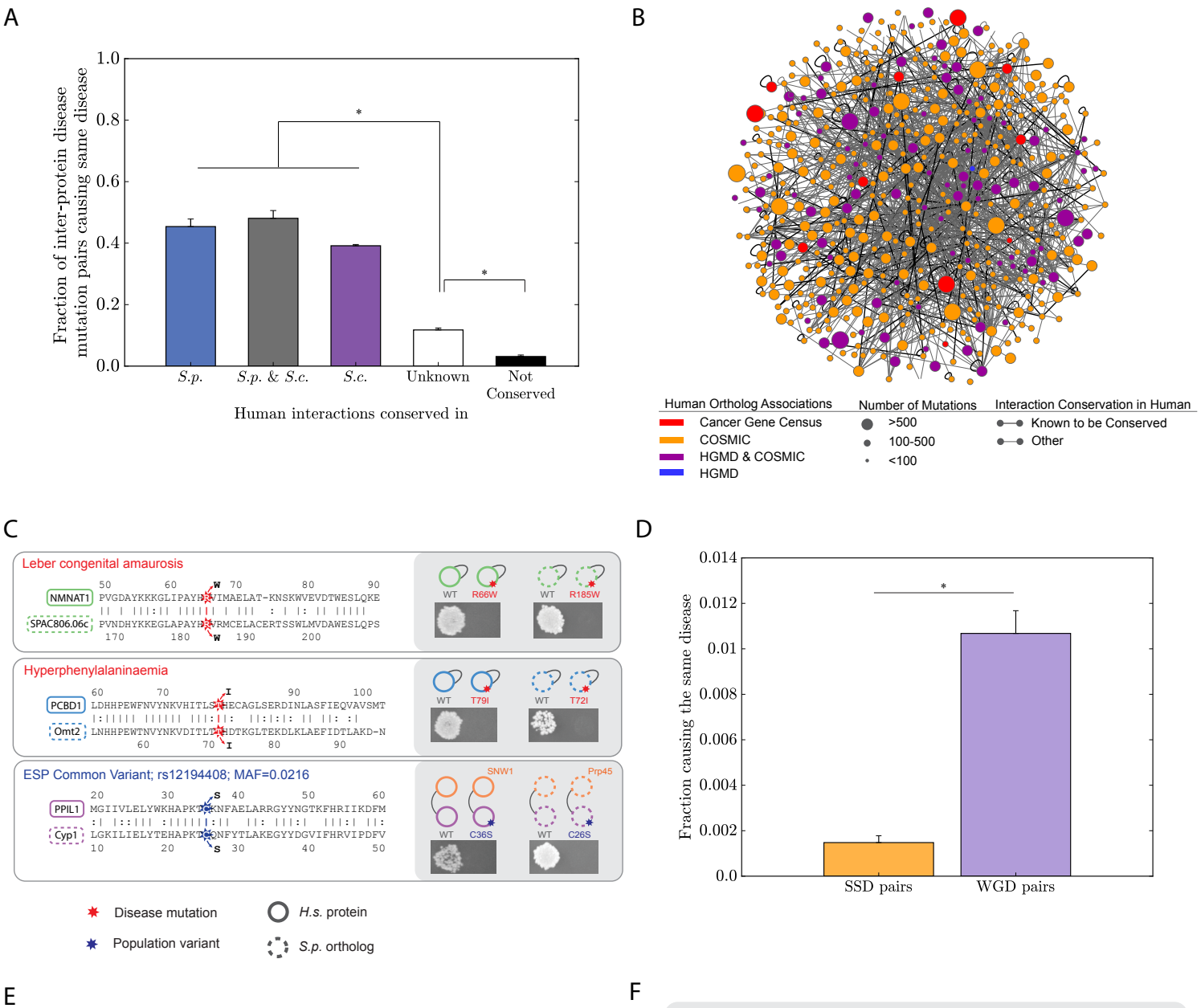

A
Fraction of inter-protein disease mutation pairs causing same disease
1.0
0.8
0.6
0.4
0.2
0.0
*
*
S.p.
S.p. & S.c.
S.c.
Unknown
Not Conserved
Human interactions conserved in
B
Human Ortholog Associations
Cancer Gene Census
COSMIC
HGMD & COSMIC
HGMD
Number of Mutations
>500
100-500
<100
Interaction Conservation in Human
Known to be Conserved
Other
C
Leber congenital amaurosis
NMNAT1
SPAC806.06c
WT
R66W
R185W
Hyperphenylalaninaemia
PCBD1
Dmt2
T79I
T72I
ESP Common Variant; rs12194408; MAF=0.0216
PPIL1
Cyp1
SNW1
Prp45
C36S
C26S
Disease mutation
Population variant
H.s. protein
S.p. ortholog
D
Fraction causing the same disease
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0.0
*
SSD pairs
WGD pairs
E
Vertebrate WGD Events
2R
1R
Human
Chimpanzee
Mouse
Chicken
Frog
Fugu
Zebrafish
Lamprey
Fly
Terrestrial Vertebrates
Aquatic Vertebrates
Invertebrates
Protostomes
Continual accumulation of SSDs
F
WGD
Gene Loss
A B C D E F G H ...
A' B' C' D' E' F' G' H'
Stoichometric constraints lead to functional redundancy
SSD
Divergence
A B C D
A B B' C D
Initial functional redundancy
Subfunctionalization and/or Neofunctionalization

Figure 8.9

Table 8.1. List of the positive reference set (PRS).

ORF_A	ORF_B	Y2H_status	PCA_status
SPAC1002.06C	SPBC1778.02	Negative	Negative
SPAC110.03	SPAC22H10.07	Negative	Negative
SPAC110.03	SPAC24H6.09	Negative	Negative
SPAC110.03	SPBC1289.04C	Negative	Negative
SPAC110.03	SPBC1604.14C	Negative	Negative
SPAC13C5.07	SPAC13C5.07	Positive	Negative
SPAC13C5.07	SPBC6B1.09C	Negative	Negative
SPAC1565.06C	SPAC222.10C	Positive	Negative
SPAC1565.06C	SPBC21.06C	Positive	Negative
SPAC16.02C	SPBC530.14C	Positive	Positive
SPAC16A10.06C	SPCC5E4.06	Negative	Positive
SPAC16A10.07C	SPAC16A10.07C	Negative	Positive
SPAC16A10.07C	SPBC1778.02	Negative	Negative
SPAC16E8.09	SPAC22H10.07	Positive	Negative
SPAC17H9.09C	SPBC1D7.05	Negative	Positive
SPAC1834.04	SPBC428.08C	Negative	Positive
SPAC18G6.02C	SPBC83.03C	Positive	Positive
SPAC19A8.12	SPBC3B9.21	Positive	Negative
SPAC19D5.01	SPAC24B11.06C	Negative	Negative
SPAC222.10C	SPAC6F6.08C	Positive	Negative
SPAC22E12.07	SPAC22E12.07	Negative	Negative
SPAC22F3.09C	SPBC2F12.11C	Positive	Negative
SPAC22F3.09C	SPBC336.12C	Positive	Negative
SPAC22H10.07	SPBC1604.14C	Negative	Negative
SPAC23A1.06C	SPAC24B11.06C	Negative	Negative
SPAC23C11.16	SPCC4B3.15	Negative	Negative
SPAC23C4.15	SPBC28F2.12	Negative	Negative
SPAC23C4.15	SPCC1442.10C	Negative	Negative
SPAC23C4.18C	SPAC6B12.11	Negative	Negative
SPAC23G3.01	SPCC1442.10C	Negative	Negative
SPAC23H3.13C	SPBC19C7.03	Negative	Negative
SPAC24B11.06C	SPBC29B5.01	Negative	Positive
SPAC24B11.06C	SPBC409.07C	Positive	Positive
SPAC24B11.11C	SPBC428.13C	Negative	Negative
SPAC24B11.11C	SPCC1739.11C	Negative	Negative
SPAC24H6.05	SPCC1259.13	Negative	Negative
SPAC24H6.05	SPCC18B5.11C	Negative	Negative
SPAC26H5.06	SPAC26H5.06	Negative	Negative
SPAC27E2.05	SPBC1734.02C	Negative	Negative
SPAC27F1.09C	SPBC146.07	Negative	Negative
SPAC30.03C	SPCC736.09C	Negative	Negative
SPAC30D11.10	SPAC30D11.10	Negative	Negative
SPAC30D11.10	SPAC644.14C	Positive	Negative
SPAC3A12.07	SPCC1442.10C	Negative	Negative
SPAC3F10.01	SPBC25D12.03C	Negative	Negative
SPAC637.07	SPBC646.09C	Negative	Positive
SPAC644.06C	SPCC18B5.03	Negative	Positive

SPAC644.14C	SPAC644.14C	Positive	Positive
SPAC664.01C	SPAC664.01C	Negative	Positive
SPAC694.06C	SPCC18B5.11C	Negative	Negative
SPAC6F12.09	SPBC83.03C	Negative	Positive
SPAC6G9.13C	SPBC1778.02	Negative	Negative
SPAC8E11.02C	SPCC1259.13	Negative	Negative
SPAC8E11.03C	SPAC8E11.03C	Positive	Positive
SPAC9E9.08	SPBC216.05	Negative	Negative
SPAP8A3.06	SPBC146.07	Positive	Negative
SPBC11B10.09	SPBC14C8.07C	Negative	Negative
SPBC11B10.09	SPBC32F12.09	Negative	Positive
SPBC11B10.09	SPCC18B5.03	Negative	Positive
SPBC11C11.08	SPBC530.14C	Negative	Positive
SPBC1289.02C	SPBC146.07	Negative	Negative
SPBC146.03C	SPBP4H10.06C	Negative	Negative
SPBC146.07	SPBC530.14C	Negative	Positive
SPBC14C8.12	SPCC1442.10C	Negative	Negative
SPBC14F5.08	SPBC31F10.04C	Negative	Positive
SPBC1604.14C	SPBC1604.14C	Negative	Negative
SPBC16D10.09	SPBC1734.02C	Negative	Negative
SPBC16H5.11C	SPBC16H5.11C	Negative	Negative
SPBC1703.06	SPBC409.05	Negative	Negative
SPBC1778.06C	SPBC32H8.12C	Negative	Negative
SPBC1921.02	SPCC18B5.11C	Negative	Negative
SPBC211.04C	SPBC25D12.03C	Negative	Negative
SPBC216.05	SPCC1259.13	Negative	Negative
SPBC216.05	SPCC18B5.11C	Negative	Negative
SPBC216.06C	SPBC30D10.04	Positive	Negative
SPBC244.01C	SPBC244.01C	Positive	Negative
SPBC25D12.03C	SPBC4.04C	Negative	Negative
SPBC28F2.07	SPBC409.03	Negative	Negative
SPBC28F2.12	SPCC1020.04C	Negative	Positive
SPBC28F2.12	SPCC1442.10C	Positive	Negative
SPBC4.04C	SPBC776.12C	Negative	Positive
SPBC409.03	SPBC409.03	Negative	Negative
SPBC409.05	SPCC18.04	Negative	Positive
SPBC646.14C	SPBC685.09	Negative	Positive
SPBC6B1.09C	SPCC338.08	Negative	Negative
SPBC725.02	SPBC887.10	Negative	Negative
SPBC776.12C	SPCC550.13	Negative	Negative
SPCC11E10.08	SPCC613.12C	Positive	Negative
SPCC1223.06	SPCC1223.06	Negative	Negative
SPCC1739.03	SPCC663.12	Negative	Negative
SPCC18B5.03	SPCC18B5.11C	Negative	Negative
SPAC15A10.03C	SPAC644.14C	Positive	Positive
SPBC1706.01	SPCC1223.06	Positive	Negative

Table 8.1. List of the negative reference set (NRS).

ORF_A	**ORF_B**	**Y2H_status**	**PCA_status**

SPAC1002.17C	SPAC2F3.09	Negative	Negative
SPAC1006.03C	SPCC1494.07	Negative	Negative
SPAC1142.07C	SPAC186.06	Negative	Negative
SPAC11D3.01C	SPBC11G11.06C	Negative	Positive
SPAC12B10.06C	SPBC800.03	Negative	Positive
SPAC13G7.10	SPBC17D1.02	Negative	Positive
SPAC1486.08	SPAC1805.06C	Negative	Negative
SPAC1705.03C	SPBC776.08C	Negative	Negative
SPAC17A2.08C	SPCC14G10.01	Negative	Negative
SPAC17C9.15C	SPBC31F10.02	Negative	Negative
SPAC1834.04	SPAC922.05C	Negative	Negative
SPAC186.07C	SPCC70.10	Negative	Negative
SPAC18B11.08C	SPAPB17E12.07C	Negative	Negative
SPAC18G6.02C	SPAC644.06C	Negative	Negative
SPAC18G6.13	SPAC1952.01	Negative	Negative
SPAC1952.02	SPAC23G3.04	Negative	Negative
SPAC1B1.01	SPBC16D10.01C	Negative	Negative
SPAC1B3.18C	SPAC890.05	Negative	Negative
SPAC1F7.11C	SPCC1840.10	Negative	Negative
SPAC1F8.02C	SPBC800.02	Negative	Negative
SPAC20G8.06	SPCC613.10	Negative	Negative
SPAC20G8.07C	SPBC19G7.13	Negative	Negative
SPAC20H4.01	SPAPB1E7.04C	Negative	Negative
SPAC227.16C	SPBC713.08	Negative	Negative
SPAC22A12.08C	SPCC1902.02	Negative	Negative
SPAC23A1.12C	SPAC458.06	Negative	Negative
SPAC23C4.07	SPBC660.08	Negative	Negative
SPAC23H4.01C	SPBC25H2.14	Negative	Negative
SPAC26A3.15C	SPBC211.09	Negative	Positive
SPAC31G5.09C	SPCC4G3.06C	Negative	Negative
SPAC323.02C	SPBC685.03	Negative	Negative
SPAC3G6.03C	SPAC4F10.05C	Negative	Negative
SPAC3H1.04C	SPCC663.14C	Negative	Negative
SPAC4A8.07C	SPAC890.08	Negative	Negative
SPAC4D7.13	SPCC965.07C	Negative	Negative
SPAC56F8.04C	SPAC869.10C	Negative	Negative
SPAC57A10.06	SPBC2D10.04	Negative	Negative
SPAC589.11	SPBC409.16C	Negative	Negative
SPAC5H10.10	SPCP25A2.02C	Negative	Negative
SPAC688.11	SPCC830.09C	Negative	Positive
SPAC688.15	SPBP22H7.03	Negative	Negative
SPAC806.07	SPCC132.03	Negative	Negative
SPAC821.10C	SPBC26H8.11C	Negative	Negative
SPAC821.11	SPBC2D10.04	Negative	Negative
SPAC869.03C	SPBC1683.03C	Negative	Negative
SPAC890.06	SPBC609.02	Negative	Negative
SPAC8C9.02	SPBPB2B2.01	Negative	Negative
SPAC8C9.10C	SPBC21H7.02	Negative	Negative
SPAC922.04	SPBC25B2.02C	Negative	Negative

SPAC9G1.05	SPBC1198.10C	Negative	Negative
SPAP14E8.04	SPBC25H2.07	Negative	Negative
SPAP27G11.03	SPBP35G2.13C	Negative	Negative
SPAP7G5.02C	SPBC1348.01	Negative	Negative
SPAP7G5.06	SPBC649.04	Negative	Negative
SPBC1105.03C	SPBC36B7.04	Negative	Negative
SPBC1105.16C	SPCC965.10	Negative	Negative
SPBC1215.01	SPBC336.04	Negative	Negative
SPBC1604.10	SPBC17A3.08	Negative	Negative
SPBC1685.01	SPCC74.04	Negative	Negative
SPBC16G5.17	SPCC1620.08	Negative	Negative
SPBC1709.04C	SPCC1223.02	Negative	Negative
SPBC17A3.08	SPBC3B8.07C	Negative	Negative
SPBC25H2.07	SPBC354.04	Negative	Positive
SPBC25H2.08C	SPBC4C3.02C	Negative	Negative
SPBC27B12.14	SPBC800.13	Negative	Negative
SPBC28E12.03	SPCC550.02C	Negative	Negative
SPBC2G2.05	SPBC3D6.04C	Negative	Positive
SPBC32H8.13C	SPCC1620.07C	Negative	Negative
SPBC336.05C	SPCC70.10	Negative	Negative
SPBC3B8.04C	SPBP8B7.20C	Negative	Negative
SPBC3B8.08	SPCC584.15C	Negative	Positive
SPBC4C3.06	SPCC297.05	Negative	Negative
SPBC947.10	SPCC16C4.12	Negative	Negative
SPBP8B7.24C	SPCC594.02C	Negative	Negative
SPBP8B7.25	SPCC2H8.05C	Negative	Negative
SPCC320.06	SPCPB16A4.06C	Negative	Negative
SPCC417.05C	SPCC4F11.01	Negative	Negative
SPCC622.21	SPCP1E11.06	Negative	Negative
SPAC1002.03C	SPBC409.20C	Negative	Negative
SPAC1039.08	SPCC417.12	Negative	Negative
SPAC1071.02	SPAC24B11.05	Negative	Negative
SPAC1071.04C	SPAC3A12.06C	Negative	Negative
SPAC1071.05	SPCC63.06	Negative	Negative
SPAC11D3.18C	SPAC212.08C	Negative	Negative
SPAC11G7.06C	SPCC11E10.04	Negative	Negative
SPAC12G12.05C	SPAC23C4.15	Negative	Negative
SPAC13C5.01C	SPCP1E11.10	Negative	Negative
SPAC13C5.05C	SPAC17H9.09C	Negative	Positive
SPAC144.03	SPBC577.05C	Negative	Negative
SPAC16.01	SPAC4H3.08	Negative	Negative
SPAC1783.07C	SPAC328.03	Negative	Negative
SPAC17D4.01	SPAC1F12.08	Negative	Positive
SPAC186.07C	SPBC25H2.07	Negative	Negative
SPAC18B11.08C	SPBC83.12	Negative	Negative
SPAC19A8.10	SPBC15D4.10C	Negative	Negative
SPAC19B12.06C	SPAC9G1.07	Negative	Negative
SPAC19G12.04	SPBC577.05C	Negative	Negative
SPAC19G12.05	SPBC19G7.18C	Negative	Negative

SPAC1B2.02C	SPBC660.08	Negative	Negative
SPAC1F7.08	SPBC1734.12C	Negative	Negative
SPAC1F8.02C	SPAPB8E5.03	Negative	Negative
SPAC222.07C	SPBC8D2.18C	Negative	Negative
SPAC227.17C	SPAC3G6.09C	Negative	Negative
SPAC22E12.11C	SPCC1682.13	Negative	Negative
SPAC22G7.10	SPAC8E11.11	Negative	Negative
SPAC23C4.11	SPCC645.02	Negative	Negative
SPAC23H4.07C	SPCC1020.11C	Negative	Negative
SPAC24B11.04C	SPBC1709.04C	Negative	Positive
SPAC2F3.13C	SPBC1677.03C	Negative	Negative
SPAC2H10.02C	SPBC1604.10	Negative	Negative
SPAC30.02C	SPCC663.09C	Negative	Negative
SPAC30.03C	SPBC691.02C	Negative	Negative
SPAC31A2.02	SPAPB1E7.10	Negative	Negative
SPAC31A2.03	SPAC4G9.17C	Negative	Negative
SPAC323.06C	SPBC685.07C	Negative	Positive
SPAC323.06C	SPCC613.04C	Negative	Negative
SPAC328.10C	SPCC895.09C	Negative	Negative
SPAC3G6.09C	SPBC3B9.03	Negative	Negative
SPAC458.04C	SPAC694.05C	Negative	Negative
SPAC4A8.14	SPBC725.03	Negative	Negative
SPAC4D7.09	SPBC17G9.08C	Negative	Negative
SPAC4G8.08	SPBC1604.07	Negative	Negative
SPAC521.03	SPAC869.01	Negative	Negative
SPAC56E4.03	SPBC17D1.08	Negative	Negative
SPAC57A7.05	SPBC16G5.18	Negative	Negative
SPAC5D6.06C	SPCC285.11	Negative	Negative
SPAC607.08C	SPBC12D12.03	Negative	Positive
SPAC6C3.04	SPBC4.06	Negative	Negative
SPAC6C3.09	SPBP23A10.09	Negative	Negative
SPAC6F12.13C	SPBC1734.15	Negative	Positive
SPAC6F6.05	SPBC557.02C	Negative	Negative
SPAC6G10.08	SPBC29A3.15C	Negative	Negative
SPAC823.01C	SPBC119.15	Negative	Negative
SPAC823.12	SPBC8D2.10C	Negative	Negative
SPAC9.10	SPBC15D4.04	Negative	Negative
SPAP27G11.04C	SPCC364.05	Negative	Negative
SPAP27G11.06C	SPBC1734.10C	Negative	Negative
SPAPB1E7.09	SPBC119.09C	Negative	Negative
SPBC106.11C	SPCC4B3.06C	Negative	Negative
SPBC1198.07C	SPBC25H2.14	Negative	Positive
SPBC13A2.03	SPBC2A9.06C	Negative	Negative
SPBC13G1.01C	SPBC17A3.01C	Negative	Negative
SPBC13G1.01C	SPBC725.14	Negative	Negative
SPBC15C4.03	SPBC21B10.08C	Negative	Negative
SPBC15D4.15	SPCC14G10.05	Negative	Negative
SPBC1683.07	SPCC1450.12	Negative	Negative
SPBC16A3.06	SPCC4F11.02	Negative	Negative

SPBC16G5.06	SPBC342.03	Negative	Negative
SPBC16H5.03C	SPCC622.19	Negative	Negative
SPBC1709.10C	SPBC1921.07C	Negative	Negative
SPBC1709.15C	SPCC126.14	Negative	Negative
SPBC17A3.03C	SPBC4B4.12C	Negative	Negative
SPBC18H10.20C	SPBC19C7.08C	Negative	Negative
SPBC19C2.02	SPBC29A3.19	Negative	Negative
SPBC1A4.02C	SPBC3D6.15	Negative	Positive
SPBC20F10.06	SPCC970.02	Negative	Negative
SPBC21C3.16C	SPBC28F2.09	Negative	Negative
SPBC23E6.06C	SPBC577.08C	Negative	Negative
SPBC2G5.01	SPBC4F6.16C	Negative	Negative
SPBC30D10.18C	SPBC31F10.04C	Negative	Negative
SPBC342.03	SPCC24B10.13	Negative	Negative
SPBC3H7.05C	SPCC1259.10	Negative	Negative
SPBC4.02C	SPCC1795.08C	Negative	Negative
SPBC405.07	SPCC132.01C	Negative	Positive
SPBP19A11.05C	SPCP1E11.02	Negative	Negative
SPBP8B7.32C	SPCC285.04	Negative	Negative
SPBPJ758.01	SPCC297.04C	Negative	Negative
SPCC1682.01	SPCC550.06C	Negative	Negative