Classifications of protein roles in the functional organization of the cell

Yury Pritykin

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Advisor: Mona Singh

September 2014

© Copyright by Yury Pritykin, 2014.

All rights reserved.

Abstract

The availability of functional genomics data sets for numerous organisms provides an opportunity to comprehensively analyze the roles proteins play in the functional organization of the cell.

In the first part of this thesis, we study how simple network features of hub proteins (i.e., those with many physical interactions) are predictive of their roles in the functional organization of the cell. We begin by examining an influential but controversial characterization of the dynamic modularity of the *S. cerevisiae* interactome that incorporated gene expression data into network analysis. We analyze the protein-protein interaction networks of five organisms—*S. cerevisiae*, *H. sapiens*, *D. melanogaster*, *A. thaliana*, and *E. coli*—and confirm significant and consistent functional and structural differences between hub proteins that are co-expressed with their interacting partners and those that are not, and support the view that the former tend to be intramodular within networks whereas the latter tend to be intermodular. However, we also demonstrate that in each of these organisms, simple topological measures are significantly correlated with the average co-expression of a hub with its partners and therefore also reflect protein intra- and inter-modularity. Further, cross-interactomic analysis demonstrates that these simple topological characteristics of hub proteins tend to be conserved across organisms. Overall, we give evidence that purely topological features of static interaction networks reflect aspects of the dynamics and modularity of interactomes as well as previous measures incorporating expression data, and are a powerful means for understanding the dynamic roles of hubs in interactomes.

In the second part of this thesis, we study the role of multifunctional genes (and the proteins they encode) in the functional organization of the cell. Many genes can play a role in multiple biological processes or molecular functions. Identifying multifunctional genes at a genome-wide level and studying their properties can shed light

on the complexity of the molecular events that underpin cellular function, leading to a better understanding of the functional landscape of the cell. However, to date, genome-wide analysis of multifunctional genes has been limited. Here we introduce a computational approach that uses known functional annotations to extract genes playing a role in at least two distinct biological processes, and compare them with the remaining annotated genes. We leverage functional genomics data sets for three organisms—*H. sapiens*, *D. melanogaster*, and *S. cerevisiae*—and show that, as compared to other genes, genes involved in multiple biological processes possess distinct physicochemical properties, are more broadly expressed, tend to be intermodular in protein interaction networks, tend to be more evolutionarily conserved and are more likely to be essential. We also find that multifunctional genes are significantly more likely to be involved in human disorders. These same features also hold for genes with multiple molecular functions. Our analysis is a step towards a better genome-wide understanding of gene multifunctionality.

Overall, the results presented in this thesis lead to a better understanding of the complex functional roles that proteins play within the cell.

Acknowledgements

First and foremost, I can hardly express in words how extremely grateful I am to my advisor Mona Singh. From the time she accepted me as her student and through all the years of my graduate studies, she has been the greatest advisor one could dream about. She has been a great mentor and teacher, and I learned a lot from her on how to do science, all the way from starting to think about a scientific problem, to developing the right methods and approaches, to emphasizing what is the most important, to writing up the results clearly and concisely. She has been also a role model for me of how to deal with many things not directly related to research.

I would like to thank my thesis committee members, Bernard Chazelle and Olga Troyanskaya, as readers, and Andrea LaPaugh and Vivek Pai, as non-readers, for taking time to serve on my committee. I also would like to thank Tom Funkhouser and Stas Shvartsman for taking time to provide feedback on the preliminary version of my dissertation work.

I would like to thank all current and past members of the Singh Lab for their help and feedback over my graduate school years: Jesse Farnham, Dario Ghersi, Borislav Hristov, Peng Jiang, Zia Khan, Daniel Munro, Shilpa Nadimpalli, Alex Ochoa, Anton Persikov, Pawel Przytycki, Jimin Song, Josh Wetzel, Tao Yue, and Jose Zamalloa. Especially I thank Anton and his family who became really good friends.

The results of Chapter 2 have been presented in April 2012 at the 16th International Conference on Research in Computational Molecular Biology (RECOMB 2012) and in March 2013 at the Cold Spring Harbor Laboratory Meeting "Systems Biology: Networks", and subsequently published [1]. The results of Chapter 3 were obtained in collaboration with Dario Ghersi.

My graduate studies were supported by a fellowship from Princeton University and grants NSF ABI-0850063, NIH GM076275, the NIH Center of Excellence grant P50 GM071508, and NSF grant CCF-0963825.

I thank all the teachers and professors from Moscow School N 57, Moscow State University, Princeton University who taught me mathematics, computer science, biology and a lot of other subjects.

I am really thankful to all the professors and instructors in Princeton University from whom I learned as a student or as a supporting teaching assistant, I remember something and in most cases a lot from each course: Sanjeev Arora, Boaz Barak, David Botstein, Léon Bottou, Moses Charikar, Bernard Chazelle, Robert Dondero, Michael Freedman, Tom Funkhouser, Alison Gammie, Coleen Murphy, Vivek Pai, Stas Shvartsman, Mona Singh, Bob Tarjan, Saeed Tavazoie, Olga Troyanskaya, Kevin Wayne, Ned Wingreen. I thank the Theory Group of the Computer Science Department where I started as a graduate student, and especially Bernard Chazelle who advised me on the early stages. The Lewis–Sigler Institute for Integrative Genomics provided a great environment for me to develop as a computational biology researcher. I am thankful to Alison Gammie for allowing me to participate in her intensive course on basic molecular biology experimental techniques, which seriously broadened my perspective of biology.

I thank the Department of Mechanics and Mathematics in Moscow State University, and in particular all the members of the Division of Mathematical Logic and Theory of Algorithms where I spent several very important and enjoyable years, especially Alexei L'vovich Semenov and Andrej Al'bertovich Muchnik (1958—2007) who advised me on my mathematics research and a lot more.

I thank all the teachers of Moscow School N 57 where it all started, especially Ivan Valerievich Yaschenko and Rafail Kalmanovich Gordin.

For being with me and enriching my life, I thank Masha, and I thank all my friends.

Most importantly, I am extremely grateful to my family for their constant love and support in everything I do.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A living cell is a very complex system, where the main players are large molecules such as DNA, RNA, and proteins. Numerous interactions of various types amongst these molecules and others underlie virtually all biochemical processes within a cell. Therefore, understanding the network composed of all of these interactions is necessary for understanding how the cell functions. The association of many of these molecules with multiple biochemical processes and biological functions further increases the complexity of the system. With the recent widespread availability of diverse functional genomics data sets for multiple organisms, it has become possible to perform systematic integrated analysis in order to better understand the functional landscape of the cell.

1.1 Proteome

Proteins are the main building blocks and functional units of the cell. The primary structure of proteins is composed of a linear sequence of small molecules called amino acids. There are twenty commonly occurring amino acids in proteins, and a protein sequence ranges in length from tens to thousands of amino acids. Combinatorially, an enormous number of amino acid sequences is possible, and typically thousands of

them are actually encoded and subsequently *expressed* in living cells. *A proteome* is the collection of all proteins encoded within a cell. The information about the proteome is stored, or encoded, in DNA, another type of macromolecule. A similar molecule, RNA, is an intermediate in the process of protein expression. Sub-regions of DNA sequence encoding proteins are called *genes*, and *the genome* is the entire DNA making up the cell. According to the central dogma of molecular biology, proteins are expressed in two stages: first, messenger RNAs (mRNAs) are produced from genes in a process called *transcription*, and then proteins are produced from mRNAs in a process called *translation*. Different sets of proteins are expressed in different cells and conditions, and this expression is regulated by proteins known as transcription factors.

The three-dimensional folding of protein sequences results in an immense versatility of structures, allowing proteins to have a variety of different functions. Proteins can be enzymes catalyzing many kinds of chemical reactions in the cell. Proteins can provide structural support or have a mechanical function in the cell. Proteins are important for cell signaling and for recognizing targets by the immune system. Many proteins are involved in more than one function and can perform their functions by interacting with each other transiently or by forming stable complexes. A protein's abundance and interactions, and consequently its function may change over time and depends upon cell type or cell developmental stage, as well as upon other conditions. This variation contributes to the complexity of proteome functioning.

For a better understanding of the functional organization of the cell as a system, one needs to study what proteins do, how they do it, and how protein functionality changes over time as a response to changing conditions. In order to gain this understanding on a proteome-wide scale, a number of high-throughput experimental techniques have been developed. For example, microarrays or RNA-seq can be used to obtain a profile of steady state mRNA abundances for all genes within the cell, as

a proxy to protein expression levels [2, 3]. Yeast two-hybrid and affinity purification followed by mass spectrometry are two major methods used to detect large numbers of physical protein-protein interactions [4]. Databases have been created for standardized aggregation and convenient use of different types of high-throughput data [5, 6]. Large consortia such as the Gene Ontology [7] develop vocabulary for the functional annotation of genes and proteins, and collect information for such annotations from a variety of data sources for a number of organisms.

Many computational methods and techniques have been developed to gain new knowledge about the biological functioning of the cell using integrative analysis of large proteome-wide or genome-wide data sets [8]. This thesis consists of two such new analyses. In the first part of this thesis, we study how simple topological properties of proteins within interaction networks reflect aspects of the dynamics and functional organization of the proteome. In the second part of this thesis, we perform a systematic analysis of the phenomenon of multifunctionality of genes and proteins. We next briefly give some background on each of these themes.

1.2 The topology and dynamics of protein interaction networks

For several model organisms and for human, there have been significant efforts in the past 15 years to build large-scale interaction datasets of various types, also sometimes called *interactomes*. A graph is a convenient and productive abstraction to represent such interaction data. For example, physical interactions between proteins (i.e., corresponding to binding events) are usually represented by an undirected graph where proteins are vertices, and edges represent physical interactions between proteins. In transcriptional regulatory networks, interactions are between genes encoding transcription factors and the genes that they regulate; these networks are represented by

directed graphs. Often for simplicity, especially when talking about interactions of different types at the same time, it is convenient to refer to genes and the proteins they encode interchangeably.

The main focus of our work is on protein-protein physical interactions. The development of experimental methods to detect thousands of protein interactions in one experiment, as well as the creation of public databases to curate these data, have accelerated research on protein interactomes [4, 9, 5]. Computational analyses have revealed a number of fundamental properties of physical protein interaction networks [10, 11]. It has been shown that a small number of proteins, often called *hubs*, have a tendency to interact with very many other proteins, more than expected by chance if interactions are distributed uniformly at random among proteins (and there has been substantial debate among researchers about models to explain these observations [12, 13, 14]). Hubs have been shown to have specific interesting properties, such as a higher rate of essentiality, more evolutionary conservation, and a higher chance of association with disease [15, 16, 17, 18]. Protein networks have been shown to have a modular structure; i.e., these networks consist of groups of tightly interconnected nodes corresponding to protein complexes and functional modules [19, 20, 21, 22]. A number of methods have been developed to integrate protein interaction data with other data sources to reveal aspects of the dynamics of protein interactomes [23]. By integrating protein interaction data with microarray expression data across a number of conditions, the concept of dynamic modularity was proposed [24]. In particular, it was argued that hubs can be categorized in one of the two groups: high-level global connectors of the interactome or low-level hubs functioning inside modules. In Chapter 2, we significantly expand our understanding of the roles proteins play in organizing the dynamics and modularity of the cell. In particular, we show that simple topological features of proteins within interaction networks reflect aspects of the

dynamics and modularity of interactomes as well as previous measures incorporating gene expression data.

1.3 Gene multifunctionality

The central dogma of molecular biology states that a gene produces a protein, and then the protein is responsible for some cellular function. For a long, time the idea of "one protein—one function" was the dominant viewpoint. However, it is getting more and more clear that life is not as simple, and that there are genes and proteins that are responsible for more than one function.

One of the first discovered examples of multifunctional genes were crystallins, structural proteins of the eye lens found to be also present in other tissues where they have an enzymatic role [25]. Since then, numerous other examples have emerged, including ribosomal proteins that function in DNA repair or as developmental regulators, thrombins that are also ligands for cell surface receptors, and many others [26]. The phenomenon of gene multifunctionality is sometimes called protein moonlighting or gene sharing [27, 26]. Gene pleiotropy is a related but different concept, where a single gene is associated with multiple phenotypes [28].

The multifunctionality of a protein can be explained by different mechanisms related to its dynamics and structural plasticity. A protein may be expressed in different parts of the cell, or inside or outside of the cell, or in different types of cells in a multicellular organism. Binding of a cofactor, as well as oligomerization, can also affect functionality. Some proteins may have several different binding sites or intrinsically unstructured regions corresponding to different functions [26, 29, 30].

There is increasing evidence that the phenomenon of multifunctionality is actually very common. We thus set out to analyze multifunctional genes at a systems level using computational approaches. However, there is a lack of detailed large-scale

experimental data about multifunctionality. Instead, using the Gene Ontology as our primary source of information about gene function, we develop an approach to study gene multifunctionality at a genome-wide scale. In Chapter 3, we propose a systematic method for detecting multifunctional genes and then analyze gene multifunctionality with respect to a number of available genome-wide data sets.

1.4 Our contributions

In this thesis, we perform computational analysis to advance our understanding of the roles proteins play in the functional organization of the cell.

In Chapter 2, we study how simple properties of hubs are predictive of the roles they play in the functional organization of cellular networks. We leverage protein-protein interaction data and microarray expression data, along with other functional genomic data, for five organisms—*S. cerevisiae*, *H. sapiens*, *D. melanogaster*, *A. thaliana*, and *E. coli*. We show that certain simple features of hubs in the network reveal important aspects of the dynamics and modularity of the interactome. One of these features is the average co-expression of a protein with its interacting partners [24]. However, surprisingly, the other features depend purely on the topology of the network. We find that these simple properties reflect intra- and inter-modularity of proteins in the network. We also perform a cross-interactomic analysis and observe that inter- and intra-modularity, as measured by these simple hub features, is conserved across organisms.

In Chapter 3, we study the role of multifunctional genes and the proteins that they encode in the functional organization of the cell. We use functional annotation information from Gene Ontology for three organisms, *S. cerevisiae*, *H. sapiens*, and *D. melanogaster*. We propose a robust method to detect genes that have two or more very distinct functions, and distinguish them from genes more likely to have

a single function. Using a number of genome-wide data sources, including protein-protein physical interaction data, we perform an analysis of multifunctional genes with respect to various biological properties, and show that they are significantly different from other non-multifunctional genes. We show that, as compared to other genes, multifunctional genes possess distinct physicochemical properties, are more broadly expressed, are intermodular in protein interaction networks, tend to be more evolutionarily conserved and are more likely to be essential. We also find that multifunctional genes are significantly more likely to be involved in human disorders. These same features also hold for genes with multiple molecular functions.

Chapter 2

Simple topological features reflect dynamics and modularity in protein interaction networks

2.1 Introduction

A better understanding of protein interaction networks would be a great aid in furthering our knowledge of the molecular biology of the cell. Towards this end, large-scale protein-protein interaction (PPI) networks have been determined for a diverse set of model organisms and for human [31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42]. Computational analyses of these networks have revealed many important aspects of cellular organization and functioning [10, 11], including a strong link between the topological characteristics of cellular networks and their underlying functioning. One fundamental finding is that PPI networks are modular: they tend to consist of groups of tightly interacting proteins corresponding to functional modules or protein complexes [19, 43, 44, 45, 46, 47]. Further, from early on, it has been apparent that hubs—proteins participating in many interactions—have special roles in PPI networks: they

tend to be more essential, more evolutionarily conserved, and in human are enriched in genes over-expressed or mutated in cancers [12, 17, 48, 16, 15, 49, 50, 18]. It is naturally interesting to combine these two well-studied views of PPI networks and to ask how hubs are positioned with respect to the modular organization of the cell.

Specific contextual information about interactions would be of great help in understanding the connection between hubs and modularity. For most interactions of a protein in a network, however, we typically do not know whether these interactions occur at the same time or under different conditions. In order to understand the dynamic roles of hubs and their relationship to network modularity, a highly influential study integrated PPI data with gene expression data measured in numerous conditions, and classified hubs based on their average co-expression with their interacting partners [24]. Hubs that have high average co-expression with their partners were classified as "party," as they are likely to interact with these other proteins at the same time. Conversely, hubs that have low average co-expression with their partners were classified as "date," as they are likely to interact with their partners at different moments of time. In an analysis of the *S. cerevisiae* interactome, date and party hubs were shown to exhibit different biological properties that imply different roles in the PPI network. In particular, date hubs were found to have more diversity in their subcellular localizations, a more drastic effect on network connectivity when removed from the network, higher centralities in a network of genetic interactions, and higher evolutionary rates [24, 51, 52]. Further, it was argued that date hubs are global connectors of different modules whereas party hubs are more local and play specific roles in modules.

Though the classification of hubs into party and date has been generally accepted [51, 16, 53, 36, 54, 4], it has also generated significant controversy [55, 56, 57]. It has been proposed that the observed date/party hub distinction may be an artefact of biases in the datasets used or of the analysis methodology. Whereas the initial

work [24] observed a bimodality in the distribution of average co-expression across hubs and used this to partition hubs into party or date, this bimodality has not been observed in subsequent studies [52, 55, 56, 57]. Further, it has been suggested that the difference in the effect on network connectivity of removing either all the date or party hubs is attributable to a difference in the total number of interactions of date and party hubs, that date and party hubs evolve at the same rate, and that there is no evidence of different centrality in the genetic network for date and party hubs [55, 56]. Finally, it has been argued that the observed differences in the topological properties between date and party hubs in the network may be attributable only to a small number of date hubs with extreme properties, while the remaining hubs are much more homogeneous [57].

The current availability of large-scale interaction networks for numerous organisms across the evolutionary spectrum provides us with an opportunity to systematically analyze whether properties of hubs are predictive of the roles they play in the functional organization of cellular networks. We use interaction networks for five organisms, *S. cerevisiae*, *H. sapiens*, *D. melanogaster*, *A. thaliana*, and *E. coli*, along with multiple mRNA expression datasets. In each of these organisms, we show that the average co-expression of a hub with its partners, independent of any categorization of hubs, reveals important properties of hubs: the average co-expression of a hub with its interacting partners is significantly positively correlated with its local clustering coefficient as well as its average biological process similarity with its interacting partners, and is significantly negatively correlated with its betweenness centrality and its participation coefficient (a topological measure that reflects the diversity of the intermodular interactions of a protein). Further, the average co-expression of a hub with its interacting partners is negatively correlated with its interaction degree in genetic networks, and positively correlated with protein essentiality. Importantly, the correlations uncovered between average co-expression and topological features—independent

of any classification of hubs—imply that the topological features of hubs by themselves reflect important aspects of the dynamics and modularity of the interactome. For example, hubs with low betweenness or high clustering coefficient will tend to have high average co-expression with their partners and fewer genetic interactions, whereas proteins with high betweenness or low clustering coefficient tend to exhibit the opposite trends. Further, as part of our study, we revisit the date-party controversy. We consider a very simple criterion to classify hubs as either party or date, and confirm significant and consistent functional and topological differences in the properties of date and party hubs across organisms. Finally, in a cross-interactomic analysis, we demonstrate that these simple topological and co-expression properties of hub proteins tend to be conserved across organisms, giving further evidence that these features reflect important aspects of cellular functioning.

2.2 Results

2.2.1 Preliminaries

We begin by briefly describing our data and analysis framework (see Section 2.4 for details). We analyze PPI networks for *H. sapiens*, *S. cerevisiae*, *D. melanogaster*, *A. thaliana*, and *E. coli* (denoted by **Human-all**, **Yeast-all**, **Fly**, **Athal**, and **Ecoli**, respectively). For human and yeast, we additionally consider networks consisting of high-confidence interactions only (denoted by **Human-hq** and **Yeast-hq**). We gather mRNA expression data for these organisms from GEO [58], and compute a co-expression score for each interaction using the Pearson correlation coefficient (PCC). We define hubs as all genes in the top 10% in each interactome by the number of interactions. For each hub, we calculate the average of the co-expression scores (avPCC) computed over all the interactions in which this hub participates. The size of each network, the number of interactions with expression data, and the number

Network	Num. genes	Num. interactions	Num. interactions with co-expression score	Hub threshold	Num. hubs (%)	Num. hubs with avPCC
Human-hq	4,750	13,102	11,781	12	481 (10.1%)	466
Yeast-hq	4,467	22,243	21,869	23	449 (10.1%)	445
Fly	8,218	36,569	36,525	23	865 (10.5%)	865
Athal	5,454	12,883	10,611	10	555 (10.2%)	506
Ecoli	3,115	17,788	17,697	23	319 (10.2%)	319
Human-all	10,229	80,651	66,102	39	1,033 (10.1%)	931
Yeast-all	5,641	59,930	59,658	49	570 (10.1%)	570

Table 2.1: Network sizes and the number of network hubs.
The number of vertices (genes) and edges (interactions) in each network, the number of interactions that were assigned a co-expression score, the degree threshold to be chosen as a hub, the number of hubs obtained using this threshold, and the number of hubs that were assigned an average co-expression score.

of hubs are listed in Table 2.1. In the main text, we focus on the human high confidence interaction network **Human-hq** unless otherwise specified, but results for all networks are given in Appendix A.

We use four measures to ascertain the functional, organizational and dynamic properties of proteins in the network: clustering coefficient, betweenness centrality, participation coefficient and functional similarity. Clustering coefficient is the density of the neighborhood of a protein in the network, and proteins with higher clustering coefficient have interactions with proteins that interact with each other. Betweenness centrality is a measure of the fraction of shortest paths passing through a node in the network, and nodes with higher betweenness are more globally central in the network. Participation coefficient shows how well interactions of a protein are distributed amongst clusters in the network, so that proteins with low participation are mostly interacting with proteins from the same cluster, whereas proteins with high participation have their interactions spread among many clusters. Functional similarity estimates to what extent a protein participates in the same biological process as its

neighbors in the network. We note that three of these measures are purely topological and do not use any information other than interaction data, whereas functional similarity also uses Gene Ontology [7] annotations.

We classify hubs in the low range of avPCC as date and hubs in the high range of avPCC as party. Despite the previously observed differences between date and party hubs [24, 51, 16, 53, 36, 54, 4], the choice of a threshold in the avPCC range between the two classes of hubs has remained a topic of disagreement. As we have many networks to consider, we choose a simple threshold criterion and later demonstrate that this choice does not matter (see Section 2.2.3). In particular, we define party hubs as the one third of hubs with the largest avPCC, and call the remaining two thirds of hubs date. Since it has been argued that the originally observed differences between date and party hubs may be attributed only to a small number of date hubs with extreme network global centrality properties [57], we classify hubs with many interactions or with high betweenness centrality into a special group called extremal hubs (18 to 89 hubs depending on the network) and exclude them from the analysis of date and party hubs.

2.2.2 Properties of date and party hubs are significantly distinct

We first analyze the differences between party and date hubs on our seven networks (Fig. 2.1 and Fig. A.1, Fig. A.2, Fig. A.3, Fig. A.4, Fig. A.5 and Fig. A.6). We confirm that date hubs tend to be more globally central in the network and to have more diverse intermodular participation, as reflected by their significantly higher betweenness centrality and participation coefficient ($p < 1e{-}23$ and $p < 9e{-}27$ respectively, in the high confidence human network, Mann–Whitney U; Fig. 2.1B). Further, party hubs tend to have denser neighborhoods consisting of genes with more similar functions, as reflected by their significantly higher clustering coefficient and functional

similarity ($p < 4e{-}30$ and $p < 4e{-}28$, respectively, in the high confidence human network, Mann–Whitney U; Fig. 2.1B).

In addition to comparing the node-level features of date and party hubs, we also compare the positioning of the set of date and party hubs in the network with respect to each other in order to better understand global network organizational features. For either the set of date hubs or the set of party hubs, we measure how well connected they are to each other by calculating the density of the subnetwork induced by them, defined as the number of interactions amongst the set of proteins, normalized by the maximum possible number of such interactions. We also measure how well spread the interactions of these hubs are in the whole network by calculating the expansion of the set, defined as the number of proteins in the network that are connected with any hub in the set, but do not belong to the set, normalized by the size of the set. We observe that party hubs have a strong tendency to interact with other party hubs, and much less so with other proteins, as reflected by their high density and low expansion in the network as compared with sets of the same size consisting of randomly selected hubs (Fig. 2.1C). On the contrary, date hubs have significantly lower density and significantly higher expansion than random sets with the same number of hubs, suggesting that they are more sparsely distributed in the network than party hubs.

As a final test to compare the topological features of date and party hubs, we compare the effect of node removal on network structure for date and party hubs. For a set of hubs (either date or party), we remove all of them from the network at once and measure the change in three representative global network characteristics: average path length, size of the largest connected component, and global clustering coefficient. We compare the effect of such a removal with the effect of a removal of random sets of the same number of hubs. We observe that date hubs are more central in the network, as their removal affects connectivity of the network much more

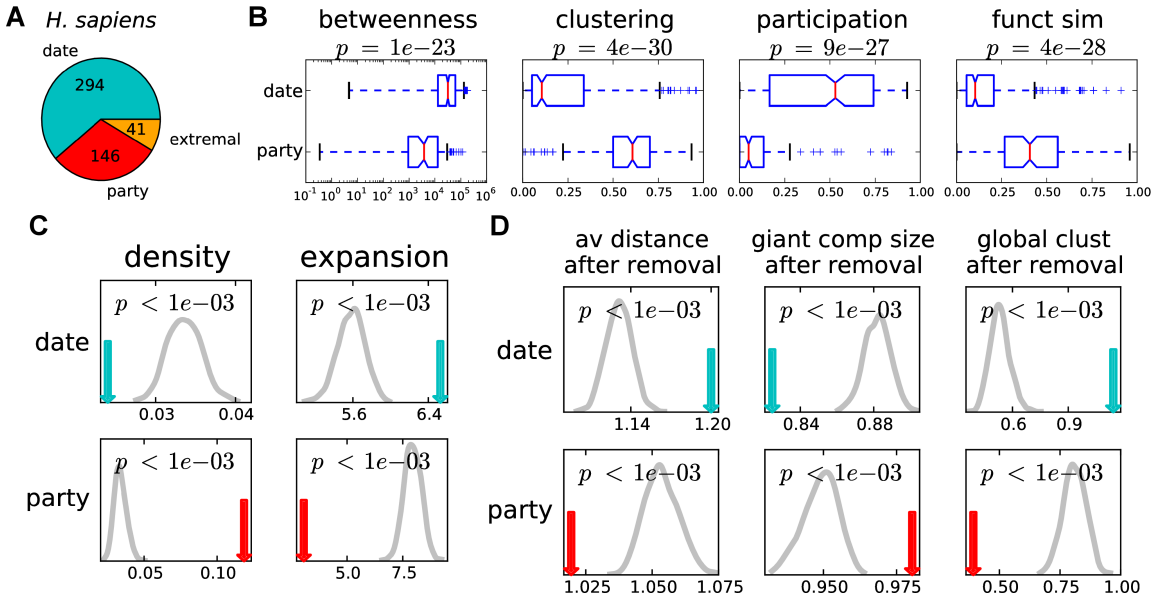

Figure 2.1: Date and party hubs have distinct functional and topological properties. Date and party hub classification analysis in the human high quality interaction network (**Human-hq**). (A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.29; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs are significantly different; p-values are computed using the Mann–Whitney U. (C) Density and expansion of date and party hubs are significantly different. The gray curves in each panel show the distributions for 1000 independent random samples of the same number of hubs, and are used to compute empirical p-values. (D) Effect of hub removal is significantly different for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. The gray curves show the distributions for 1000 independent random samples of the same number of hubs, and are used to compute empirical p-values. See Section 2.4 for more details.

significantly than removal of random hubs, as reflected by the average path length of the network and the size of the largest connected component (Fig. 2.1D). At the same time, removal of party hubs makes the network much less clustered than the removal of random hubs, as reflected by the effect on the global clustering coefficient.

The results of these analyses have qualitatively the same trends across the five organisms (Fig. A.1, Fig. A.2, Fig. A.3, Fig. A.4, Fig. A.5 and Fig. A.6), if extremal hubs are included in the analysis (Fig. A.7), or if all proteins with at least three interactions in the network are considered hubs (see Section A.1.2 and Fig. A.8, Fig. A.9 and Fig. A.10). Taken together, our analysis over seven networks suggests significant and consistent differences between proteins characterized based on avPCC with respect to topological, intermodular and functional features.

2.2.3 Hub characteristics capture functional and organizational properties of the interactome

We next show that the avPCC measure is an interesting biological measure independent of any threshold one could use to define date and party hubs. That is, while there has been significant previous controversy concerning how an avPCC threshold should be chosen to categorize hubs into date and party [55, 56], we show here that the avPCC measure is itself correlated with other characteristics of hubs in the network. In particular, we compute the Spearman rank correlation (SRCC) between avPCC and our topological and functional measures (Fig. 2.2, top row, and Table A.1). Across the organisms, we find consistent positive correlations of avPCC with clustering (SRCCs ranging from 0.30 to 0.72 depending upon the network) and functional similarity (SRCCs from 0.25 to 0.59) and negative correlations with betweenness (SRCCs from -0.20 to -0.55, except **Ecoli**) and participation (SRCCs from -0.26 to -0.69). These correlations are consistent with the original claims [24] that hubs in the high avPCC range are more local and play more central roles within modules and

complexes, and thus have higher clustering coefficients and higher average functional similarities with their interacting partners, whereas hubs in the low avPCC range play more global roles in organizing other proteins' functioning and thus are more globally central in the PPI network (as evidenced by higher betweenness centrality) and have more diverse participation in interactions with different processes and modules (as evidenced by higher participation coefficient).

Given the significant and consistent correlations between avPCC and clustering, betweenness, participation and functional similarity, we also compute the SRCCs amongst these measures. As expected from our above analysis, these measures are also correlated with each other in a consistent manner across the seven networks (Fig. 2.2 and Tables A.2, A.3 and A.4). Comparing the three purely topological measures with each other, we find that betweenness is positively correlated with participation, while both are negatively correlated with clustering. We further note that because the functional similarity measure aggregates information from Gene Ontology, we can also use it as an independent means of assessing whether the topological measures based purely on interaction data reflect properties of protein functioning. We find that functional similarity is significantly positively correlated with clustering (SRCC from 0.26 to 0.71, except **Ecoli**), while negatively correlated with betweenness (SRCC from -0.09 to -0.62, except **Ecoli**) and participation (SRCC from -0.21 to -0.65, except **Ecoli**). This suggests that measures based purely on the topology of the network can reflect interesting functional properties of proteins.

As a control to confirm that the information coming from protein-protein interactions is crucial, we randomize each network in a degree-preserving manner [59], and recompute the node-level topological and functional measures using the randomized interactions. Correlations between these measures aggregated over 20 random networks (Fig. 2.2) have substantially lower absolute values than for real interaction networks and sometimes show a completely opposite trend. We note, however, that

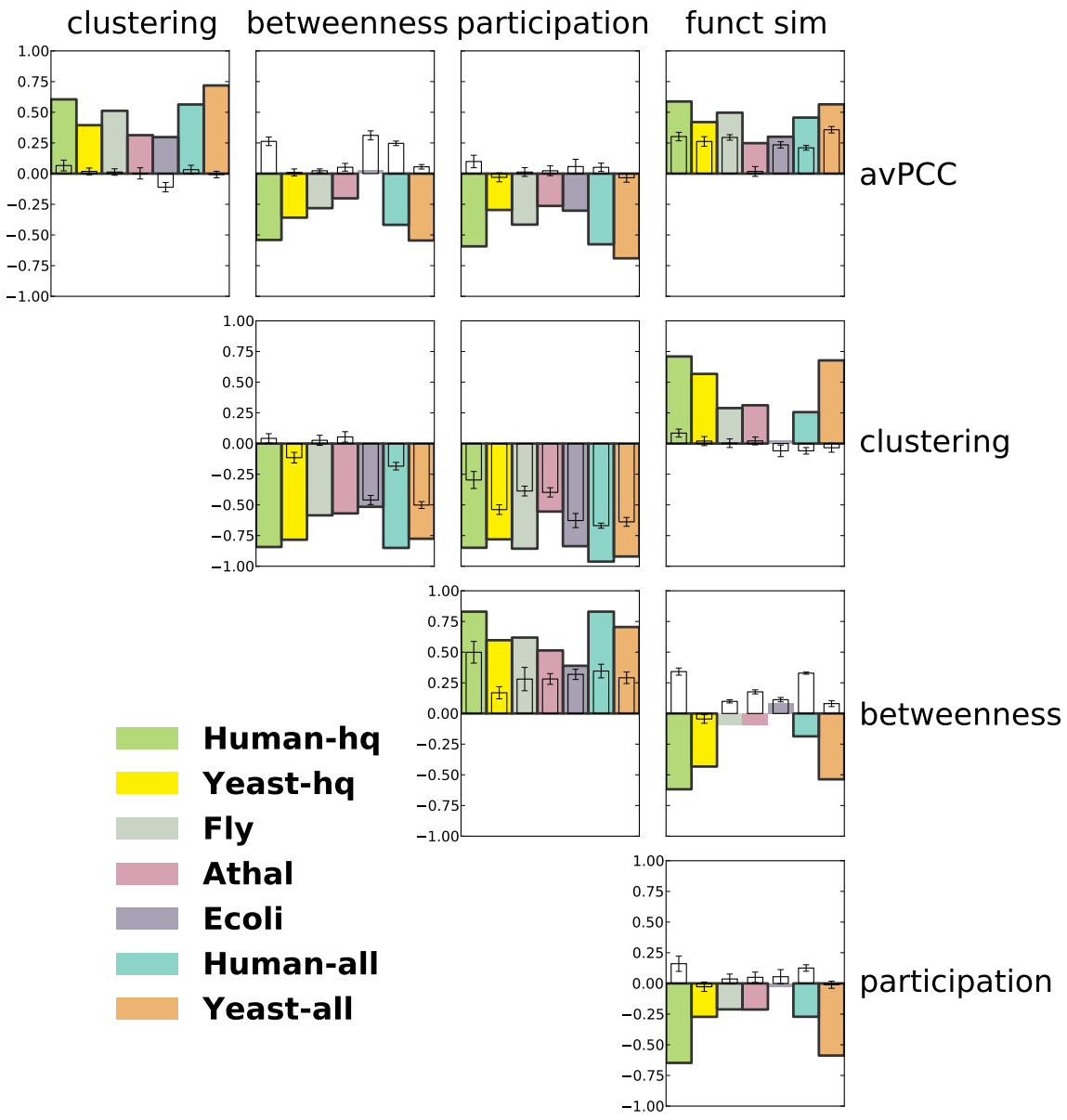

Figure 2.2: Functional and topological characteristics of hubs are significantly correlated with each other in a consistent manner in protein-protein interaction networks. Every colored bar represents a Spearman correlation between two characteristics of hubs in one of the networks. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges. See Tables A.1, A.2, A.3 and A.4 for exact values. Smaller uncolored bars show average correlations in 20 degree-preserving random networks (with error bars depicting standard deviations).

these measures can still have significant and meaningful correlations in random networks. For example, a remarkably high correlation is found between avPCC and functional similarity in random networks, though it is still noticeably lower than in real networks. This is an indication of the strong signal in expression data itself that does not arise from physical interactions. Indeed, it is not surprising that even for arbitrary pairs of genes, not necessarily physically interacting, the more often they are expressed together, the more likely that they are functionally related.

Potentially confounding factors in our correlation analysis include the protein degree threshold used to identify hubs in the networks, correlations of hub features with degree, bias from extremal hubs, and study bias. To demonstrate that none of these significantly affect our results, we also perform this analysis when using different degree thresholds, when computing partial correlations with correction for degree, when excluding extremal hubs, and when focusing on high-throughput networks in yeast and human (see Sections A.1.2 and A.1.3, and Fig. A.11, Fig. A.12, Fig. A.13, Fig. A.14 and Fig. A.15).

2.2.4 Distinct functions are enriched in hub classes

To determine whether party and date hubs (as well as hubs partitioned into groups based on topological measures) tend to participate in different biological functions, we performed GO enrichment analysis on each set of hubs in **Human-hq** using the most general terms in each of the three ontologies (i.e., the terms that are immediate children of roots of the ontologies), and used all annotated hubs to provide the background functional distribution. We found that terms enriched for date and party hubs are very different: date hubs are associated with global tasks such as "biological regulation" and "signaling," while party hubs are enriched in local and module- and complex-specific terms such as "macromolecular complex" and "metabolic process" (Fig. 2.3).

Interestingly, very similar terms are enriched when, instead of a date-party classification based on avPCC, hubs are classified in a 2-to-1 proportion using clustering coefficient, betweenness centrality, participation coefficient or functional similarity (Fig. 2.3). That is, the same functional terms have similar enrichments when hubs are classified based purely on topological measures, suggesting that these topological properties can reflect the functional roles of hubs in the interactome as well as avPCC does. In general, we obtain similar results for the other networks (see Section A.1.4 and Fig. A.16, Fig. A.17, Fig. A.18, Fig. A.19, Fig. A.20 and Fig. A.21).

2.2.5 Hubs that are more globally central in physical interaction networks have more genetic interactions

We next consider the relationship between various properties of hubs in physical interaction networks and their number of genetic interactions. In the initial publication on date and party hubs [24], it was observed that date hubs are involved in more genetic interactions than party hubs, and it was proposed that their phenotypic link to many proteins was due to their connecting different biological processes to each other [24]. As yeast remains the only organism with a sufficiently large number of known genetic interactions, our analysis on genetic interactions is limited to this organism. We note, however, that the current dataset aggregated in BioGRID [60] is two orders of magnitude larger than the one used previously by Han *et al.* [24].

We compute SRCCs between the number of genetic interactions a hub has and its avPCC, its clustering coefficient, its betweenness centrality, its participation coefficient and its functional similarity in the physical interaction network. For both the **Yeast-all** and **Yeast-hq** networks, avPCC is significantly negatively correlated with genetic interaction degree (Fig. 2.4). Further, we find that the number of genetic interactions of a hub is positively correlated with betweenness and participation in both networks, while negatively correlated with clustering and functional similarity.

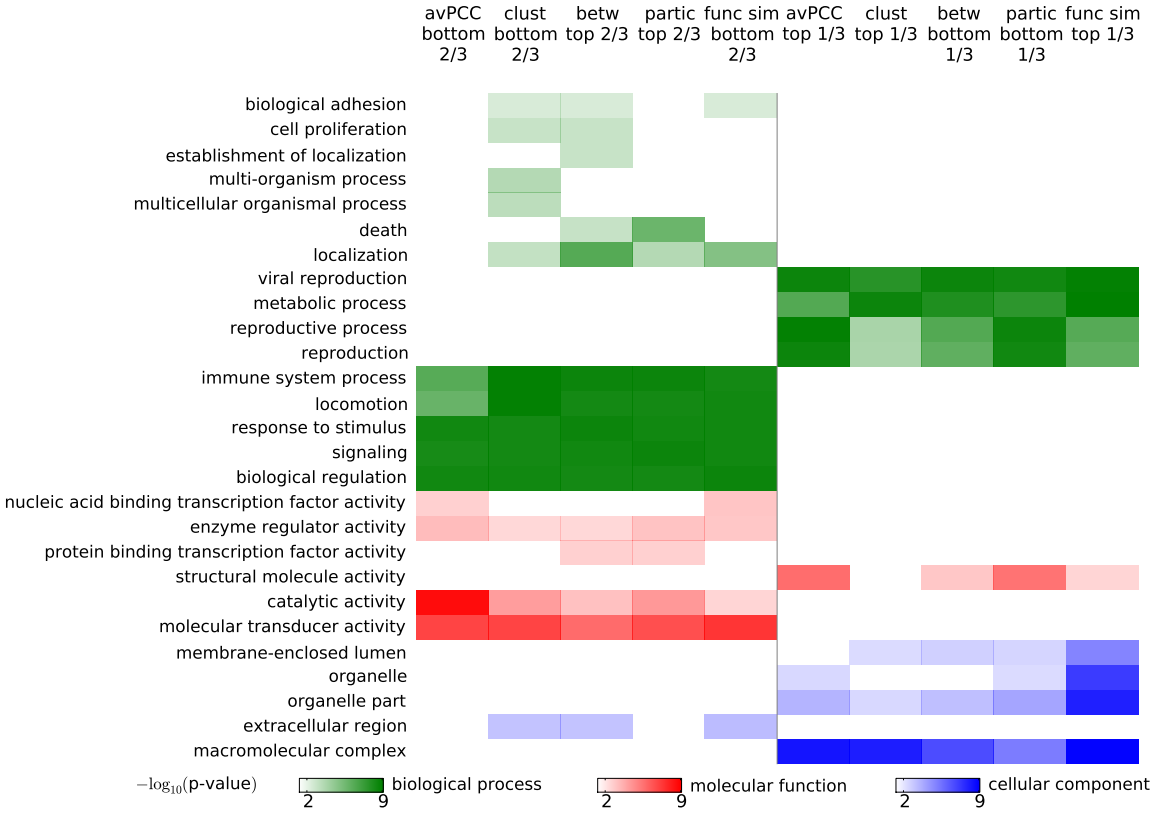

Figure 2.3: Different hub characteristics produce classifications of hubs with similar functional properties.
Hubs are divided in a 2-to-1 proportion using either avPCC, clustering coefficient, betweenness centrality, participation coefficient or functional similarity scores in **Human-hq**. Broad GO terms that are enriched at Bonferroni-corrected significance threshold 0.05 are shown with colors indicating the ontology of the term and color intensity indicating the p-value of enrichment. Classifications of hubs based on different hub characteristics produce classes of hubs with similar functional annotations.

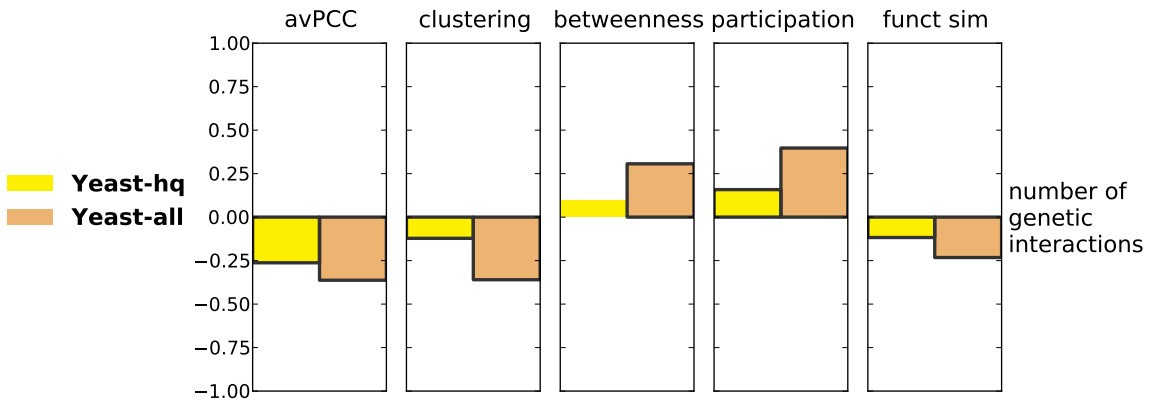

Figure 2.4: Characteristics of hubs in protein physical interaction networks are significantly correlated with their number of genetic interactions.
Every bar represents a Spearman correlation between a hub characteristic in one of the protein physical interaction networks for yeast and degree in the yeast genetic interaction network. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges.

We also confirm that these correlations are significant even when compared with those in random networks (see Section A.1.8 and Fig. A.34). Finally, to directly compare with the original study [24], we also verify that date hubs are involved in many more genetic interactions than party hubs (Fig. A.22AB).

These observations are also largely confirmed when negative and positive genetic interactions are considered separately (Fig. A.23), as well as when computing partial correlations corrected for essentiality (see Section A.1.5, Fig. A.22CD and Fig. A.24). As is the case when considering all genetic interactions together, the trends are stronger in **Yeast-all** than in **Yeast-hq**.

Overall, we find that not only avPCC, but also other hub characteristics, including those that are purely topological, are significantly correlated with centrality in the genetic interaction network. These results support the original observations that hubs with the role of global connectors and organizers of the interactome, as identified by avPCC or (as we show here) by other topological measures, are related in their effect on phenotype with many more genes than are local hubs from modules and complexes.

2.2.6 Role of yeast two-hybrid and co-complex interactions

Physical protein-protein interactions obtained using different methods can differ in their characteristics [36, 61]. In particular, the two high-throughput methods that account for the largest number of interactions in our networks, yeast two-hybrid (Y2H) and affinity purification followed by mass spectrometry, tend to detect different types of interactions. The former are more likely to detect direct, transient binary interactions between proteins whereas the latter tend to detect more stable co-complex interactions that may or may not correspond to direct interactions.

It was previously observed that for a fixed avPCC threshold in the definition of date and party hubs, date hubs are much more prevalent in Y2H networks, while party hubs are more prevalent in co-complex networks [36]. Therefore it was suspected that the observed distinction between date and party hubs may be attributable to the fact that interaction networks are typically compiled of interactions of both types, and this may artificially imply the date/party distinction [57]. In order to rule out these concerns regarding our observations about topological features of hub proteins, we apply the same analysis to networks of only Y2H or of only co-complex interactions.

We find that correlations between different hub characteristics for networks formed by either only yeast two-hybrid or only co-complex interactions are qualitatively the same as in networks with interactions of all types combined (Fig. A.25, as compared with Fig. 2.2). The date/party distinction for hubs in these networks separately is also qualitatively the same as in networks with interactions of both types combined (Fig. A.26, Fig. A.27, Fig. A.28, Fig. A.29, Fig. A.30 and Fig. A.31, compare with Fig. 2.1 and Fig. A.1, Fig. A.2, Fig. A.3, Fig. A.4, Fig. A.5 and Fig. A.6). Thus, simple hub characteristics consistently reflect principles of network structure and functioning even when applied to networks comprised of either Y2H or co-complex interactions.

Despite the consistency in correlations amongst hub features between Y2H or co-complex networks with the network compiled of interactions of both types, when

analyzing just the latter combined network, topological properties of hubs are consistently and oppositely correlated with the number of Y2H interactions as opposed to the number of co-complex interactions. The avPCC measure is negatively correlated with the number of Y2H interactions, while positively correlated with the number of co-complex interactions (Fig. 2.5). Accordingly, date hubs participate in more Y2H interactions, while party hubs participate in more co-complex interactions (p-value from $5e{-}08$ to $3e{-}24$, Mann–Whitney U; Fig. A.32). Betweenness and participation are positively correlated with the number of Y2H interactions, while negatively correlated with the number of co-complex interactions, which suggests that these two measurements are indeed capturing the centrality of hubs and their tendency to interact one-to-one with other proteins. Clustering and functional similarity are negatively correlated with the number of Y2H interactions, while positively correlated with the number of co-complex interactions, which suggests that these two measurements are capturing the tendency of hubs to participate in complexes and functionally homogeneous modules. Thus, we find that more globally central hubs (as specified by either betweenness or participation) tend to have more yeast two-hybrid interactions whereas more module-specific hubs (as specified by either avPCC, clustering coefficient or functional similarity) tend to have more co-complex interactions.

2.2.7 Hubs involved in modules and clusters are more likely to be essential

In the initial study [24], it was observed that in yeast, party hubs are more likely to be essential than date hubs (though the observed difference was not significant). We revisit this question with our newer and larger data set.

We compute the SRCC between essentiality represented as an indicator vector (i.e., 1 if a gene is essential and 0 otherwise) and other characteristics of hubs in the physical interaction network. For both the **Yeast-all** and **Yeast-hq** networks,

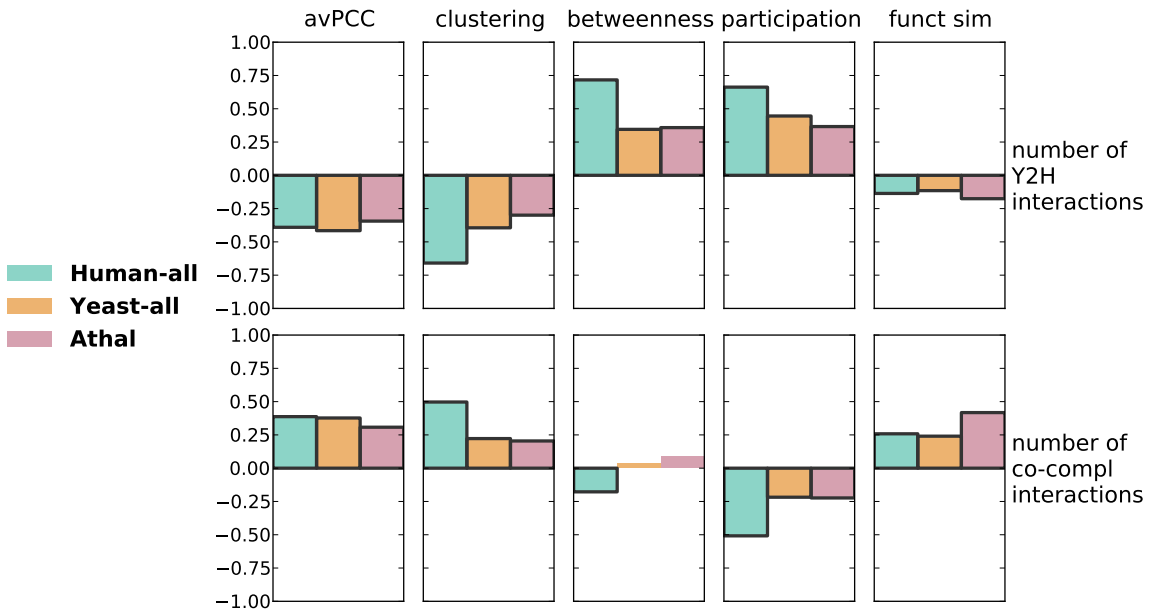

Figure 2.5: Hubs with different roles in interactomes are involved in interactions of different types.
For hubs and their characteristics determined from the full networks **Human-all**, **Yeast-all** and **Athal**, we measured the Spearman correlation between hub characteristics and the number of interactions of the type yeast two-hybrid or co-complex. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges. Hubs with higher avPCC, clustering coefficient and functional similarity tend to have more co-complex interactions, while hubs with higher betweenness and participation coefficient tend to have more yeast two-hybrid interactions.

avPCC is significantly positively correlated with essentiality (Fig. 2.6). To directly compare with the original study [24], we also compare date and party hubs and find a significantly larger fraction of essential genes in the set of party hubs than in the set of date hubs, as determined by the hypergeometric test (Fig. A.33). We further show that the correlation of avPCC and essentiality is significantly high even when compared with that found in random networks (see Section A.1.8 and Fig. A.35). We also find that essentiality is positively correlated with clustering and functional similarity, while negatively correlated with betweenness and participation (though this correlation is not significant for betweenness in **Yeast-all**). This is in agreement with recent evidence of the tight relationship of a protein's essentiality with modularity

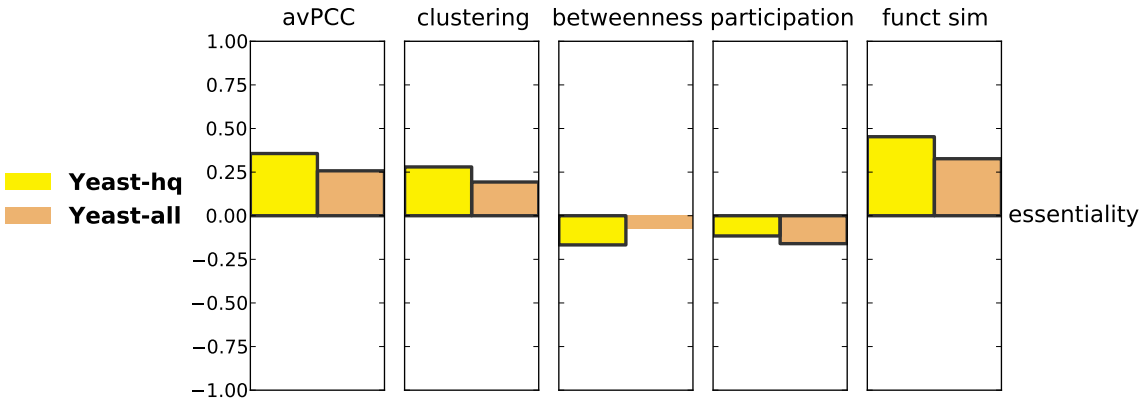

Figure 2.6: Hub characteristics in yeast protein physical interaction networks are correlated with protein essentiality.
Each bar represents a Spearman correlation between a hub characteristic and hub essentiality in one of the yeast networks. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges.

and its involvement in essential complexes [17, 48], as hubs with high avPCC, clustering, or functional similarity, and correspondingly low betweenness and participation, are likely to play key roles in modules and complexes.

2.2.8 Hub roles in the interactome are evolutionary conserved

In order to compare hub characteristics for similar genes from different organisms, we obtain sets of orthologous proteins from P-POD [62] for all organisms under consideration. As genes from the *E. coli* network have only a few orthologs in P-POD in the other networks, we focus this analysis on the four eukaryotic species. For each pair of networks of different organisms, we calculate the SRCCs of hub characteristics (avPCC, clustering, betweenness, participation and functional similarity, computed as described above independently for each network) over all pairs of orthologous hubs. For **Human-all** and **Yeast-all** we observe highly significant positive correlations which range from 0.38 for functional similarity to 0.76 for participation

Hub characteristic	ρ	p-value	Empirical p-value
avPCC	**0.55**	$3e{-}36$	< 0.001
clustering	**0.72**	$3e{-}70$	< 0.001
betweenness	**0.44**	$1e{-}22$	< 0.001
participation	**0.76**	$1e{-}82$	< 0.001
func. sim	**0.38**	$4e{-}16$	< 0.001

Table 2.2: Spearman correlation for characteristics of orthologous hubs in **Yeast-all** and **Human-all**.
Five hub characteristics for all 437 orthologous pairs between 291 hubs in **Yeast-all** and 299 hubs in **Human-all** are significantly positively correlated, as measured by Spearman's rho (ρ) and the correspondingly determined p-values and empirical p-values for 1000 random permutations of hubs. See main text and Section 2.4 for details.

(Table 2.2), and for **Human-hq** and **Yeast-hq** they range from 0.23 for avPCC to 0.62 for clustering (Table A.5). We note that some proteins may be involved in many orthologous pairs and therefore we also validate the significance of the observed correlations by randomly permuting hubs, and find that these results remain significant (see Section 2.4 for more details). These features are largely consistently positively correlated when comparing ortholog pairs between different pairs of networks (Tables A.6, A.7, A.8, A.9, A.10), though at varying levels of statistical significance. Further, we observe that purely topological features such as clustering coefficient and betweenness centrality are much more consistently conserved between pairs of networks than avPCC (Tables 2.2, A.5, A.6, A.7 and A.8), which is additional evidence that these topological features can reflect hub roles in the interactome.

One may suspect that the observed high correlation of hub features between organisms may be explained by conservation of modules that correspond to higher values of avPCC, clustering and functional similarity and to lower values of betweenness and participation. To exclude this possibility, for a pair of organisms, we first determine in each the hubs with the highest and lowest third of scores, according to any given hub measure. Next, we determine how many ortholog pairs are found between each of

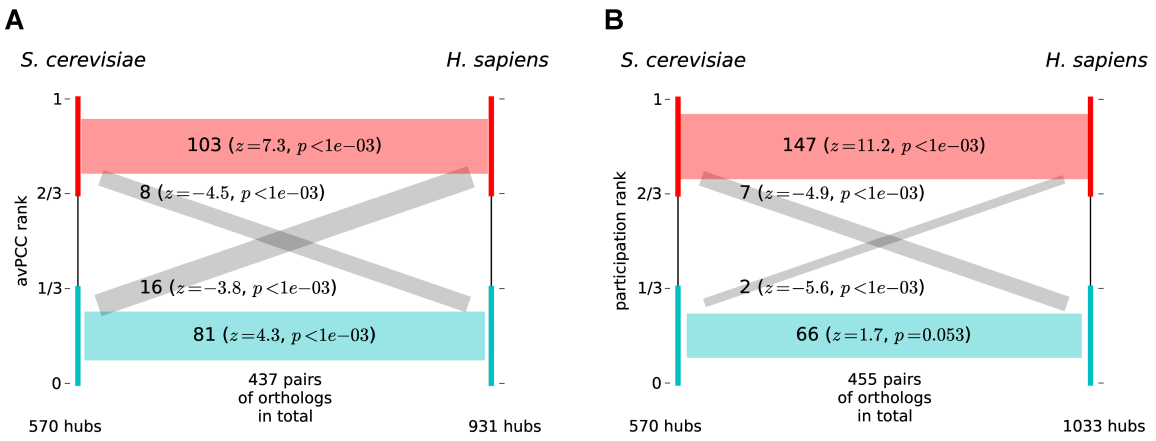

Figure 2.7: Characteristics of hubs are conserved across networks.
The top (respectively, bottom) third of hubs in each of **Yeast-all** and **Human-all**, as determined by (A) avPCC and (B) participation coefficient, are enriched in the number of orthologs between them. The number of orthologs between each of the groups is given, along with a z-score and p-value derived empirically from random samples of proteins for each group. Red bars indicate orthologous relationships between proteins in the top third of hubs, blue bars indicate orthologous relationships between proteins in bottom third of hubs, and gray bars indicate orthologous relationships between proteins that are in opposing groups in the two organisms.

the top and bottom groups in both organisms. We compare these numbers with the same values expected if these top and bottom sets of hubs were selected at random, rather than according to the hub score. We expect more ortholog pairs between the top thirds as well as between bottom thirds, and fewer orthologs between the top and bottom thirds. Indeed, this is what is observed for our hub measures (see, for example, avPCC when comparing the **Yeast-all** and **Human-hq** networks in Fig. 2.7A and participation for these networks in Fig. 2.7B).

2.3 Discussion

We have confirmed in protein interaction networks across a range of organisms that if hubs are partitioned into two classes according to their tendency to be co-expressed with their interacting partners, they exhibit significantly different properties and roles

in the interactome. In one class, hubs tend to have higher average co-expression with their interacting partners, higher clustering coefficients and higher functional similarities, but lower betweenness centralities and participation coefficients. These hubs are more often interacting with each other, and are enriched with co-complex interactions. Simulated removal of these hubs from the network does not greatly affect the connectivity of the network. These properties suggest that such hubs may act locally inside functional modules and protein complexes. In another class, hubs tend to have lower average co-expression, clustering, and functional similarity, but higher betweenness and participation. These hubs more often participate in genetic interactions, and are more often detected in yeast two-hybrid interactions, which are presumably enriched in binary transient interactions. These hubs tend to interact with each other less, and with other proteins more. After these hubs are removed, the network becomes more disconnected and clustered. These properties suggest that such hubs tend to be global connectors and coordinators of different modules in the interactome.

Initially, it was proposed that the distribution of the hubs' average co-expressions with their neighbors was bimodal, and it was argued that this naturally implied a categorization of all hubs into two classes with hypothetically different roles. Furthermore, Han *et al.* [24] proposed a view of the interactome with mostly non-intersecting independent modules, and certain proteins outside of these modules that connect and coordinate their functioning. This model, as well as the existence of the two classes of hubs with correspondingly different roles, has been the subject of some controversy. We argue that even though we believe the two classes of proteins in the interactome can be distinguished from each other and their roles can be recognized as different, it is not necessarily the case that all proteins or even just all hubs can be classified into one of the classes. Rather, the network is almost certainly more complex, with highly overlapping modules, multifunctional proteins, and proteins of mixed and not easily

detectable roles that depend upon conditions and time. A better understanding of the structure and functioning of the interactome will require large-scale annotation of interactions and interacting proteins with information about concentrations and the strength, condition and timing of when, where and how these interactions occur. These annotations are currently not available at a large scale, but may be obtained experimentally in the future, or by developing new methods for analysis of existing data. With uncertainty about the exact role of each particular protein in the interactome, measuring and analyzing their properties on a continuous scale may be more appropriate than trying to extract firm classes.

A significant amount of computational research has been devoted to uncovering the dynamics of protein interactions via integration with other types of data [63, 64, 45, 65, 23]. Currently, the most common approach to glean information about the dynamics of hubs and their interactions is by integrating interaction data with gene expression data, as is done here and previously using the measure of a protein's average co-expression with its interacting partners. However, we have shown that very similar information is reflected in the interaction data itself. Even though it is highly unlikely that just topological network features can describe all of the structure and dynamics of interactomes, analysis of topological characteristics in networks may be of great help in furthering our understanding of network dynamics.

As more large-scale protein interaction networks have become available, certain of their aspects and properties have been shown to be conserved across interactomes of different organisms [66, 67, 68, 69, 70]. Such conservation is strong evidence that a network feature reflects an important aspect of interactomes. We have shown that a protein's average co-expression over neighbors in its PPI network is conserved for orthologous hubs across different organisms, and have further confirmed that it is a biologically meaningful measure for understanding hub roles. At the same time,

we have shown that hub characteristics that depend purely on network topology are conserved at least as well as average co-expression.

Following previous work, in our analysis we have focused almost exclusively on hubs, a small fraction of proteins within interactomes. However, we have also demonstrated, by reducing the number of interactions required to call a protein a hub, that our observations hold when we consider many more proteins in the network, so it may be possible to classify not just hubs based on topological features or co-expression properties, but also proteins in general. Moreover, we have shown that our analysis is robust to noise in interaction data, as the trends we report are consistent not only across networks of lower coverage where interactions are additionally selected for high quality, but also across larger networks without additional quality filtering that are likely to contain more noise but also have higher coverage.

We have shown that topological features of proteins in the network capture functional and structural properties of networks. Therefore, the distribution of these features also, to some extent, characterizes the whole interactome. In the future, depending upon the application, it may be desirable to take these features into account when building and analyzing models for protein interaction networks, and in particular, within algorithms that are used for generating random networks in order to compare them with real data. Existing approaches for randomizing protein interaction networks have preserved local properties such as degree and local clustering coefficient, small subgraphs and schemas, as well as some evolutionary constraints [71, 72, 70, 73, 74, 75]. In addition to these features, in the future, randomization algorithms may try to also preserve measures such as betweenness centrality and participation coefficient, as we have demonstrated that these features capture additional information about network structure.

In sum, our observations provide a better understanding of the dynamic interactome of the cell. As more specific, high-quality and high-coverage protein-protein

interaction data become available, we believe our approaches to analyze these data can reveal further details about the structure, function and evolution of interactomes.

2.4 Materials and methods

2.4.1 Interaction networks

Seven interaction networks for five organisms are considered in our analysis. We briefly describe the networks below; further details can be found in Section A.2.2. In all networks, self-loops and duplicate interactions are deleted. The size of each network is shown in Table 2.1.

S. cerevisiae: The network **Yeast-all** consists of all yeast protein physical interactions from BioGRID [60] version 3.1.78. The high quality network **Yeast-hq** consists of all binary and co-complex interactions from HINT [76]. Yeast genetic interactions are obtained from BioGRID version 3.1.78 (123707 interactions).

H. sapiens: We use two human protein-protein physical interaction networks, both compiled by [77]. The first, **Human-all**, is their comprehensive network aggregated from numerous sources, and the second is their high quality subnetwork **Human-hq**.

D. melanogaster: **Fly** combines all interactions in DroID [78] version 2011_02 with those from DPiM [38].

A. thaliana: **Athal** consists of protein-protein interactions obtained from IntAct [9], BioGRID, and from the supporting material of [39].

E. coli: **Ecoli** consists of protein-protein physical interactions extracted via the PSICQUIC View application [79].

2.4.2 Network topology analysis

We briefly describe the topological measures that we utilize in our study.

The **degree** of a vertex is the number of interactions the corresponding protein has. In each network, we consider hubs to be proteins in the top 10% by degree, where the precise degree threshold to be called a hub is chosen such that at least 10% of vertices are hubs. These thresholds and the number of hubs for each network are shown in Table 2.1.

The **betweenness centrality** of a vertex v in a network is the number of shortest paths between all pairs of vertices in the network that pass through v, with the shortest paths between two genes s and t weighed inversely to the total number of distinct shortest paths between s and t.

The **clustering coefficient** of a node is defined as the ratio of the number of triangles containing that node to the number of triples centered on it; i.e., for a protein, this measures the number of interactions among its interactors, normalized by the maximum number of possible interactions.

The **participation coefficient** [80, 57] of a vertex with respect to a set of clusters in a network is defined as $P = 1 - \sum_i \left(\frac{k_i}{k}\right)^2$, where the summation is over all clusters, k is the degree of the vertex, and k_i is the number of edges going from the vertex to vertices in cluster i. Note that $P = 0$ if all edges from a vertex go to a single cluster, and P is closer to 1 if edges from the vertex are more uniformly distributed over clusters. To find clusters in the network, we used the SPICi clustering algorithm [22] with parameters optimized with a simple exhaustive search procedure to approximately maximize Newman's modularity measure [81]. See Section A.2.4 for details.

The **density** of a set of vertices S is the ratio of the actual number of edges between vertices in S to the maximum possible such number $|S| \cdot (|S|-1)/2$. The neighborhood of a set of vertices S is the set of all vertices that are connected to some vertex from S but are not themselves members of S. The **expansion** of a set of vertices S is the ratio $|N|/|S|$, where N is the neighborhood of S. When measuring the density or expansion of a class of hubs, we compare it with the density or expansion of a

random subset of the same number of background hubs as in the class in question. We consider 1000 independent samples, and report the empirical p-value of the actual value as compared to the distribution of random values.

The **average path length** for a network is measured as the average over all pairs of vertices of the lengths of the shortest paths between them. (For a disconnected network, only pairs of vertices connected by a path are considered.) The **relative size of the giant component** is calculated as the ratio of the size of the largest connected component in the network to the number of vertices in the network. The **global clustering coefficient** of a network measures the tendency of network vertices to cluster together. It is defined as thrice the number of triangles divided by the number of connected triples of vertices in the network.

In a **hub removal** experiment for a class of hubs, we remove all vertices of the class with their interactions from the network at once and measure the fold change of certain characteristics of the remaining network as compared with the initial network (e.g., if the average path length increased 1.23 times, then the fold change is 1.23). To compute an empirical p-value of this fold change value, we compare it with the distribution of the same values obtained after 1000 independent removals of random subsets of the same number of background hubs. We use the average path length, the size of the giant component, and the global clustering coefficient as global characteristics of network structure. By removing all hubs at once and comparing computed values with removals of random subsets of the same size, our hub removal experiment does not depend on the order in which hubs are removed or the size of the set of hubs considered, two issues which were raised previously [55, 56]. The results of these experiments can be compared for different classes of hubs, as in each case we compare the effect for a class of hubs relative to random subsets of the same size.

All topological measures are computed based on the python interface to the igraph library, version 0.5.4 (`http://igraph.sourceforge.net/`). We utilize degree-

preserving network randomizations, as implemented in the `igraph.Graph.Degree_Sequence()` method with the "vl" option [59].

2.4.3 Expression

Expression compendia for each organism consist of datasets collected from online databases and papers, as described in detail in Section A.2.3, and for each organism cover a wide range of conditions and/or tissue and cell types (where applicable). Each dataset is processed independently as follows: all replicates are merged (gene expression values averaged over replicates of the same experiment); genes with less than 50% known values are removed; the $\log_2$-transformation is applied to all values if absolute signal values are given; for each matrix column corresponding to a single genome-wide experiment, the values of the column are transformed to z-scores.

For each organism, for each interacting pair of genes, we compute their co-expression via the Pearson correlation coefficient (PCC) of their expression profiles as follows. For genes with incomplete expression profiles within a dataset, only dimensions where values for both genes in the pair are known are used when computing the PCC of this pair. If the expression compendium for an organism consists of several datasets, the PCC is computed for each dataset independently, and then these PCC values are averaged with weights proportional to the number of expression datapoints that the dataset contributed to the compendium (in case of incomplete data, this is only over datasets where the PCC could be successfully computed), to obtain a final co-expression interaction score.

Some proteins in the networks are not included in any expression datasets. These proteins are not used to compute PCCs and avPCCs (see below), but may still contribute to degree or other topological properties of proteins. The number of interactions in the networks for which co-expression values are computed is shown in Table 2.1.

2.4.4 Hub scores and classifications

For each hub, the average co-expression score (avPCC) is computed as the average of its co-expression interaction scores [24]. More precisely, the avPCC of a hub is the sum of all defined co-expression scores for interactions of the hub divided by hub degree (thus unknown edge scores are effectively assumed to be 0). Hubs are scored with avPCC only if they have at least 3 interactions with defined co-expression score.

Extremal hubs are defined as hubs in the top 5% by either degree or betweenness centrality amongst all hubs. For most networks, these two subsets of hubs are highly intersecting, so the union contains much less than 10% of all hubs. These hubs are excluded from the classification of hubs into date and party, and the corresponding analysis of this classification, but may still contribute to properties of other genes, particularly other hubs. Further, the background set of hubs, from which random sets of hubs are chosen to compute empirical p-values of several properties (as described above), does not include extremal hubs. Note, however, that extremal hubs are not excluded when doing correlation analysis of hub characteristics.

Party hubs are defined as the top one third by avPCC amongst all non-extremal hubs, and the remaining non-extremal hubs are defined as date hubs.

2.4.5 Gene ontology analysis

For our functional analysis, we use Gene Ontology (GO) [7] terms and gene association data for each organism, not including associations with evidence codes IEA, RCA, IPI, ND or the qualifier NOT (downloaded from `http://www.geneontology.org/` on July 25, 2011). The **functional similarity** of a pair of genes is computed as described in [57]. First, the information content of a term t is defined as $s(t) = -\log \frac{|t|}{|G|}$, where $|t|$ is the number of genes annotated with the term, and $|G|$ is the total number of genes in the organism annotated with at least one term. Then if $T(g)$ and $T(h)$ are the sets of terms annotating genes g and h respectively, functional similarity is computed

as $f(g,h) = \frac{\sum_{t \in T(g) \cap T(h)} s(t)}{\sum_{t \in T(g) \cup T(h)} s(t)}$. For functional similarity, all GO Biological process terms of depth ≥ 2 annotating at least 3 and at most 1000 genes are considered. The functional similarity of a vertex in a network is the average of functional similarity of this gene with all its interacting partners; proteins not annotated with one of the terms under consideration lead to functional similarities of 0.

We perform GO annotation enrichment test using the code of the project `goatools` (`https://github.com/tanghaibao/goatools`). We apply it to groups of hubs in the top and bottom one third or two thirds by avPCC, clustering, betweenness, participation and functional similarity in each network. For this analysis, we use broad functional terms that are direct children of roots of all three ontologies: biological process, 28 terms; cellular component, 13 terms; and molecular function, 20 terms. We use the set of all annotated hubs as the background population, and report terms with a Bonferroni-corrected p-value of less than 0.05. For each network, we test enrichment for each ontology (e.g., Biological process ontology) independently, and restrict the analysis only to the hub proteins that have at least some annotation with terms other than the root (e.g., Biological process) in this ontology.

2.4.6 Essential genes

The 1222 essential genes for *S. cerevisiae* are obtained from the Saccharomyces Genome Deletion Project webpage (file `http://www-sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt`).

2.4.7 Orthologs

We use protein ortholog information from version 4 of the Princeton Protein Orthology Database (P-POD) [62] (`ftp://gen-ftp.princeton.edu/ppod/`). We consider

two proteins in different networks to be orthologous if they are categorized in the same family by P-POD using either OrthoMCL or MultiParanoid.

For each pair of networks for two different organisms, we consider each pair of hubs (H_1, H_2) where H_1 and H_2 are non-extremal hubs in the networks of organisms 1 and 2, respectively, that are reported to be orthologous. A hub can appear in several pairs if it has more than one ortholog in another species. The Spearman correlation coefficient is computed over hub pairs for various network characteristics (avPCC, clustering coefficient, etc.).

Since a hub may contribute to several pairs of orthologs, in addition to using the standard computation of the p-value for the Spearman correlation coefficient, we also calculate an empirical p-value in the following way: the actual Spearman's rho is compared with the distribution of Spearman's rho values calculated in exactly the same manner as above, but for the characteristic (say, avPCC) among hubs randomly permuted in each of the two networks (as opposed to permuting vector components that contain repetitions). We report the empirical p-value of the actual correlation with respect to the distribution of correlations from 1000 instances of randomized data.

We test if the low range of a hub feature (say, avPCC) is evolutionary conserved as much as the high range. For two networks of different organisms, we extract hubs in the top one third and bottom one third as ranked by the feature computed in each network. We calculate for each pair of hub groups (top third from the first organism vs. top third from the second organism, top third from the first organism vs. bottom third from the second organism, etc.) how many ortholog pairs are observed between them. Then we compare this number with the number calculated in exactly the same manner, but for 1000 random samples of the same number of hubs in each network, and report the corresponding z-score and empirical p-value of the actual number of orthologs compared with the distribution of the numbers for random data.

Chapter 3

Genome-wide detection and analysis of multifunctional genes

3.1 Introduction

Multifunctionality can be defined as the involvement of a gene in multiple cellular processes [82]. This can come about either because the protein coded by the gene is capable of performing distinct molecular functions [25, 26, 83, 84, 85], or as a result of the same molecular function being reused in different contexts [28, 86]. Pioneering experimental work led to the surprising finding that crystallins—the proteins responsible for the optical properties of the eye lens—can also play non-refractive roles and have enzymatic activity in other tissues [25]. This evolutionary strategy was named "gene sharing" [27]. Further examples of proteins performing multiple molecular functions were subsequently described: a uracil-DNA glycosylase that can also function as glyceraldehyde-3-phosphatase dehydrogenase, or an enzyme like thrombin that can moonlight as a ligand for surface receptors [26]. More recently, a large-scale screening of mutants in yeast was performed to measure the pleiotropic effects of genes under different conditions [87]. In the case of pleiotropy, a gene may perform only one

molecular function, but it can be involved in multiple biological processes, and its perturbation can therefore have pleiotropic consequences.

Based on existing functional annotations of genes, it is likely there are numerous multifunctional genes within organisms. Despite the prevalence of multifunctional genes, multifunctionality remains a poorly understood phenomenon. Identifying multifunctional genes at a genome-wide level and studying their properties can shed light upon the complexity of the molecular events that underpin cell function, leading to a new understanding of the functional landscape of genes.

Earlier computational studies have attempted to identify multifunctional genes from the availability of functional annotations for genes in different organisms. Several previous works measured multifunctionality by simply counting the number of distinct Gene Ontology (GO) biological process terms annotating a gene product [88, 89, 90]. While straightforward, this approach does not always guarantee that a gene annotated with more than one GO term is indeed involved in two distinct biological processes. In particular, this is an incorrect assumption not only when one term is a direct descendant of another term in the GO hierarchy, but also even when two terms are in completely different branches of the ontology, as idiosyncrasies in GO may lead to similar processes being categorized in distinct places in the ontology. Other approaches have used protein-protein interaction data and defined as multifunctional those proteins that are located at the intersection of overlapping clusters [91]. However, using interaction data to identify multifunctional genes has the obvious drawback of preventing an unbiased analysis of their network properties, as well as uncertainty due to the computationally derived clusters themselves.

We develop a computational approach to leverage GO functional annotations of genes in a systematic and robust manner. To handle similar terms that appear in distant places in GO, we explicitly select sets of terms that co-occur less frequently than expected by chance; these terms are then used to identify multifunctional genes.

We apply our procedure to detect multifunctional genes to three organisms—human, fly and yeast—and then in each organism compare the properties of multifunctional genes against those of other genes. Our results across these species consistently show that, as compared to other genes, multifunctional genes possess distinct physicochemical properties, are more broadly expressed across cell types and tissues, tend to be more evolutionarily conserved, are more likely to be essential, and are topologically distinct in protein-protein interaction networks, in regulatory transcription factor–gene networks and in genetic interaction networks. We also find that multifunctional genes are significantly more likely to be involved in human disorders than other genes. The same observations also hold for genes with multiple molecular functions.

3.2 Results

3.2.1 Genome-wide detection of multifunctional genes

We use functional annotations of genes in three organisms, *H. sapiens*, *D. melanogaster*, and *S. cerevisiae*, to identify multifunctional genes in each of them at a genome-wide level. To accomplish this, we use Biological Process GO annotations [7]. We define as multifunctional all genes that have two or more distinct annotations by GO terms. To be able to detect truly distinct terms, we require that they have comparable specificity. The method is shown schematically in Fig. 3.1, and is briefly described below (see Section 3.4 for details).

The Biological Process GO is a hierarchy of terms representing different aspects of biological processes in a living organism. The terms range from very general (the most general being `biological process`) to very specific, with a relationship between terms indicating if a term implies another term. That is, each term annotates a set of genes, and a term should annotate all genes that are annotated by the terms indicated as more specific than this one in the hierarchy.

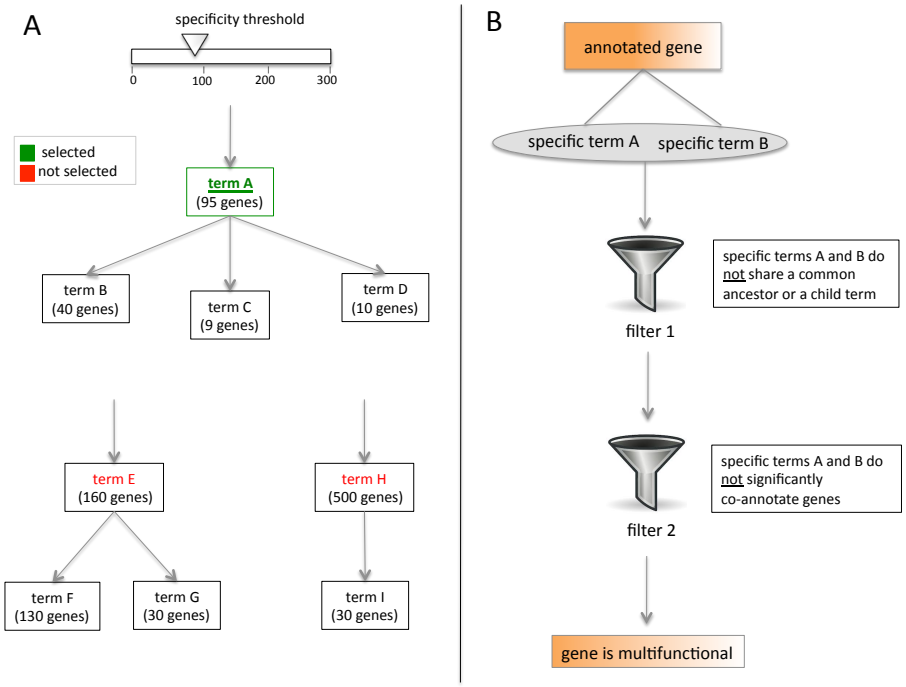

Figure 3.1: Schematic representation of the pipeline to identify multifunctional genes. We define as multifunctional all genes that have two or more annotations by distinct terms of comparable specificity. (A) First, we extract a subset of Gene Ontology terms at a comparable level of specificity. For a specificity threshold N, we select all terms which annotate $\geq N$, but $< 2N$ genes, and each of their descendant terms annotates $< N$ genes. For example, if $N = 90$, then term A is selected because it annotates more than 90 genes and less than 180 genes, and each of its descendant terms annotates less than 90 genes. In contrast, term E is rejected, because its descendant term F annotates more than 90 genes. Term H is also rejected, because it annotates more than 180 genes. (B) Once the terms at a certain specificity level have been selected, we extract all genes annotated with at least two such terms. In order to consider annotations by distinct terms only, from the collection of all pairs of terms selected at the chosen level of specificity, we filter out those that either share a common ancestor (other than the root) or have a common descendant term in the GO graph. Further, we remove all pairs of terms that co-annotate more genes than expected by chance, as measured by the hypergeometric test. All genes co-annotated by some pair of terms (chosen at any considered level of specificity) passing these two filters are considered multifunctional.

We start by selecting a subset of terms that annotate roughly the same number of genes. The set of specific terms can be chosen at different specificity levels, represented by a parameter N roughly corresponding to the number of genes annotated by a term. Lower values of this parameter produce larger numbers of more specific terms, and higher values result in smaller numbers of more general terms (Fig. B.1). We will consider several distinct levels of specificity in identifying multifunctional genes and call multifunctional all genes for which we find evidence of multifunctionality at some specificity level.

Once the terms have been selected at a particular specificity level, we extract all genes annotated with at least two such terms. In order to select only pairs of distinct terms and make sure a gene annotated by both terms is truly multifunctional, we apply several filters to pairs of terms. From the collection of all pairs of terms at a particular specificity level, we filter out those that either share a common ancestor (other than the root) or have a common descendant term in the GO graph, as these events indicate that the terms are semantically related. However, this is not sufficient to claim that the remaining pairs of terms are distinct. For example, the terms `aerobic respiration` and `mitochondrial translation` do not have any ancestral or descendant term in common in the GO hierarchy graph besides the most general `biological process` term, but often co-annotate mitochondrial ribosomal proteins and capture semantically distinct aspects of the same function. Therefore, we further remove all pairs of terms that co-annotate more genes than expected by chance (detected by hypergeometric test). All genes co-annotated by some pair of chosen terms passing these two filters, for any set of chosen terms at each specificity level N considered, are called multifunctional. The more general terms allowed (i.e., the higher the upper bound on N), the more multifunctional genes are detected (Fig. B.1).

In what follows, we compare multifunctional genes with all other annotated genes in fly, human, and yeast in order to uncover significant differences in biological prop-

Organism	Number of multifunctional genes detected	Total number of annotated genes
D. melanogaster	1509	6354
H. sapiens	2517	9664
S. cerevisiae	876	4682

Table 3.1: Number of multifunctional genes
For each organism, we show the number of multifunctional genes detected by our method and the total number of annotated genes (annotated by one of the terms used to detect multifunctionality; see Fig. 3.1 and Section 3.4).

erties between the two groups. The number of multifunctional genes and the total number of annotated genes for each organism is given in Table 3.1.

We note that multifunctional genes may appear more often in results of various experiments and thus be more actively studied by researchers, and this could potentially introduce a study bias in our analysis. In order to avoid this, in what follows, we mostly focus on analysis involving unbiased high-throughput and whole-genome data sets. When looking at association of multifunctional genes using manually curated data which could potentially be biased, we directly correct for the study bias in order to observe significant biological differences of multifunctional genes.

3.2.2 Proteins encoded by multifunctional genes are longer, have more domains and have a higher fraction of disordered residues

We start the analysis by studying some basic physicochemical properties of proteins. First, we hypothesize that multifunctional proteins may be longer in order to accommodate more functional domains. To test this hypothesis, we compare the lengths of the proteins encoded by multifunctional and other annotated genes in *D. melanogaster*, *H. sapiens*, and *S. cerevisiae*, and find that multifunctional genes are

significantly longer than other genes (p-values $1e{-}39$, $1e{-}9$, and $8e{-}12$, respectively, Mann–Whitney U test), on average by 39%, 16%, and 15%, respectively (Fig. 3.2). We also observe that proteins encoded by multifunctional genes have significantly higher numbers of distinct domains per protein (p-values $2e{-}7$, $1e{-}10$, and $2e{-}4$, respectively), on average by 17%, 13%, and 8%, respectively (Fig. 3.2). However, note that longer proteins have more domains, so the difference in length between multifunctional and non-multifunctional genes could explain the observed difference in the number of domains; we cannot conclusively distinguish the cause from the consequence (see Section B.1.1).

Another mechanism that has been proposed to explain multifunctionality is the presence of intrinsically unstructured regions [29]. To determine whether multifunctional proteins tend to be more disordered, we predict the fraction of disordered residues using the IUPred program [92, 93], and find that multifunctional genes in *D. melanogaster*, *H. sapiens*, and *S. cerevisiae* have a significantly higher fraction of predicted disordered residues (p-values $6e{-}21$, $7e{-}4$, and $3e{-}14$, respectively), on average by 26%, 5%, and 31%, respectively (Fig. 3.2).

Overall, we find that proteins encoded by multifunctional genes are longer, have more domains and are more disordered than other annotated genes.

3.2.3 Multifunctional genes are expressed more broadly in fly and human

Differential gene expression is key in tissue and cell specificity. A gene expressed in different contexts may have different functions depending upon how and when it is expressed. Therefore we hypothesize that a gene associated with several functions may be expressed in a larger number of contexts. In order to assess the relationship between gene expression and gene multifunctionality, we use genome-wide mRNA expression data and count in how many conditions, tissues or cell types each gene

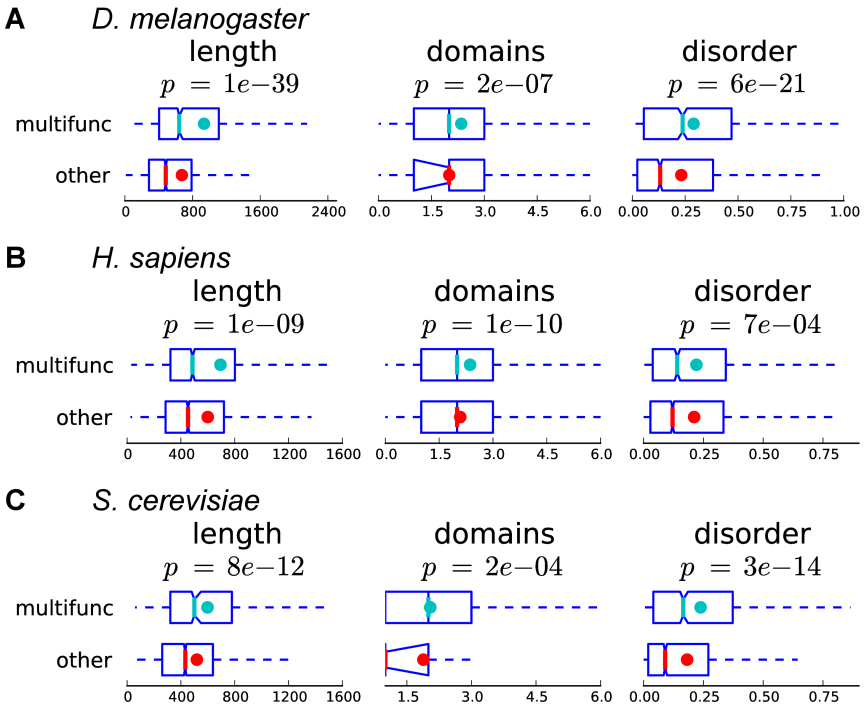

Figure 3.2: Proteins encoded by multifunctional genes are longer, have more domains and are more disordered.
Boxplots for length, number of domains, and fraction of disordered residues in proteins encoded by multifunctional and other annotated genes are shown for (A) fly (B) human and (C) yeast. Colored dots show the means, notches show bootstrap-generated 95% confidence intervals around the medians, boxes show quartile ranges, whiskers extend to the most extreme data points within 1.5 times the size of the inner quartile range. For genes in fly and human, if a gene had more than one protein isoform, the longest isoform was considered. Multifunctional genes are significantly longer, have significantly larger number of domains, and are significantly more disordered (Mann–Whitney U test).

is expressed. For fly, we use two datasets: FlyAtlas [94], the Drosophila microarray gene expression atlas across different tissues in larva and adult, and RNA-seq data from modENCODE across many different tissues and development time points, as aggregated by FlyBase [95, 96]. For human, we use information about organism parts in which genes are expressed, obtained from Ensembl BioMart [97]. We observe that in both human and fly, multifunctional genes are expressed more broadly than other

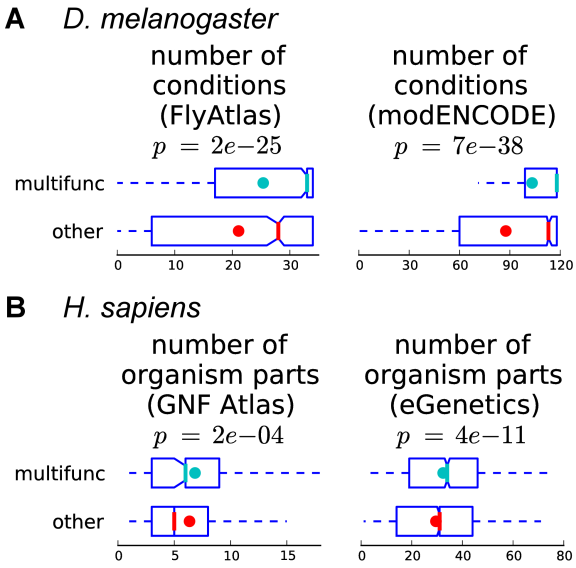

Figure 3.3: Multifunctional genes are more broadly expressed.
Boxplots of the number of organism parts and/or conditions in which multifunctional and other annotated genes are expressed are shown for (A) fly (microarray expression data from FlyAtlas and RNA-seq expression data from modENCODE) and (B) human (GNF atlas and eGenetics expression data obtained from Ensembl). Colored dots show the means, notches show bootstrap-generated 95% confidence intervals around the medians, boxes show quartile ranges, whiskers extend to the most extreme data points within 1.5 times the size of the inner quartile range. Multifunctional genes are expressed in a significantly larger number of conditions than other annotated genes (Mann–Whitney U test).

annotated genes; that is, they are expressed in significantly larger number of tissues or organism parts (p-values from $7e{-}38$ to $2e{-}4$, Mann–Whitney U test; Fig. 3.3A-B).

A potential mechanism of gene multifunctionality is the production of multiple protein isoforms with different functions using alternative splicing. Indeed, we observe that multifunctional genes have a significantly larger number of known isoforms in fly and human (Fig. B.2). If different isoforms of a gene have different expression patterns, this gene may be detected as broadly expressed in genome-wide assays, which currently report expression only at the gene level, merging information about expression of different isoforms. Indeed, we observe a strong correlation between the number of isoforms per gene and the number of tissues or organism parts in which

it is expressed (Table B.1). However, when comparing genes with an equal number of known isoforms, we still observe that multifunctional genes are expressed in larger numbers of tissues or organism parts (although most p-values for human are above our significance threshold of 5%; Fig. B.2). This indicates that multifunctional proteins are more broadly expressed regardless of the number of isoforms.

We conclude that multifunctional genes are consistently more broadly expressed than other annotated genes.

3.2.4 Multifunctionality is evolutionarily conserved

Acquiring multiple functions may constitute a special evolutionary strategy and limit gene evolutionary rates; indeed it has been previously suggested that genes with multiple functions or associated phenotypes tend to be more evolutionary conserved [27, 88, 28, 98]. In order to study the evolutionary dynamics of gene multifunctionality at a genome-wide level and in an unbiased manner, we use evolutionary conservation scores from phastCons [99]. Scores in phastCons are computed using phylogenetic hidden Markov models of multiple sequence alignments of *D. melanogaster* with 14 other insect genomes, *H. sapiens* with 99 other vertebrate genomes, and *S. cerevisiae* with 6 other yeast species. For each nucleotide of the genome, phastCons produces a score between 0 and 1, where higher values indicate stronger evolutionary conservation. For each gene, we average the scores of all nucleotides of each isoform of the gene, and then average over all isoforms of the gene to obtain a single value for each gene, which is an estimate of how evolutionarily conserved the gene is. Previously, a positive correlation between the number of biological process GO terms and evolutionary conservation was observed for yeast [88, 28, 98]. In agreement with this, we find that in fly, human, and yeast, multifunctional genes are significantly more evolutionarily conserved than other annotated genes (p-values $5e{-}13$, $6e{-}10$, 0.02, respectively, Mann–Whitney U test; Fig. 3.4).

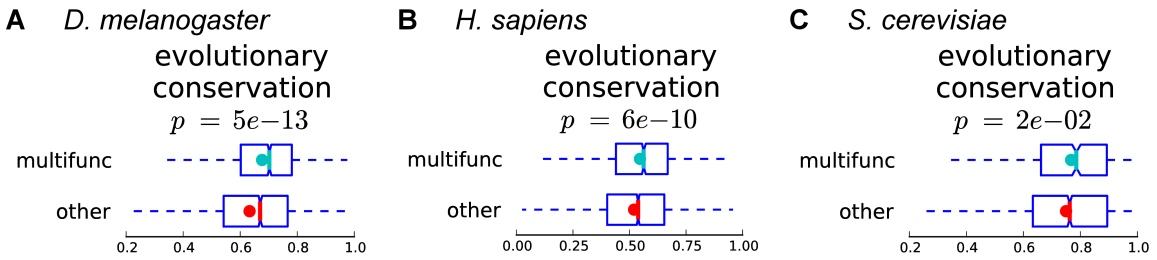

Figure 3.4: Multifunctional genes are more evolutionarily conserved. Boxplots of evolutionary conservation (estimated by phastCons [99] for each nucleotide, averaged over nucleotides of each gene) of multifunctional and other annotated genes are shown for (A) fly (B) human and (C) yeast. Colored dots show the means, notches show bootstrap-generated 95% confidence intervals around the medians. Multifunctional genes are significantly more evolutionary conserved than other genes (Mann–Whitney U test).

Having showed that multifunctional genes evolve more slowly, we next hypothesized that some of them may have become multifunctional early in evolutionary history. In order to test this, we compare the property of multifunctionality for orthologous proteins from different organisms. We use information about protein orthology from P-POD [62] and count how many orthologs are observed between proteins encoded by multifunctional genes from different organisms. Between fly and human, we observe 1725 orthologous pairs of genes where one gene in a pair is classified as multifunctional in fly and another gene in the pair is classified as multifunctional in human. To assess significance, we compute the same number when randomly reshuffling multifunctional and non-multifunctional genes from orthologous pairs in each organism, and observe on average only 845.1 ± 90.0 orthologous pairs, the actual value being 2.0 times higher (empirical p-value $< 1e{-}3$). For fly and yeast, we find 388 orthologous pairs between multifunctional genes (2.1 times higher than 184.7 ± 20.2 expected by chance, $p < 1e{-}3$). For human and yeast, we find 576 orthologous pairs between multifunctional genes (2.2 times higher than 267.2 ± 32.6 expected by chance, $p < 1e{-}3$). We conclude that the property of multifunctionality is conserved across orthologous genes of different organisms.

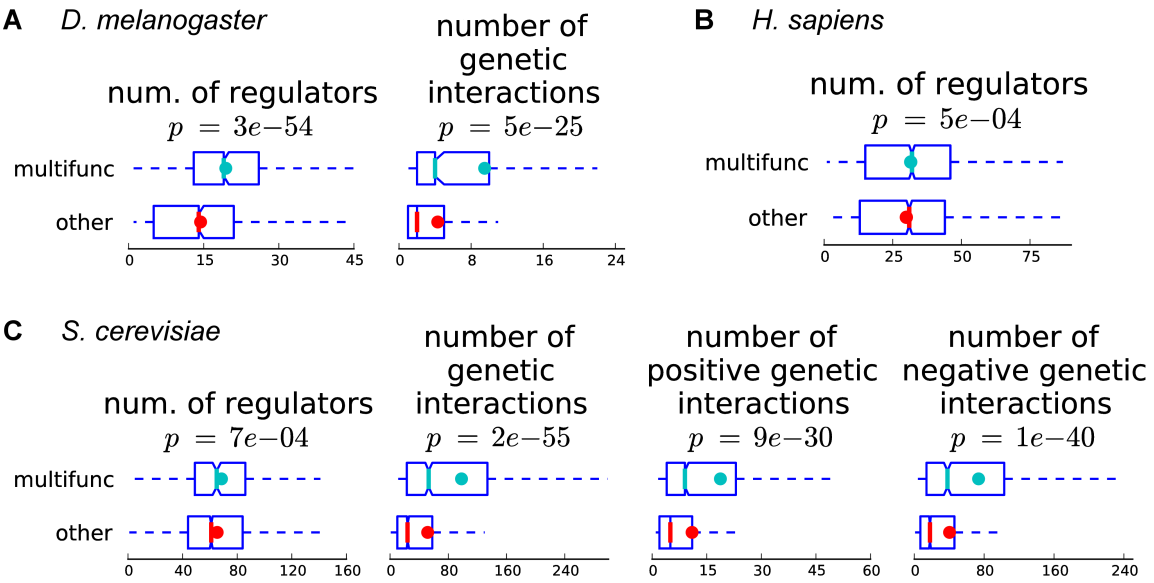

Figure 3.5: Multifunctional genes are involved in a significantly larger number of regulatory and genetic interactions.
Boxplots of the number of regulatory and/or genetic interactions for multifunctional and other annotated genes are shown for (A) fly, (B) human, (C) yeast. Colored dots show the means, notches show bootstrap-generated 95% confidence intervals around the medians, boxes show quartile ranges, whiskers extend to the most extreme data points within 1.5 times the size of the inner quartile range. Multifunctional genes are involved in significantly many more regulatory and genetic interactions (Mann–Whitney U test).

3.2.5 Multifunctional genes are involved in more regulatory and genetic interactions

Genes responsible for multiple functions in a cell may require more complex regulatory programs to differentiate functions across multiple tissues or conditions. In order to study how regulated multifunctional genes are, we use regulatory interactions from high-throughput ChIP experiments [78, 100, 101, 102]. For each gene, we count the number of transcription factor–target interactions this gene participates in as a target. In all three organisms, we observe that multifunctional genes are regulated by a significantly larger number of transcription factors than are other annotated genes (p-values from $3e{-}54$ to $7e{-}4$, Mann–Whitney U test; Fig. 3.5).

In addition to requiring more complex regulatory programs, multifunctional genes may also be associated with more complex phenotypes that require interactions across many other genes. In order to compare the distribution of genetic interactions between multifunctional and other annotated genes, we use a collection of genetic interactions curated by FlyBase [96] for fly and by BioGRID [5] for yeast. Previously, a positive correlation between the number of biological process GO annotations and the number of genetic interactions was observed for yeast [89]. In agreement with this, we observe that in fly and yeast, the number of genetic interactions is significantly higher for multifunctional genes than for all other annotated genes (p-values $5e{-}25$ and $2e{-}55$, respectively; Fig. 3.5). Moreover, in a more refined comparison for yeast, we observe that both the number of positive and the number of negative genetic interactions are significantly larger for multifunctional than for non-multifunctional genes (p-values $9e{-}30$ and $1e{-}40$, respectively; Fig. 3.5).

3.2.6 Multifunctional genes are more often essential

A gene associated with multiple functions may be more important for the normal functioning of the cell and therefore may potentially be more critical for survival than a gene associated with a single function. In order to test this hypothesis, we consider the relationship between gene essentiality and multifunctionality.

For fly, we call essential all genes with lethal phenotype (as curated by FlyBase [96]) and observe that 74% of multifunctional genes are essential, while only 44% of other annotated genes are essential ($p < 2e{-}86$, hypergeometric test; Fig. 3.6A). In addition, we use data from a genome-wide RNAi screen in cell lines [103] and observe that, even though only a small fraction of genes in the study overall are detected as essential, multifunctional genes have a significantly higher fraction of essential genes than other annotated genes (3.8% and 2.9%, respectively, $p < 0.046$; Fig. 3.6B).

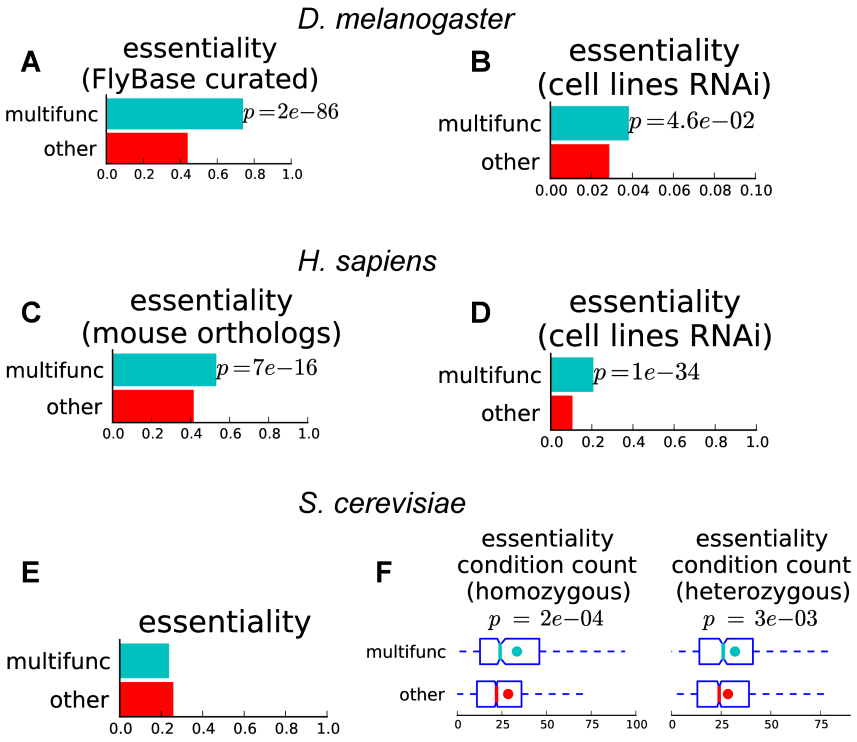

Figure 3.6: Multifunctional genes are more likely to be essential.
Barplots showing the fraction of multifunctional and other annotated genes that are essential in (A–B) fly (A, essentiality data from FlyBase; B, genome-wide RNAi screen in cell lines, note different scale on x-axis) (C–D) human (C, orthologs of essential genes in mouse; D, genome-wide RNAi screen in cell lines) and (E) yeast (essentiality screen in rich medium). In fly and human, multifunctional genes are essential significantly more often, whereas in yeast the difference is not significant (hypergeometric test). (F) Boxplot showing, for yeast genome-wide homozygous and heterozygous gene deletion screen across a variety of conditions, the number of conditions in which a gene is essential. Multifunctional genes are essential in significantly larger number of conditions.

For human, we call essential all genes which have a mouse ortholog with a lethal phenotype (according to MGI [104]). We find that 53% of multifunctional genes are essential, whereas only 42% of other genes are ($p < 7e{-}16$; Fig. 3.6C). Using data from a genome-wide RNAi screen in human mammary cells [105], we also observe that multifunctional genes are essential significantly more often ($p < 1e{-}34$; Fig. 3.6D). In a more detailed analysis using quantitative data about essentiality in 72 human

cancer cell lines [106, 107], we confirm that in all 72 cell lines, multifunctional genes are more essential (Fig. B.3).

In contrast, for yeast, when using information about essentiality for growth in rich medium, we do not observe a significant difference in essentiality: 24% of multifunctional genes and 26% of other annotated genes are essential ($p = 0.11$; Fig. 3.6E). However, in a genome-wide screen of yeast homozygous and heterozygous deletion strains across a variety of conditions, up to 97% yeast genes are reported as essential in at least one condition [108]. Using these data, we count in how many conditions each gene is detected as essential, and find that multifunctional genes are essential in a significantly larger number of conditions than other annotated genes (p-values $2e{-}04$ and $3e{-}03$ for homozygous and heterozygous screens, respectively; Fig. 3.6F).

Overall, we observe that multifunctional genes are more likely to be essential than other annotated genes.

3.2.7 Multifunctional genes are more often involved in human disorders

Being more critical than other genes for the survival and normal functioning of the cell, multifunctional genes may potentially also be more likely to be associated with human diseases. To address the relationship between gene multifunctionality and involvement in human disorders, we use the gene-disease "morbid map" from the Online Mendelian Inheritance in Man (OMIM) catalog [109], and calculate the fraction of genes with an OMIM annotation among multifunctional genes found for human. We find that 32% of all multifunctional genes are involved in at least one Mendelian disorder, whereas the fraction of other annotated genes involved in at least one Mendelian disorder is 21% ($p < 8e{-}30$, hypergeometric test; Fig. 3.7A).

To further investigate the relationship between gene multifunctionality and involvement in human disorders, we look at genes involved in multiple distinct dis-

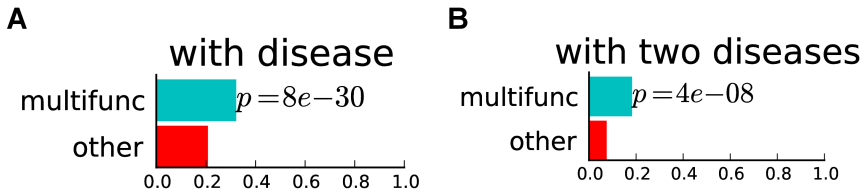

Figure 3.7: Multifunctional genes in human are associated with more diseases. (A) Barplot showing the fraction of multifunctional and other annotated human genes that are associated with disease. (B) Barplot showing for the genes associated with disease, the fraction of multifunctional and other annotated human genes that are associated with two or more diseases. Multifunctional genes are associated with significantly larger number of diseases (hypergeometric test).

orders. We map OMIM terms onto the Disease Ontology [110] and identify genes with at least one pair of disjoint OMIM terms (i.e., diseases that fall into separate branches of the Disease Ontology). We consider these genes to be involved in two or more distinct diseases. When considering the genes involved in at least one disease from Disease Ontology, we find that 18% of multifunctional genes are involved in at least two diseases, while only 8% of other such genes are involved in at least two diseases ($p < 4e{-}8$; Fig. 3.7B).

One might expect that genes involved in more disorders, as well as multifunctional genes, are more actively studied by the research community, and this could potentially introduce a study bias in our observations. Using the number of PubMed publications associated with a gene as a proxy to how well studied the gene is, we indeed confirm that multifunctional genes are more actively studied (Fig. B.4), but show that our observations still hold when correcting for this bias. Namely, we observe that the fraction of multifunctional genes associated with disease is higher than that for non-multifunctional genes with the same number of associated publications as for multifunctional genes (Fig. B.5).

Overall, we observe that multifunctional genes are associated with diseases significantly more often than other annotated genes.

3.2.8 Multifunctional genes tend to be intermodular in protein interaction networks

Genes associated with multiple functions may potentially play a more central role in the global functional organization of the cell. Large-scale networks of physical protein-protein interactions provide a good view of the cellular functional landscape. In order to study how multifunctional genes are positioned in protein interaction networks, we use the interaction data curated by BioGRID [5]. We use three measures of centrality: degree, betweenness centrality, and participation coefficient. Degree is the number of interactions in which a protein is involved. Betweenness centrality is the number of shortest paths passing through a node in the network, and nodes with higher betweenness are more globally central in the network. Participation coefficient shows how well a protein's interacting partners are distributed among clusters in the network, so that proteins with low participation are mostly interacting with proteins from the same cluster, whereas proteins with high participation have their interactions spread among many clusters.

We observe that with respect to all three considered measures, multifunctional genes are significantly more central than other genes (p-values from $2e{-}13$ to $3e{-}50$, Mann–Whitney U; Fig. 3.8). However, not surprisingly, degree is correlated with betweenness and participation (Fig. B.6), and the correlation of multifunctionality with degree could potentially be the only explanation for the correlation with the other two more complex measures. In order to test for this, we perform the comparisons of betweenness and participation between multifunctional and other annotated genes correcting for degree distribution, and still observe that multifunctional genes have significantly larger betweenness and participation (Fig. B.6 and Table B.2).

In order to show that our observations are not affected by potential study biases, we repeat the comparisons of degree, betweenness, and participation between multifunctional and other annotated genes in networks containing only interactions

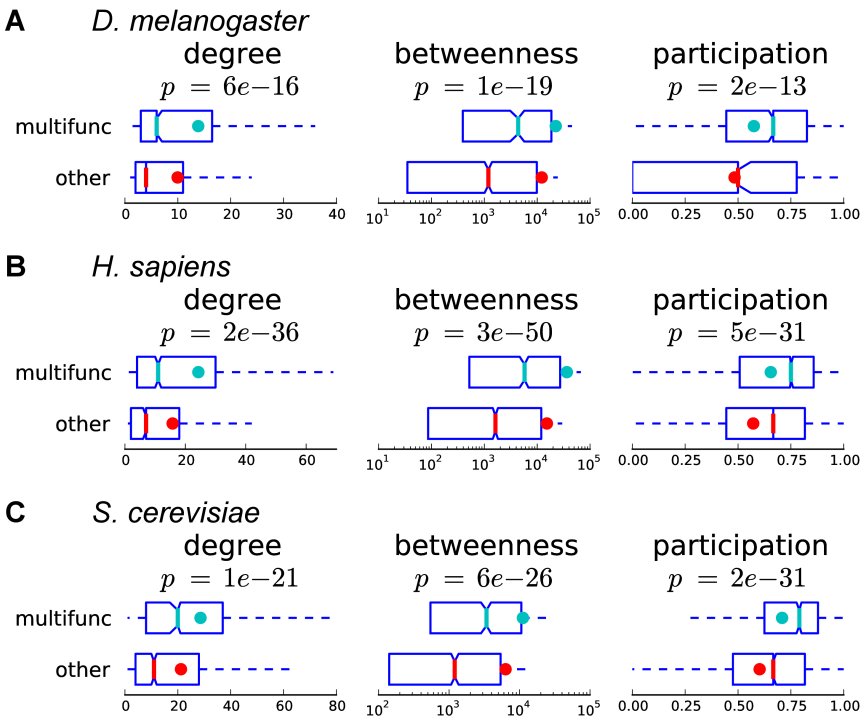

Figure 3.8: Multifunctional genes are more central in protein physical interaction network.
Boxplots of degree (number of interactions), betweenness centrality, participation coefficient of multifunctional and other annotated genes in the protein interaction network are shown for (A) fly, (B) human, (C) yeast. Colored dots show the means, notches show bootstrap-generated 95% confidence intervals around the medians, boxes show quartile ranges, whiskers extend to the most extreme data points within 1.5 times the size of the inner quartile range. According to all three measures of centrality, multifunctional genes are significantly more central than other genes (Mann–Whitney U test).

from high-throughput experiments, as reported in BioGRID [5] or HINT [76], and observe similar results (Fig. B.7). Furthermore, in order to show that potential bias in selection of baits in these high-throughput experiments does not affect our conclusions, we compare the number of bait-to-prey interactions only, as reported in these high-throughput experiments. In particular, we only compare multifunctional and non-multifunctional genes which are baits in these experiments, and observe the same trends (Fig. B.7). Overall, we conclude that multifunctional genes are more

centrally positioned in the protein interaction network, which may suggest an intermodular role in the interactome.

To further investigate the tendency for intermodularity of multifunctional genes, we integrate the protein interaction network with GO annotations in order to determine if multifunctional genes tend to interact with several proteins having distinct functions (i.e., be intermodular). For a GO term, we consider as a functional module the set of all genes annotated by this term. Recall that by definition, a gene is multifunctional if it is annotated with a pair of dissimilar GO terms of comparable specificity. Therefore for a multifunctional gene and a pair of dissimilar terms annotating it (as by our multifunctionality definition), we call this gene intermodular if it interacts with the two modules in the network formed by genes annotated by these two terms.

We observe that for fly, out of 1075 multifunctional genes in the protein-protein interaction network, 267 genes are detected as intermodular according to the above definition. When we repeat the computation in degree- and annotation-preserving random networks (see Section 3.4), only 4.3 ± 2.2 multifunctional genes are detected as intermodular, confirming that the actual number is highly significant. Similarly, in human, out of 2160 multifunctional genes in the network, 828 genes are detected as intermodular, while only 27.2 ± 5.2 are detected as such in random networks. In yeast, out of 833 multifunctional genes in the network, 519 genes are detected as intermodular, while only 21.8 ± 4.7 are called intermodular in random networks (Table B.3).

Genes with many interactions are more likely to interact with any functional modules regardless of whether these genes are multifunctional or non-multifunctional. Multifunctional genes tend to have many interactions (Fig. 3.8), and this could potentially be the only explanation for their tendency for intermodularity in the above

analysis. However, we repeat this analysis focusing only on genes with high degree, and still observe that multifunctional genes tend to be intermodular (Table B.3).

Overall, this analysis confirms the strong tendency for the intermodularity of multifunctional genes.

3.2.9 Multifunctionality with respect to molecular function

The main focus of our analysis so far has been on multifunctional genes detected using the Biological Process ontology (BP-multifunctional). However, the same procedure for detecting multifunctional genes can be applied to the Molecular Function ontology (MF-multifunctional) instead, providing an orthogonal view of gene multifunctionality.

We identify sets of MF-multifunctional genes for each organism and observe that MF-multifunctional genes have all the same distinct biological properties when compared with other annotated genes as has been reported in previous sections for the BP-multifunctional genes (although some p-values for yeast were above our significance threshold of 5%; see Fig. B.8, Fig. B.9, Fig. B.10).

In order to see if the involvement of a gene in multiple biological processes can be explained by multiple functions of the gene at the molecular level, we directly compare the two sets of multifunctional genes derived from the two ontologies. We observe that between 12% to 35% of BP-multifunctional genes are also MF-multifunctional, which constitutes a significant overlap ($p < 6e{-}18$, Table B.4), while the remainder may potentially be explained by other modes of gene multifunctionality. In contrast, a gene involved in multiple molecular functions might be expected to have these molecular functions while performing different biological processes, and indeed most MF-multifunctional genes are also BP-multifunctional (56% to 78%; Table B.5). Note, however, that the total number of MF annotations is lower than the total number of BP annotations (Tables B.4 and B.5), and thus the total number of genes identified

as MF-multifunctional is lower than the total number of genes identified as BP-multifunctional (Table B.5).

3.3 Discussion

Most proteins are—to a certain extent—multifunctional. Even within this context, previous experimental studies have identified proteins that perform remarkably different molecular functions [25, 26, 83, 84, 85], or that affect several distinct biological processes [87, 28, 86]. These findings suggest the existence of a subset of genes that are endowed with a particularly high degree of functional plasticity. Identifying such genes and studying their properties can help elucidate the functional organization of the cell. In this study, we have introduced a computational approach to systematically identify multifunctional genes using Gene Ontology annotations, and we have shown that they are characterized by distinct properties as compared to other genes. With respect to other studies, our approach specifically addresses some previous weaknesses in handling the Gene Ontology, such as the prior use of distinct GO terms that nonetheless convey closely related functions to define multifunctionality. Further, we have carried out inter-species comparisons, observing similar trends across three different organisms and thereby minimizing the effects of organism-specific annotation biases. Special care was also taken in gauging the effects of study bias, particularly in the case of interaction network properties and disease genes.

The main conclusion of our study is that gene multifunctionality is associated to several distinct properties, including a higher number of protein domains, a higher proportion of disordered regions in the protein sequence, more regulatory interactions, and a tendency to occupy more central and intermodular regions in the interactome. Determining which of these properties (or combinations of properties) represent the main mechanism underlying the functional plasticity of a gene is of great interest. It

is also possible to speculate that multifunctionality may be achieved via class-specific mechanisms; i.e., certain mechanisms may be at play only for a given class of genes.

Another important aspect that needs to be addressed in greater depth is the role played by context on protein function. In other words, what subset of functions are carried out by a gene in a given spatio-temporal context? Being able to tease apart the conditions under which a specific function is performed by a gene could lead to the development of a context-specific Gene Ontology vocabulary. In this ontology, the terms used to annotate genes could be qualified with other terms specifying the cell type, the developmental stage, or the stage in the cell-cycle in which a given function is most likely to be carried out by a gene.

In conclusion, a comprehensive understanding of gene and protein function has been a major goal of computational biology since the emergence of the field. In this work, we developed a computational method for genome-wide detection of multifunctional genes using existing functional annotations. This allowed us to make a number of novel observations about gene multifunctionality across several organisms, as well as to confirm some previous findings (including in some cases where only anecdotal evidence existed). Overall, our work contributes to a better systematic understanding of the functional landscape of the proteome, and can be the basis for future work in this direction as more specific and detailed functional genomics data become available.

3.4 Materials and Methods

3.4.1 Multifunctional genes

Gene Ontology (GO) [7] terms and gene association data for each organism were downloaded from

`http://www.geneontology.org/` on July 12, 2013. We excluded associations with evidence codes IEA ("Inferred from Electronic Annotation") , RCA ("Inferred from

Reviewed Computational Analysis"), IPI ("Inferred from Physical Interaction"), ND ("No Biological Data Available") or the qualifier NOT.

We call multifunctional every gene which is annotated with at least "two sufficiently distinct functional terms of comparable specificity", as explained next.

To define terms of about equal specificity, we start with the notion of informative terms used previously in the literature [111, 112, 113, 50], which selects for a given N all terms that annotate $\geq N$ genes, but whose descendants annotate $< N$ genes. However, a very general term annotating many genes may have all descendant terms annotating only small numbers of genes. In this case, with this definition of informative terms, a general term may be selected as informative for too small a value of the parameter N, for which all other informative terms selected are much more specific. For example, a fly term `imaginal disc-derived wing morphogenesis` (`GO:0007476`) annotates 508 genes, but its descendant terms annotate no more than 82 genes each (248 genes in total), and it may be undesirable to call this term informative for $N \approx 100$, as it is actually a much more general term than other terms which annotate approximately 100 genes. To overcome this problem, for a certain ontology (e.g., Biological Process) and for a certain value of the parameter N indicating specificity level, we select all terms which annotate $\geq N$ genes, but $< 2N$, and whose every descendant term annotates $< N$ genes.

For each N, for the terms at the specificity level N, we further select pairs of terms that are sufficiently distinct. Terms annotating similar sets of genes may correspond to similar functions, so first we filter out all pairs of terms which annotate significantly overlapping sets of genes (hypergeometric test, $p < 0.1$). Then we remove all pairs of terms with semantic similarity larger than zero [114]; in other words, we select only the pairs of terms for which their least common ancestor is the root of the ontology. Finally, we filter out pairs of terms that have a common descendant term, as this may

be an indication of similarity between the terms. We consider genes annotated by pairs of terms at the specificity level N selected by this procedure as multifunctional.

We call multifunctional all genes for which we find evidence of multifunctionality at a certain specificity level. We also would like to focus on more specific biological process terms and avoid considering less informative more general terms annotating a lot of genes. Namely, the final set of **multifunctional genes** is given by the union of all sets obtained for different N, where N ranges from 10 up to certain upper bound M, with increment of 10. We choose $M = 120$ in the main text. We compare multifunctional genes with all other genes that are annotated with terms at the specificity level N such that $10 \leq N \leq M$.

For GO analysis, we use code from the project goatools (`https://github.com/tanghaibao/goatools`).

3.4.2 Data for comparison of multifunctional and other genes

Physicochemical properties of genes. Proteomes of *D. melanogaster*, *H. sapiens*, and *S. cerevisiae* were downloaded from UniProt (September 2013 release). For proteins encoded by all genes, we computed their length and average disorder. For fly and human, we considered the longest protein isoform encoded by a gene (for yeast, there was only one isoform per gene in the database). Domain information was obtained from Pfam 27.0 [115], using the annotations contained in the swisspfam file. We note that multiple instances of the same domain in a sequence were explicitly ignored in the calculations of the number of domains. Prediction of disordered residues was carried out using the IUPred program [92, 93], with default parameters. The average fraction of disordered residues was then computed as the average fraction of residues with a IUPred score above 0.5.

Expression. *D. melanogaster*: FlyAtlas project data [94] was downloaded from GEO [58] (accession number GSE7763), and it consists of four replicates

per condition. A gene is considered present in a tissue if it is detected as present in all four replicates, as reported in the dataset. We also use RNA-seq data from modENCODE [95] processed by FlyBase [96] as described in `http://flybase.org/reports/FBrf0221009.html`, file `gene_rpkm_report_fb_2013_05.tsv.gz`. A gene with non-zero RPKM in a tissue was considered present in this tissue. *H. sapiens*: Expression data for human was downloaded using Ensembl BioMart, release 73 [97], using data sources "GNF/Atlas organism part" for GNF Atlas [116] and "Anatomical System (egenetics)" for eGenetics [117].

Evolutionary conservation. Evolutionary conservation scores from phastCons [99] were downloaded from the UCSC genome browser website on December 10, 2013, for *D. melanogaster*, *H. sapiens*, and *S. cerevisiae*, from `http://hgdownload.soe.ucsc.edu/goldenPath/dm3/phastCons15way/`, `http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons100way/hg19.100way.phastCons/`, and `http://hgdownload.soe.ucsc.edu/goldenPath/sacCer3/phastCons7way/`, respectively. Conservation scores were averaged over nucleotides of exons of each isoform and then averaged over isoforms. (For yeast, there was only one isoform per gene.) For comparison across orthologs, we use protein ortholog information from version 4 of the Princeton Protein Orthology Database (P-POD) [62] (`ftp://gen-ftp.princeton.edu/ppod/`). We consider two proteins from different organisms to be orthologous if they belong to the same family, as detected by P-POD using either OrthoMCL or MultiParanoid.

Regulatory interactions. Regulatory TF–gene interactions were obtained from DroID [78], version v2013_07 for *D. melanogaster*, from ENCODE [100] (file `enets1.Proximal_raw.txt` from `http://encodenets.gersteinlab.org/`) for

H. sapiens, and from YeastMine [118] (downloaded November 24, 2013, high-throughput interactions attributed to [101] or [102]) for *S. cerevisiae*.

Genetic interactions. Genetic interactions for fly were obtained from FlyBase [96], version v2013_07, and for yeast from BioGRID [5], version 3.2.102. For yeast, positive (evidence codes Positive Genetic, Synthetic Rescue) and negative (evidence codes Negative Genetic, Synthetic Growth Defect, Synthetic Lethality) genetic interactions were also considered separately.

Essentiality. Phenotype data was obtained for fly (FlyBase [96], version v2013_07), human (mouse ortholog phenotype data from MGI [104], downloaded October 3, 2013), and yeast (from YeastMine [118], downloaded September 26, 2013). Essential genes were defined as genes with "lethal" phenotype for fly, with any phenotype containing "lethal" in its name for human, and with "inviable" phenotype for yeast. When applying a hypergeometric test for enrichment of essential genes in multifunctional genes, the set of all genes with any reported phenotype was used as a background (only genes from this background set were considered). In addition, sets of essential genes detected in genome-wide RNAi screens in cell lines were obtained from OGEE [119] for fly [103] and human [105]. In addtion, a more detailed analysis reporting a score of essentiality for each gene in a genome-wide screen in each of 72 tested human cancer cell lines was obtained from COLT-Cancer [106, 107] (file `GARP-score.txt.tar.gz` downloaded from `http://dpsc.ccbr.utoronto.ca/cancer/download.html`). For yeast, we also used the data from genome-wide heterozygous and homozygous gene deletion screens across multiple conditions [108] from files `hom.z_tdist_pval_nm.pub` and `het.z_tdist_pval_nm.goodbatch.pub` downloaded from `http://chemogenomics.stanford.edu/supplements/global/download.html`, for each gene counting the number of conditions for the corresponding deletion strain with p-value below 0.01.

Disease data. We used BioMart [120] to obtain gene-disease associations from the Online Mendelian Inheritance in Man (OMIM) catalog [109]. Out of the 9,664 human genes with at least one specific GO term, 2,299 had at least one OMIM association. To further probe the similarity between diseases involving the same genes, we used Disease Ontology [110], a knowledge base of human disorders that are hierarchically organized in a directed acyclic graph. We mapped OMIM terms to Disease Ontology terms using the OBO file available at `http://disease-ontology.org/downloads`. Out of 2,299 genes with an OMIM association, 1,148 had at least one Disease Ontology term. We focused on these genes with at least one Disease Ontology term, and extracted from them all genes that had at least two Disease Ontology terms with only the root node in common; this resulted in 135 genes, which we considered as genes associated with at least two distinct diseases.

Physical protein-protein interactions. Physical protein-protein interactions were obtained from BioGRID [5], version 3.2.102. Proteins with more than 200 interactions were iteratively removed (i.e., the protein with the highest number of interactions removed one at a time), in order to avoid experimental artifacts due to "sticky" proteins. For extraction of high-throughput interactions, we considered only the interactions indicated as high-throughput in the database and only from publications contributing interaction data with at least 100 baits. For human and yeast, we also considered high quality high-throughput interaction datasets from HINT [76].

PubMed publications. The number of PubMed publication IDs associated with each gene was downloaded from NCBI at `http://www.ncbi.nlm.nih.gov/gene` on September 18, 2013.

3.4.3 Comparison across orthologs

For each pair of organisms, we count how many orthologous pairs of multifunctional genes are found where one gene in a pair is from one organism and the other gene in

the pair is from the other organism. To assess significance, we repeat the computation 1000 times randomly and independently re-assigning genes in each organism to the two gene classes of the same size as classes of multifunctional and non-multifunctional genes, but preserving the orthology relationship between genes of different organisms. In this randomization, only genes from orthologous pairs between the two organisms are considered. Then we compute the average and standard deviation of the counts in random trials and an empirical p-value of the real count with respect to the randomized counts.

3.4.4 Network analysis

The **degree** of a vertex is the number of interactions that the corresponding protein has in the network. The **betweenness centrality** of a vertex v is the number of shortest paths between all pairs of vertices in the network that pass through v, with the shortest paths between two genes s and t weighed inverse to the total number of distinct shortest paths between s and t. The **participation coefficient** [80, 57] of a vertex with respect to a set of clusters in a network is defined as $P = 1 - \sum_i \left(\frac{k_i}{k}\right)^2$, where the sumation is over all clusters, k is the vertex degree, and k_i is the number of edges going from the vertex to vertices from the cluster i. The rationale is to have $P = 0$ if all edges from the vertex go to a single cluster, and to have P closer to 1 if edges from the vertex are more uniformly distributed over clusters. To find clusters in the network, we used the SPICi clustering algorithm [22] with parameters optimized with a simple exhaustive search procedure to approximately maximize Newman's modularity [81], as described in Section 2.4. For network analysis, we use the python interface to the igraph library, version 0.6.5 (`http://igraph.sourceforge.net/`).

We integrate protein interaction networks with Gene Ontology to define intermodular multifunctional genes. We call a multifunctional gene g intermodular if there is a pair of dissimilar terms T_1 and T_2 annotating g (as detected for g according to the

definition of multifunctionality given above) such that g has interactions with at least one other gene annotated by T_1 and at least one other gene annotated by T_2. Here genes in the network annotated by T_1 and T_2 (including g for each of these terms) are considered as belonging to the corresponding functional modules. For assessing significance, we repeat the computation in 200 degree-preserving random networks (as implemented in method `igraph.Graph.Degree_Sequence()` with option "vl" [59]) while preserving GO annotations of all genes, and compute the average and standard deviation of the number of multifunctional intermodular genes in these random networks. We also repeat all the analysis focusing only on genes in top 20% or in top 5% in the network by the number of interactions.

Chapter 4

Conclusion

In this thesis, we developed computational approaches for systematically analyzing large-scale genomic and proteomic data in order to gain new knowledge about the functioning of the cell.

In Chapter 2, we studied how simple properties of hub proteins are predictive of their roles in the functional organization of cellular networks. For this, we leveraged functional genomic data for five organisms, *S. cerevisiae*, *H. sapiens*, *D. melanogaster*, *A. thaliana*, and *E. coli*. We showed that simple features of hubs in the network reveal important aspects of the dynamics and modularity of the interactome. We showed that this holds not only for the feature of average co-expression previously studied in this respect [24], but also for other features that depend purely upon the topology of the network, such as betweenness centrality, clustering coefficient, and participation coefficient. We found that these features reflect intra- and inter-modularity of proteins in the network. Working with data for several different organisms allowed us to perform a cross-interactomic analysis. We showed that inter- and intra-modularity, as measured by these simple hub features, is conserved across organisms.

In Chapter 3, we studied the role of multifunctional genes and the proteins that they encode in the functional organization of the cell. For this, we used functional

annotations for three organisms, *S. cerevisiae*, *H. sapiens*, and *D. melanogaster*. We proposed a robust method to detect multifunctional genes, and distinguished them from genes more likely to have a single function. We performed an analysis of multifunctional genes with respect to a number of different biological properties, and showed that, as compared to other genes, multifunctional genes are longer, are more disordered, are more broadly expressed, are more intermodular in protein interaction networks, are regulated by larger number of transcription factors, tend to be more evolutionarily conserved and are more likely to be essential. We also found that mutations in multifunctional genes are significantly more likely to be associated with human disorders.

In sum, our observations provide a better understanding of the functional organization of the cell. As more specific, high-quality and high-coverage genome-wide and proteome-wide data become available, we believe our approaches to analyze these data can reveal further details about the structure, function, and evolution of the cell.

Appendix A

Supplementary information for Chapter 2

A.1 Supplementary results

A.1.1 Hub classification analysis

Our classification of hubs into party, date and extremal for different networks (Fig. A.1–A.6) yields results qualitatively similar to those reported for the **Human-hq** network (Fig. 2.1).

The results of classification of all hubs are qualitatively the same as the results of classification with extremal hubs excluded (Fig. A.7, compare with Fig. 2.1).

A.1.2 Analysis for a relaxed definition of hubs

For the three largest networks, we also considered a more relaxed definition of hubs, where all genes with degree ≥ 3 are considered as hubs, instead of just the top 10%. This results in 6762 hubs in **Human-all** (66.1% of all vertices), 4716 (83.6%) hubs in **Yeast-all** and 4992 (60.7%) hubs in **Fly**. Results of our hub classification analyses are largely the same as for hubs defined with a more selective definition (Fig. A.8–

A.10). Furthermore, the results of correlation analysis of hub characteristics stay largely the same as with a more selective definition (Fig. A.11).

We do however observe a higher betweenness for party hubs in **Yeast-all** when considering as hubs all genes of degree ≥ 3 (Fig. A.9), which is the opposite trend of when a higher hub threshold is used. This may be explained by the correlation of degree with avPCC in this case, and the typically observed correlation between degree and betweenness. Indeed, the SRCC between degree and avPCC is 0.32 ($p < 1e{-}110$), the SRCC between degree and betweenness is 0.81 ($p = 0$), and the SRCC between avPCC and betweenness is 0.21 ($p < 9e{-}46$). However, the partial SRCC of avPCC and betweenness corrected for degree is -0.10 ($p < 3e{-}11$), which is consistent with all previous observations (Fig. A.12).

A.1.3 Potential biases and confounding factors in the correlation analysis of hub characteristics

The number of interactions of a protein in the network could be significantly correlated with avPCC and other topological measures, and this may be a confounding factor in the analysis [24]. Sometimes we indeed observe a correlation (Fig. A.13A). To control for this, we calculate the Spearman partial correlation with a correction for degree. High correlations of hub characteristics remain significant (see Fig. A.13B and compare with Fig. 2.2).

In order to show that hubs with extremal properties do not bias the analysis of correlations between hub features, we perform the same analysis, but with extremal hubs excluded, and observe very similar results (see Fig. A.14 and compare with Fig. 2.2).

A bias towards more studied genes could also be responsible for some of the observed correlations [24]. In order to avoid that, we also perform the correlation

analysis on high-throughput networks for yeast and human and observe the same trends as for our main networks (see Fig. A.15, compare with Fig. 2.2).

A.1.4 GO annotations of hubs

We performed GO enrichment analysis for date and party hubs, as well as for classes of hubs specified by other hub characteristics. These results are shown in Fig. 2.3 and in Fig. A.16–A.21. See Section 2.4 for details.

For the **Fly**, **Athal** and **Ecoli** networks, we observe results that are in general similar to those for the human and yeast networks, though fewer terms are enriched. A possible explanation for the fewer number of enriched terms may be that hubs in these networks have fewer annotations than hubs in the networks of yeast and human. We show in Table A.11 the fraction of hubs annotated with terms other than the root in each ontology, and these numbers are considerably smaller for **Fly**, **Athal** and **Ecoli** than for the yeast and human networks.

A.1.5 Correction for essentiality when studying the number of genetic interactions of genes

Non-essential genes participate in a significantly larger number of genetic interactions than essential genes, or in other words, essentiality is correlated with the number of genetic interactions (Fig. A.24A). However, even after removing all essential genes from consideration, the numbers of genetic interactions date and party hubs are involved in are still significantly different (Fig. A.22CD). The partial correlation of avPCC and the number of genetic interactions with correction for essentiality is almost as high as without correction (see Fig. A.24B and compare with Fig. 2.4).

A.1.6 Yeast two-hybrid and co-complex interaction networks

The date and party hub analysis on networks of only yeast two-hybrid or only co-complex interactions for *H. sapiens*, *S. cerevisiae*, and *A. thaliana* (Fig. A.26–A.31) confirms that the date/party distinction is observable in these networks as well, though it is not as stringent for yeast two-hybrid as it is for co-complex networks.

A.1.7 Comparison of network topology properties for orthologs between organisms

We compute the Spearman correlation of various hub characteristics across networks. The results for networks **Human-all** and **Yeast-all** are shown in Table 2.2, the results for networks **Human-hq** and **Yeast-hq** are shown in Table A.5, and the results for the other networks are shown in Tables A.6–A.9 (organized per hub feature, rather than per a pair of networks). We observe that clustering coefficient, as well as betweenness centrality and participation coefficient, are highly correlated for networks of different organisms; this suggests that the placement and role of proteins within networks tend to be conserved and are biologically meaningful properties. Surprisingly, we do not observe the degree in the network, which is simply the number of physical interactions, to correlate in most cases: the only significant correlations were $\rho = 0.23$ ($p < 0.01$; empirical $p = 0.006$) for **Yeast-all** and **Athal**, $\rho = 0.14$ ($p < 0.003$; empirical $p = 0.002$) for **Yeast-all** and **Human-all**; this may be due to which proteins are studied more extensively in different networks.

A.1.8 Correction for signal from random networks for genetic interactions and essentiality

For correlations between hub characteristics (Fig. 2.2), we compared real correlations with those observed in random networks (as reported in Secton 3.2). We also perform

the same analysis for correlations between hub characteristics in yeast and the number of genetic interactions. That is, for the yeast networks, we compute the average correlation of the number of genetic interactions with avPCC, clustering, betweenness, participation and functional similarity in 100 random networks generated to preserve the number of physical interactions for each gene (see Fig. A.34A and compare with Fig. 2.4 which shows the same bars for real networks). In all cases, correlations in random networks are smaller by absolute value than significant correlations in real networks.

We noticed, however, a surprisingly strong significant negative correlation of avPCC and genetic degree in random counterparts of the network **Yeast-all**. We noticed no such relationship when the network was randomized and the degree distribution was not preserved (data not shown). We hypothesized that if the degrees of genes in random networks are restricted to be the same as in the real physical interaction network, certain properties of the resulting randomized networks may preserve structures and correlations in ways that are not entirely understood. For example, we observed that hubs have a preference to interact with the same genes they interact with in real networks when performing degree-preserving randomizations. This may result in a positive correlation between avPCC in a random network and avPCC in the real network, that leads to correlations between avPCC in the random network and other traits (such as genetic interaction degree). To correct for the behavior of avPCC in random networks, we generate another 100 random networks (separately from those used for the plots) and compute for each hub the average of avPCC scores in these networks. We denote the resulting hub score as avPCC-rand. We confirm a positive Spearman correlation between avPCC in the real network and avPCC-rand (0.60 in **Yeast-hq** and 0.74 in **Yeast-all**), though there is a large difference in the magnitude of avPCC and avPCC-rand (mean of avPCC 0.14 vs mean of avPCC-rand 0.03 over all hubs in **Yeast-hq**, and mean of avPCC 0.16 vs mean of avPCC-rand

0.04 over all hubs in **Yeast-all**). Then, we compute partial Spearman correlations of the genetic degree and avPCC, clustering, betweenness, participation, and functional similarity corrected for avPCC-rand, and the same values computed in random networks (Fig. A.34B, compare with Fig. A.34A). The correlation between avPCC and the genetic degree is significant even after this correction, and is close to zero in random networks. Note again that random networks used for plots are different from those used to calculate avPCC-rand.

We also perform the same analysis for essentiality and obtain similar results (Fig. A.35); that is, there is a significant correlation of essentiality and other hub features including avPCC even after correction for avPCC from random networks.

A.2 Supplementary materials and methods

A.2.1 Gene IDs

The following gene names were used as identifiers in networks and expression datasets.

S. cerevisiae: locus names such as YDL229W or YLR438C-A.

H. sapiens: Ensembl gene ids such as ENSG00000008988 or ENSG00000141510.

D. melanogaster: locus names such as CG14228 or CG9986.

A. thaliana: locus names such as AT1G66410 or AT5G42190.

E. coli: locus names such as B0015 or B4142.

All other gene identifiers were mapped to these using files from Saccharomyces Genome Database (SGD)[1], Profiling of Escherichia coli chromosome (PEC) database[2], EcoCyc project[3], Arabidopsis Information Resource (TAIR)[4], Database of Drosophila

[1] `SGD_features.tab` from `http://www.yeastgenome.org/`
[2] `PECData.dat` from `http://www.shigen.nig.ac.jp/ecoli/pec/`
[3] `gene-links.dat` from `http://ecocyc.org/ecocyc/index.shtml`
[4] `gene_aliases.20101027` from `http://www.arabidopsis.org/`

Genes & Genomes (FlyBase)[5], Drosophila Interactions Database (DroID)[6], and gene mapping files downloaded using BioMart MartView interface[7] for different organisms.

A.2.2 Interactions

The following interaction networks for five organisms are considered. In all networks, self-loops and duplicate interactions were deleted. The size of each network is shown in Table 2.1.

S. cerevisiae: Based on evidence types from BioGRID, interactions in **Yeast-all** were annotated as 'yeast two-hybrid' (7810 in **Yeast-all**) and 'co-complex' (44610 in **Yeast-all**), see Table A.12. Annotations of genetic interactions were taken from BioGRID evidence types: `Negative Genetic`, `Synthetic Growth Defect`, `Synthetic Haploinsufficiency`, `Synthetic Lethality` for negative (96142 interactions in total) and `Positive Genetic`, `Synthetic Rescue` for positive (20068 interactions).

H. sapiens: Based on evidence types of interactions from [77], in the network **Human-all** 14633 interactions were annotated as 'yeast two-hybrid' and 50390 were annotated as 'co-complex', see Table A.12.

D. melanogaster: The network of physical protein-protein interactions **Fly** was obtained by combining all interactions from DroID [78] version 2011_02 (25948 interactions, annotated 'yeast two-hybrid'), and from DPiM [38] (10623 coAP-MS interactions reported as high-quality in the publication, annotated 'co-complex').

A. thaliana: The network of protein-protein interactions **Athal** was formed from datasets downloaded from IntAct [9] and BioGRID, as well as from the recent publication [39]. First, 4707 interactions were obtained from BioGRID represented by 881 publications, then from IntAct 2620 interactions were obtained from those 272 publications (out of total of 603) that were not present in BioGRID, in order to avoid

[5] `gene_map_table_fb_2011_06.tsv.gz` and `fbgn_annotation_ID_fb_2011_06.tsv.gz` from `http://flybase.org/`

[6] `FLY_GENE_ATTR.txt` from `http://www.droidb.org/`

[7] `http://www.biomart.org/biomart/martview`

duplicate representation of interactions from the same publications with different gene ids. These interactions were annotated as 'yeast two-hybrid' (3086 interactions) and 'co-complex' (3148 interactions) based on evidence types provided by BioGRID and IntAct. All 6045 non-redundant interactions from [39], AI-1 dataset, were annotated as 'yeast two-hybrid', see Table A.12.

E. coli: The network of physical protein-protein interactions **Ecoli** was collected from different databases via PSICQUIC View application [79] using query `(taxidA:83333 AND taxidB:83333) AND (type:physical OR detmethod:(biophysical OR biochemical OR "two hybrid" OR affinity OR "pull down"))`. This network consists mostly of co-complex data.

A.2.3 Expression datasets

The expression compendia for the five organisms are as follows:

H. sapiens: the GNF Atlas project data over 79 cell or tissue types [116] (downloaded from GEO, accession number GDS596) is used as the source of expression data for human.

S. cerevisiae: In the GEO [58] database, aiming to construct an unbiased representative expression compendium and following the approach of Han *et al.* [24], we searched for keywords "stimulus response OR stress response OR cell cycle" while limiting the search to Series data (GSE) having from 20 to 100 datapoints (upper limit to avoid bias from large datasets), and publication date from 2006/07 to the 2011/07 (corresponding to the five years directly prior to when we gathered this data). We took only genome-wide datasets that used only *S. cerevisiae* in microarray experiments, and only those that after merging replicates would provide at least 10 datapoints. This resulted in a compendium of 20 datasets with the total of 540 expression datapoints (see Table A.13).

D. melanogaster: An expression compendium was formed from a collection of GEO datasets as was done for yeast (see above). Genome-wide RNA-seq data from the modENCODE project [121], as analyzed and published by FlyBase [122, 123], was added as well. This resulted in the compendium of 9 datasets with a total of 199 datapoints, from different types of cells including embryonic and various adult fly tissues, and under different conditions including development and stress response (see Table A.14).

A. thaliana: A compendium consisting of development data [124] (79 datapoints, from various tissues) and stress response data [125] (149 datapoints, from cells from roots and shoots, as well as from cell cultures) from AtGenExpress project was formed. Expression datasets were downloaded from the web page of the project[8].

E. coli: An expression compendium of 362 datapoints was formed from two smaller ones: a dataset consisting of 240 datapoints from different conditions with several timepoints for each was obtained from [126] as a log ratio data file, and the dataset of 122 datapoints corresponding to different conditions [127] was obtained from GEO, accession number GSE6836.

A.2.4 Clustering the network for computing participation coefficient

In order to compute the participation coefficient for hubs in a protein-protein interaction network, we first had to find clusters in the network. For this, we used the SPICi clustering algorithm [22] with parameters optimized with a simple exhaustive search procedure to approximately maximize Newman's modularity [81].

Namely, SPICi has two main parameters: the minimum density threshold parameter d and the minimum increment ratio r. We run SPICi many times with different parameters, and optimize for the resulting value of modularity. At the first stage we

[8] `http://www.weigelworld.org/resources/microarray/AtGenExpress/`

run the algorithm with parameters $d = 0.2, 0.4, 0.6, 0.8, 1.0$ and $r = d$ and select the preliminary value of $d = d_0$ that produces the maximum modularity. Then we do a binary search for the optimal value of d in the segment $[d_0 - 0.14, d_0 + 0.14]$ with granularity $1/2^{15}$, and for each hypothetical value of d, we optimize r in the segment $[0, d]$ with a step $d/15$.

This method produces, for example, parameters $d = r = 0.09487$ resulting in a Newman's modularity measure of 0.573881 for the network **Human-hq**, or parameters $d = r = 0.20215$ resulting in a modularity measure of 0.253280 for **Yeast-all**.

A.3 Supplementary figures

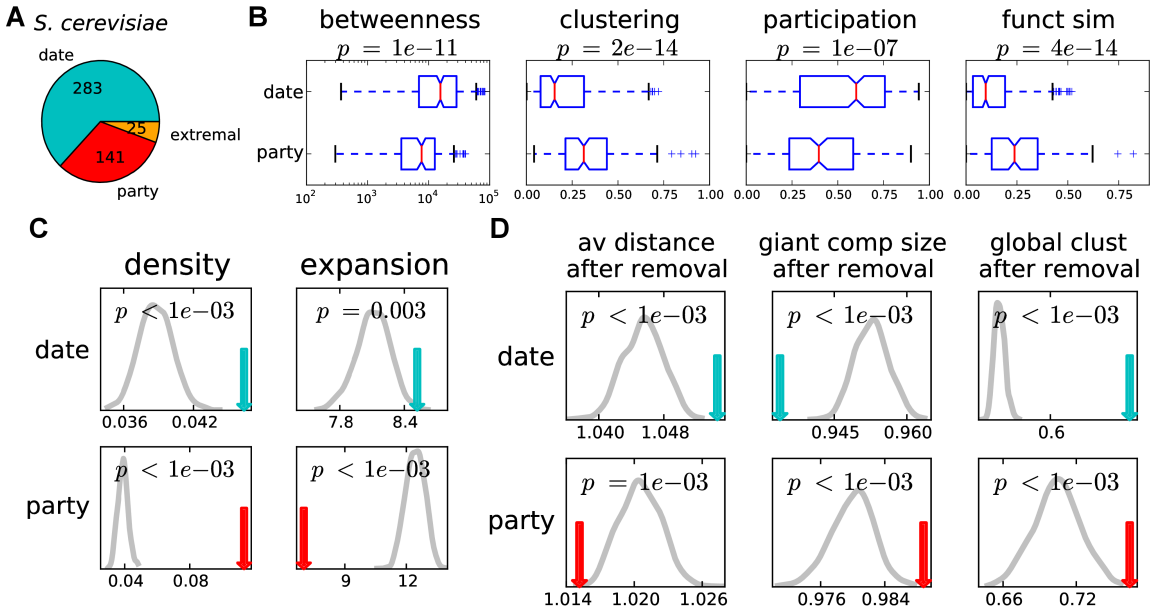

Figure A.1: Date and party hub classification analysis in yeast high quality network (**Yeast-hq**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.14; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

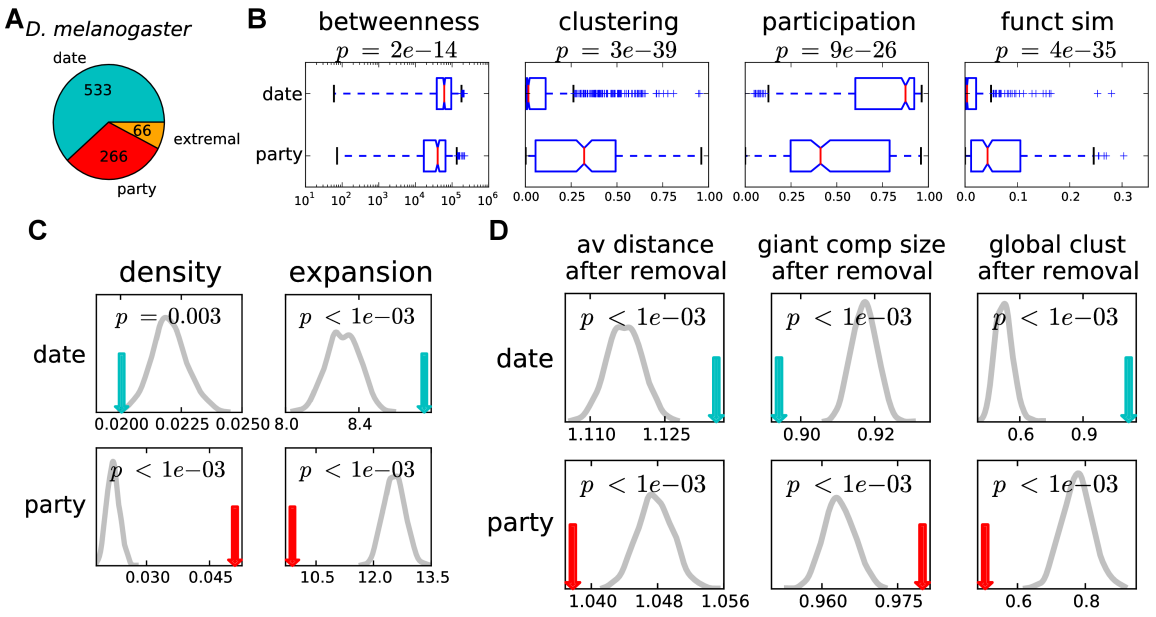

Figure A.2: Date and party hub classification analysis in fly network of all physical interactions (**Fly**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.12; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

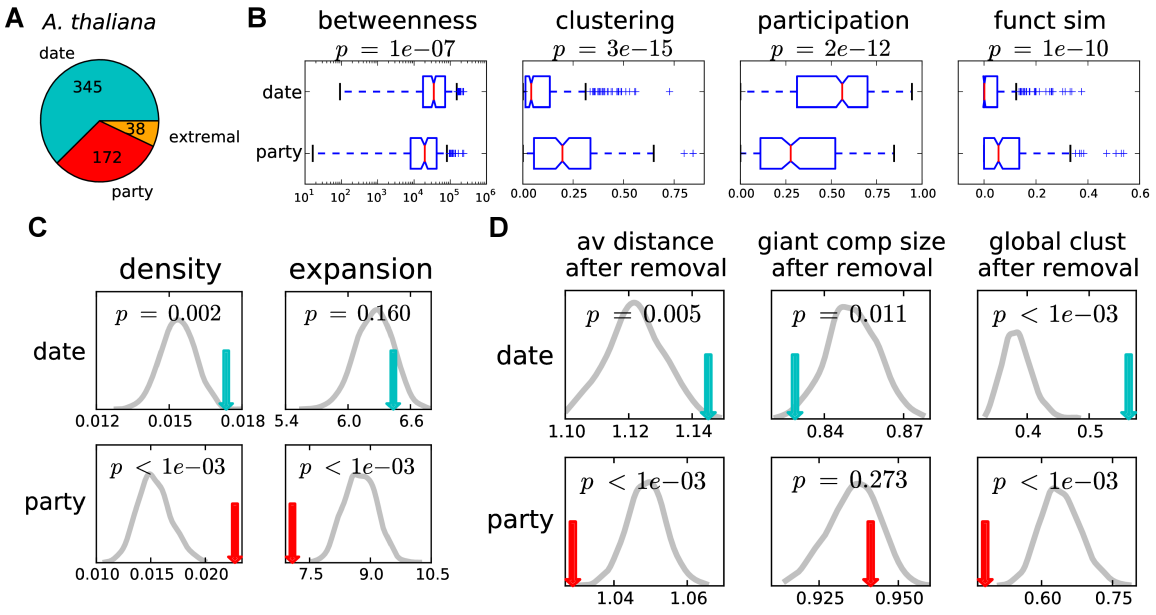

Figure A.3: Date and party hub classification analysis in Arabidopsis network (**Athal**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.15; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

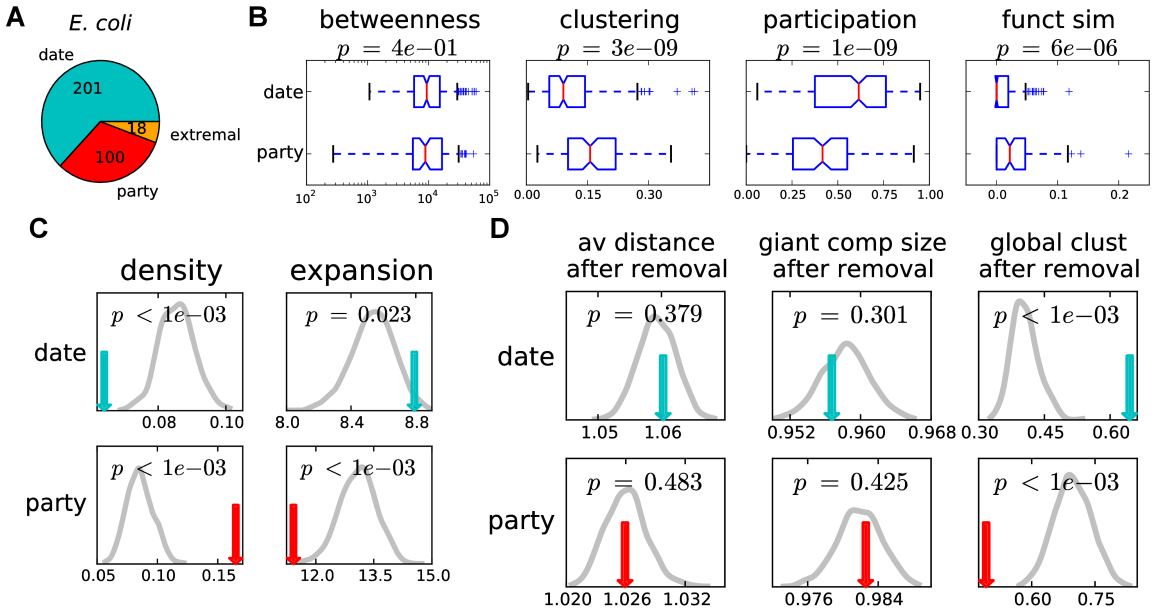

Figure A.4: Date and party hub classification analysis in *E. coli* network (**Ecoli**). (A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.13; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

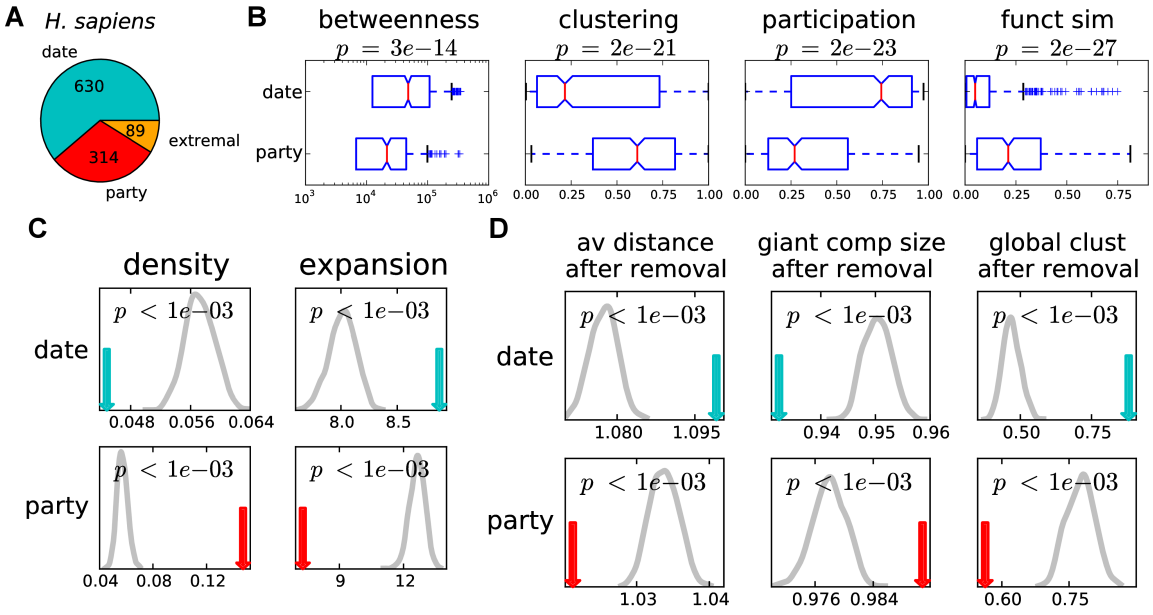

Figure A.5: Date and party hub classification analysis in human network of all physical interactions (**Human-all**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.24; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

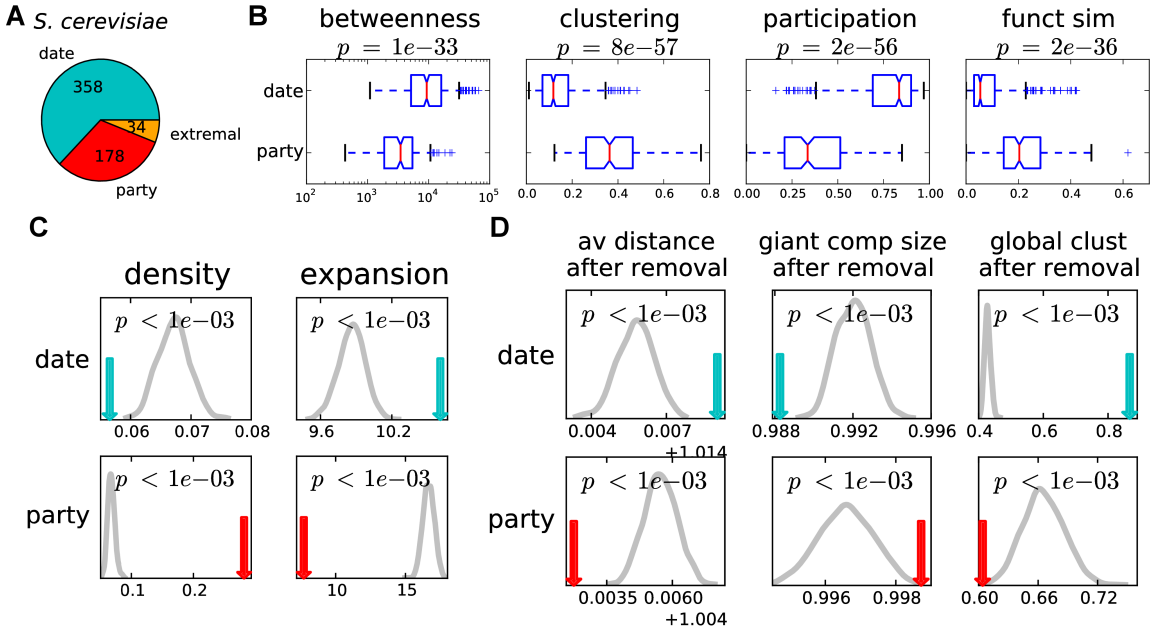

Figure A.6: Date and party hub classification analysis in yeast network of all physical interactions (**Yeast-all**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.21; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

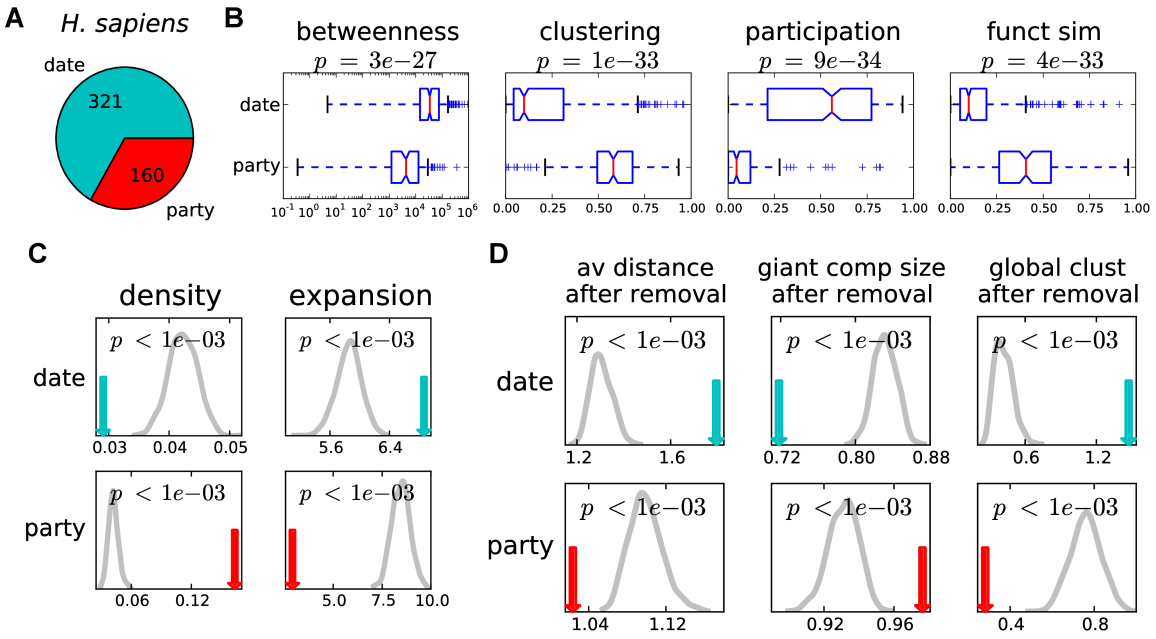

Figure A.7: Date and party hub classification analysis in human high quality network (**Human-hq**) with extremal hubs included.
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.30; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

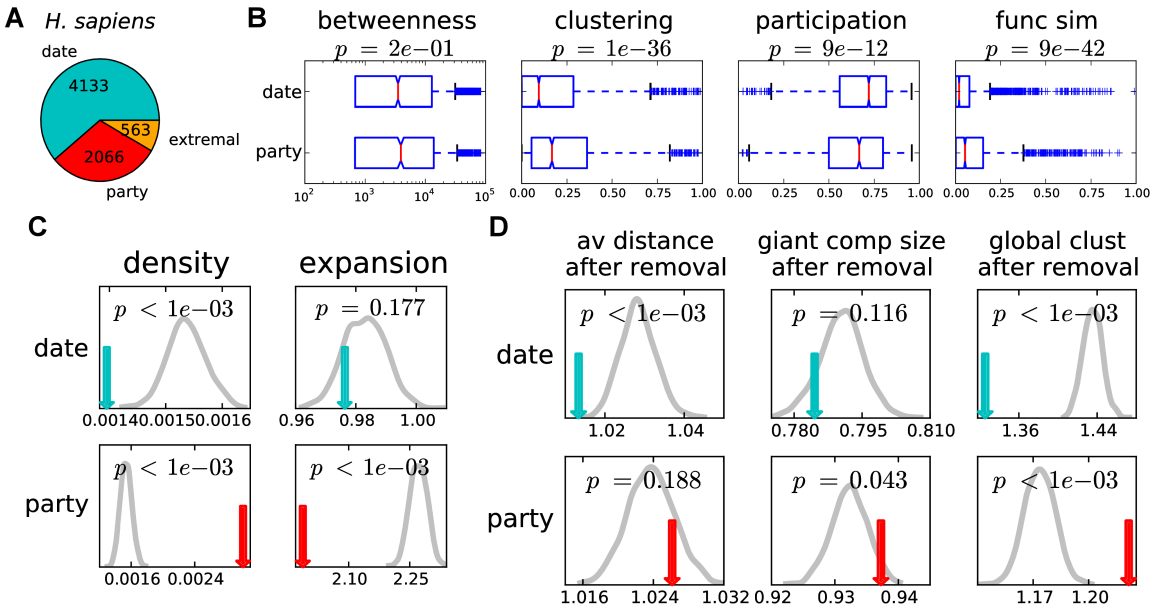

Figure A.8: Date and party hub classification analysis in human network of all physical interactions (**Human-all**), with all genes of degree ≥ 3 as hubs.
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.12; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

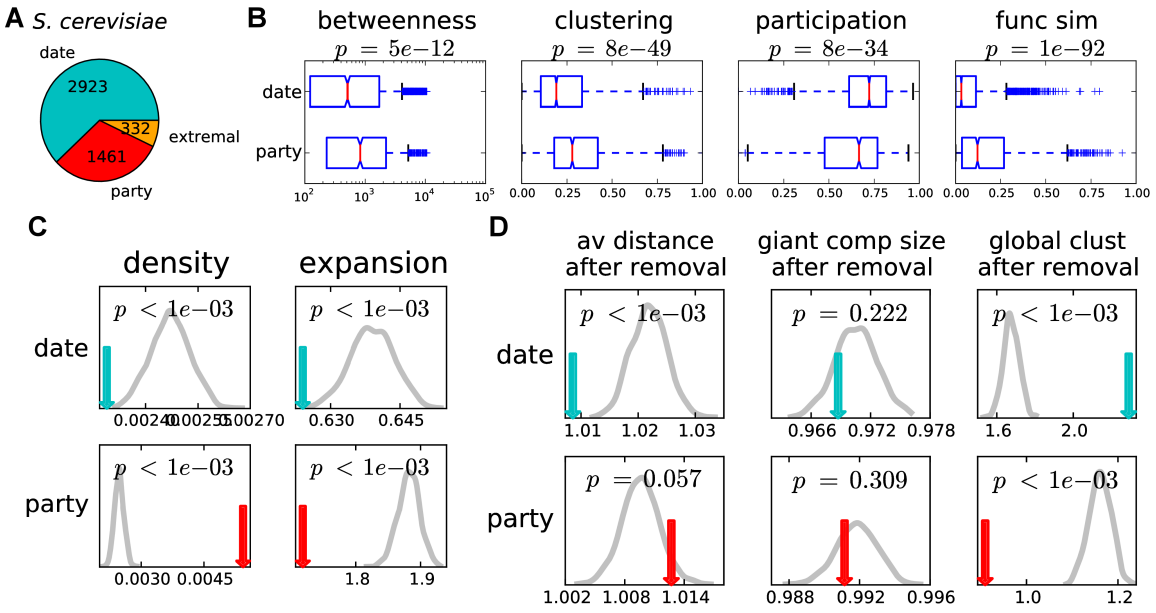

Figure A.9: Date and party hub classification analysis in yeast network of all physical interactions (**Yeast-all**), with all genes of degree ≥ 3 as hubs.
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.08; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

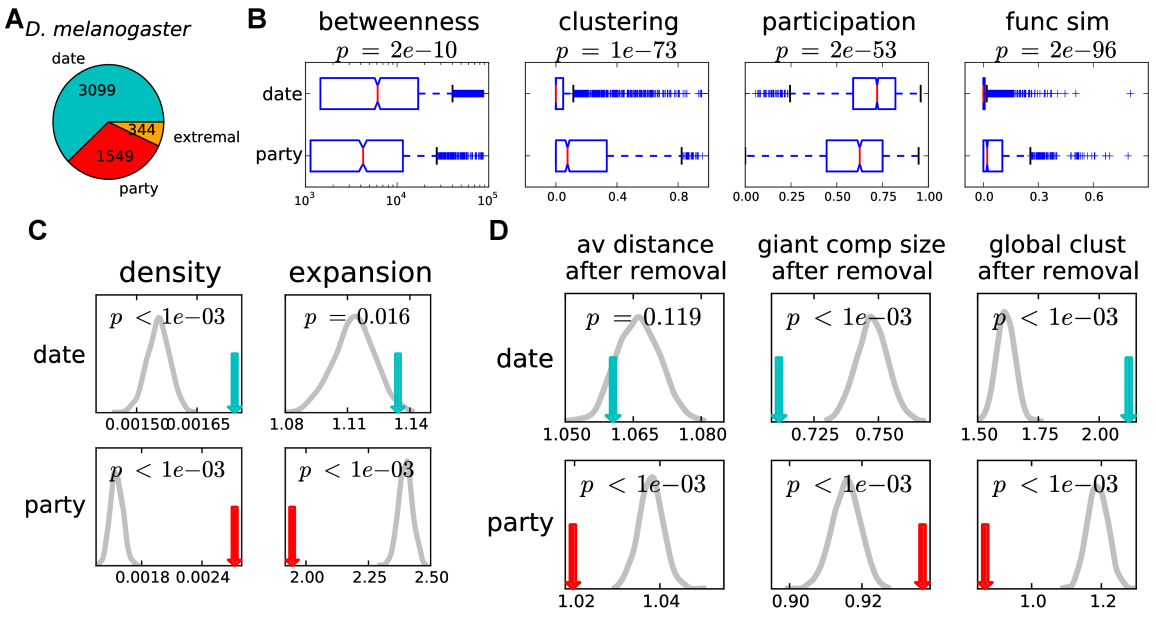

Figure A.10: Date and party hub classification analysis in fly network of all physical interactions (**Fly**), with all genes of degree ≥ 3 as hubs.
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.12; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

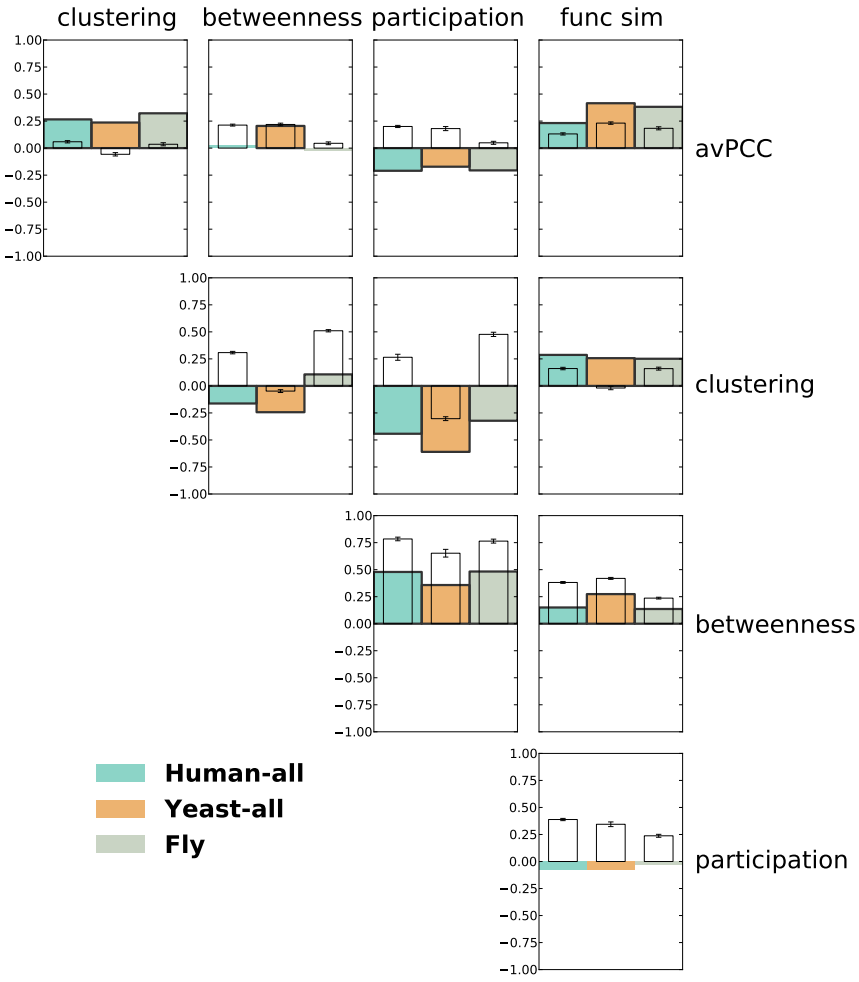

Figure A.11: Spearman correlation of hub characteristics in interaction networks, with all genes of degree ≥ 3 as hubs.
Every bar represents a Spearman correlation between two characteristics of hubs in one of the networks. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges. Smaller uncolored bars show average correlation (with error bars depicting the standard deviations) in 20 random networks on the same genes with the same number of interactions for each.

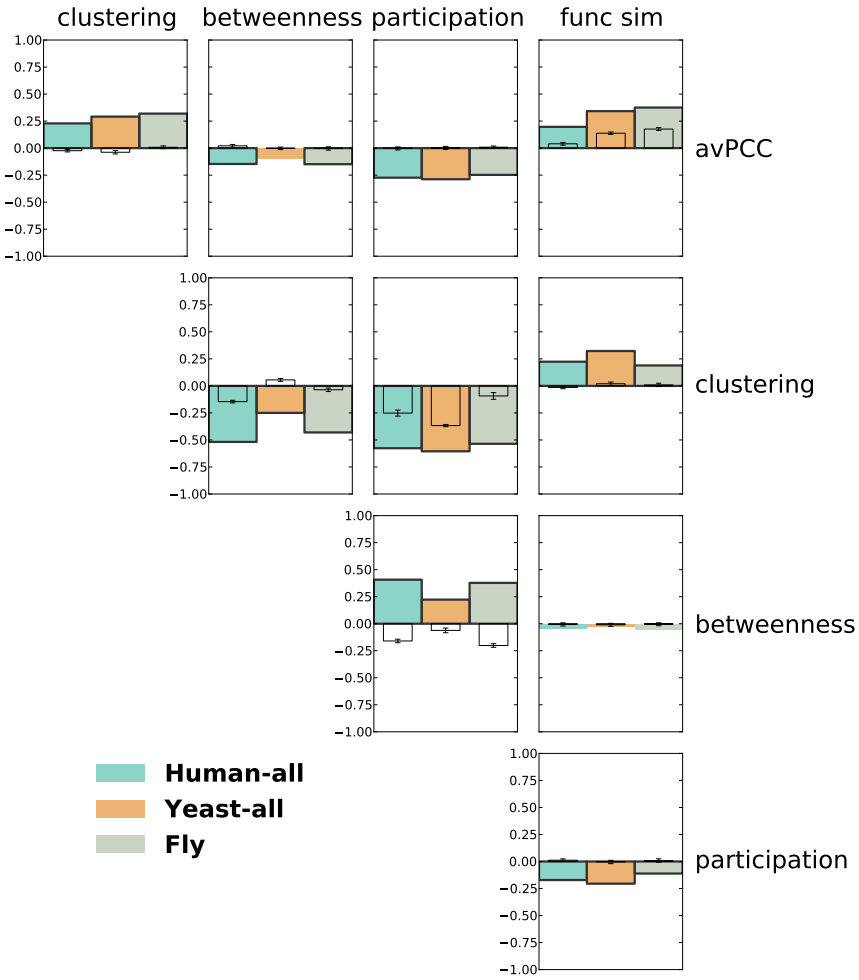

Figure A.12: Spearman correlation of hub characteristics in interaction networks, with all genes of degree ≥ 3 as hubs and with correction for degree.
Every bar represents a partial Spearman correlation corrected for degree between two characteristics of hubs in one of the networks. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges. Smaller uncolored bars show average correlation (with error bars depicting the standard deviations) in 20 random networks on the same genes with the same number of interactions for each.

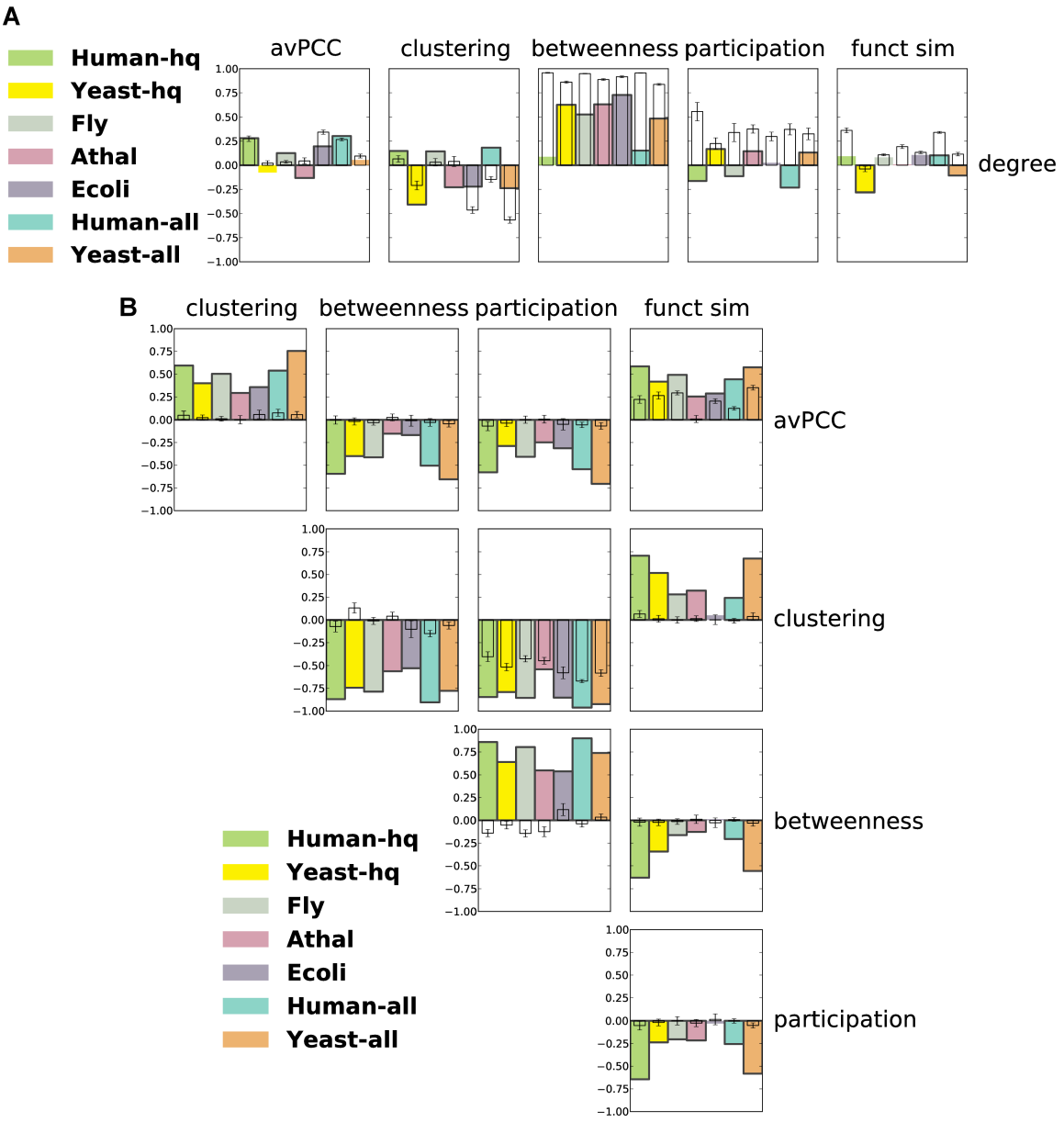

Figure A.13: Correlation with degree is not a confounding factor in the correlation analysis of hub characteristics.
(A) Every bar represents a Spearman correlation between a hub characteristic and degree (the number of interactions) for hubs in one of the networks. (B) Every bar represents a partial Spearman correlation corrected for degree between two characteristics of hubs in one of the networks. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges. Smaller uncolored bars show average correlation (with error bars for standard deviations) in 20 random networks on the same genes with the same number of interactions for each.

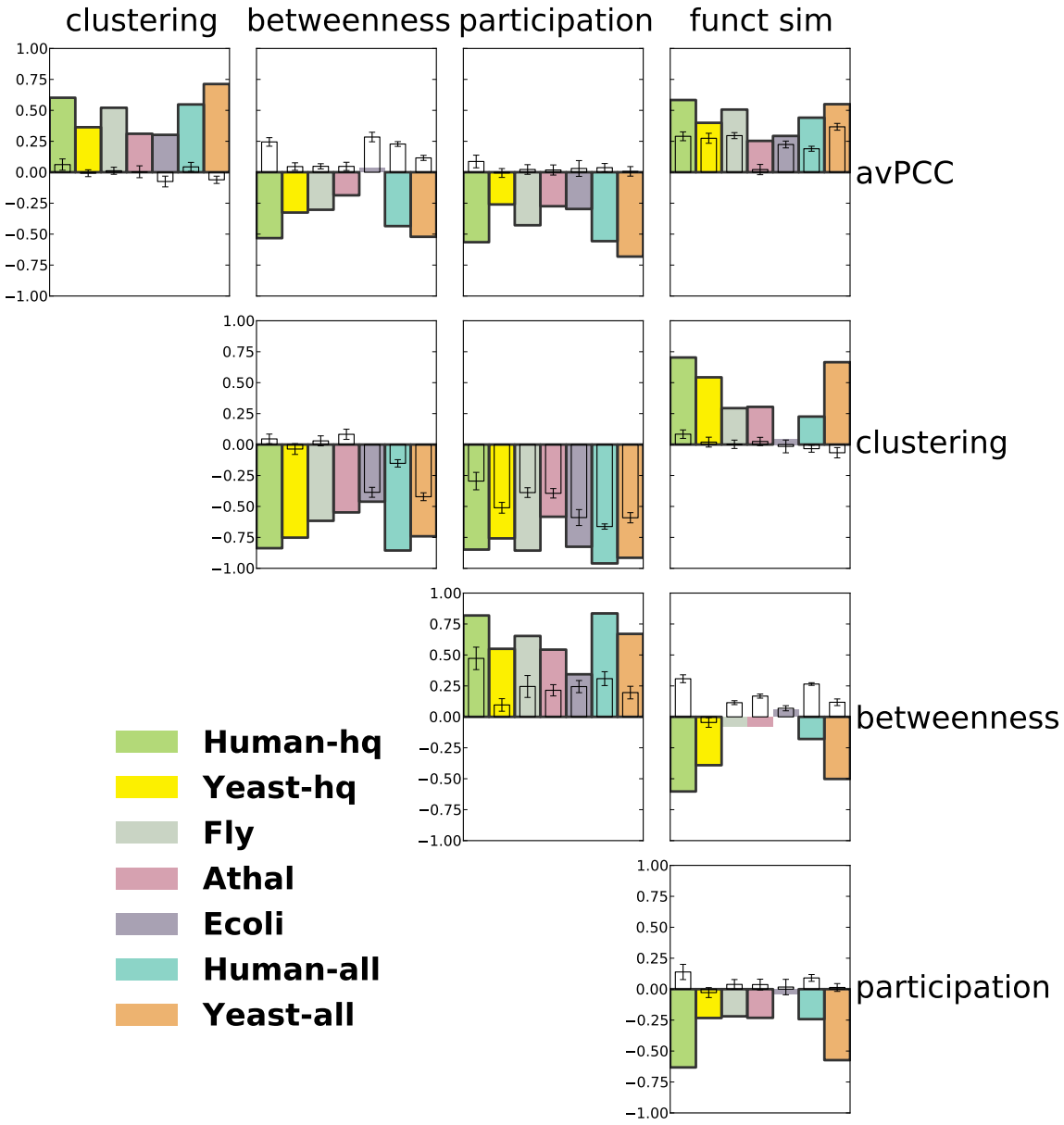

Figure A.14: Hubs with extremal properties do not bias the correlation analysis of hub characteristics.

Every bar represents a Spearman correlation between two characteristics of non-extremal hubs in one of the networks. Bars of significant correlations (absolute value $>$ 0.1, p-value $<$ 0.05) have black edges. Smaller uncolored bars show average correlation (with error bars for standard deviations) in 20 random networks on the same genes with the same number of interactions for each.

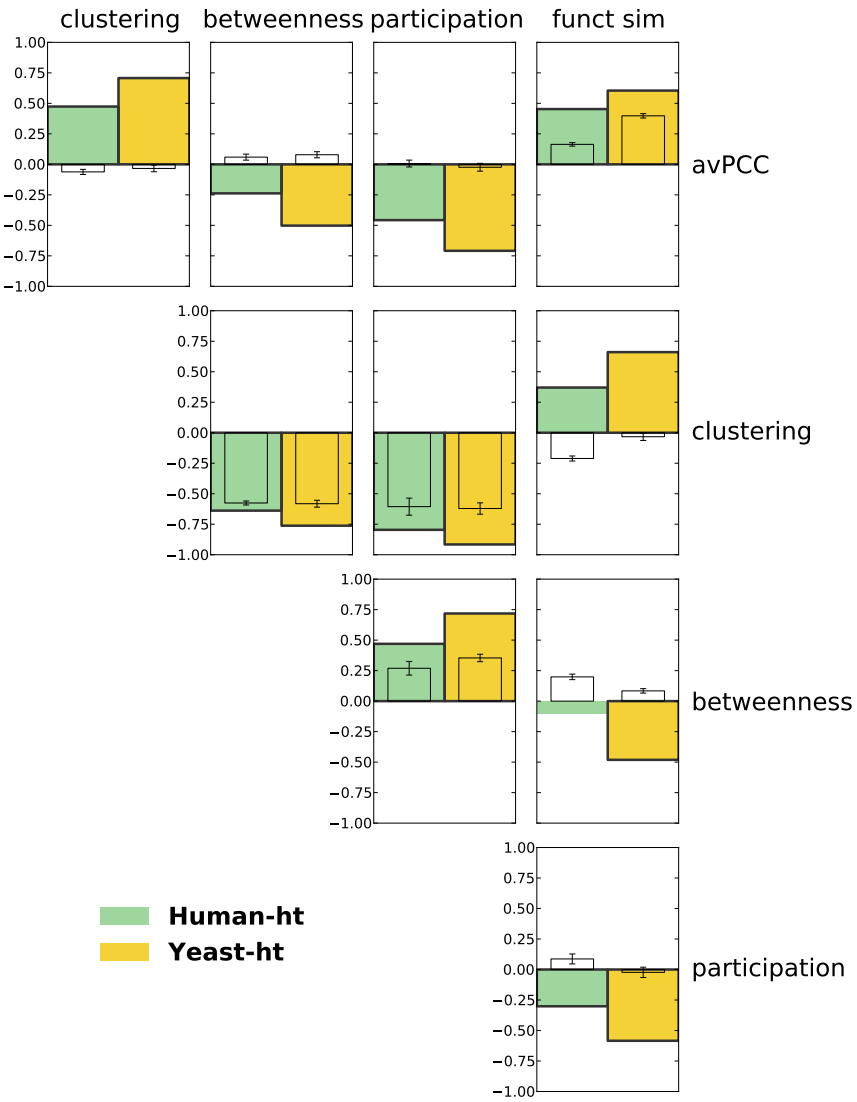

Figure A.15: Spearman correlation of hub characteristics in high-throughput interaction networks for human and yeast.
Every bar represents a Spearman correlation between two characteristics of hubs in one of the networks. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges. Smaller uncolored bars show average correlation (with error bars for standard deviations) in 20 random networks on the same genes with the same number of interactions for each.

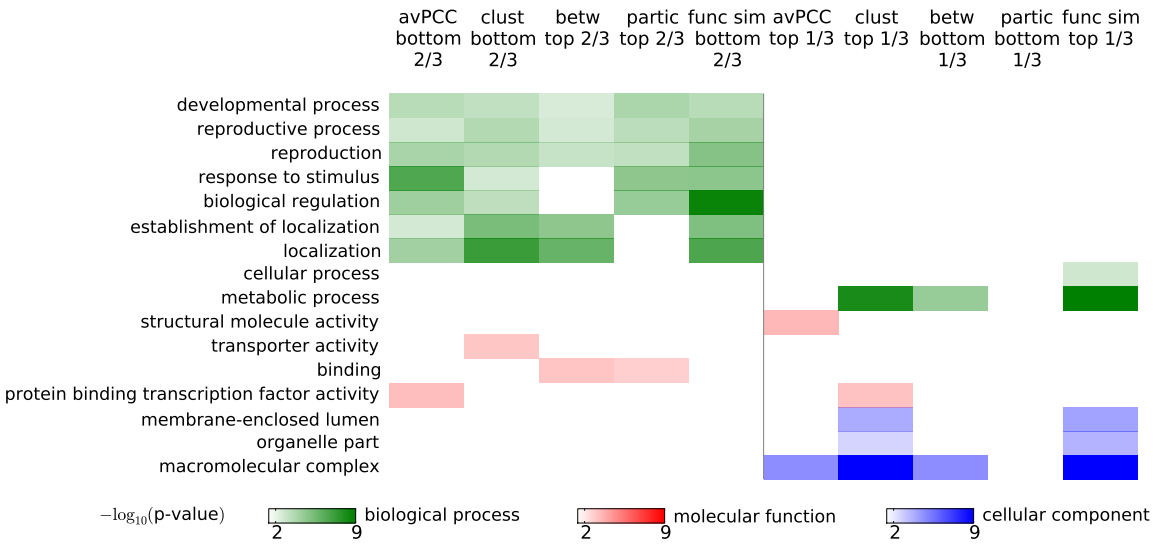

Figure A.16: GO annotation enrichment analysis of hubs in **Yeast-hq**. GO annotation enrichment analysis of hubs divided in a 2-to-1 proportion by avPCC, clustering, betweenness, participation and functional similarity scores in **Yeast-hq**. See Fig. 2.3 for details.

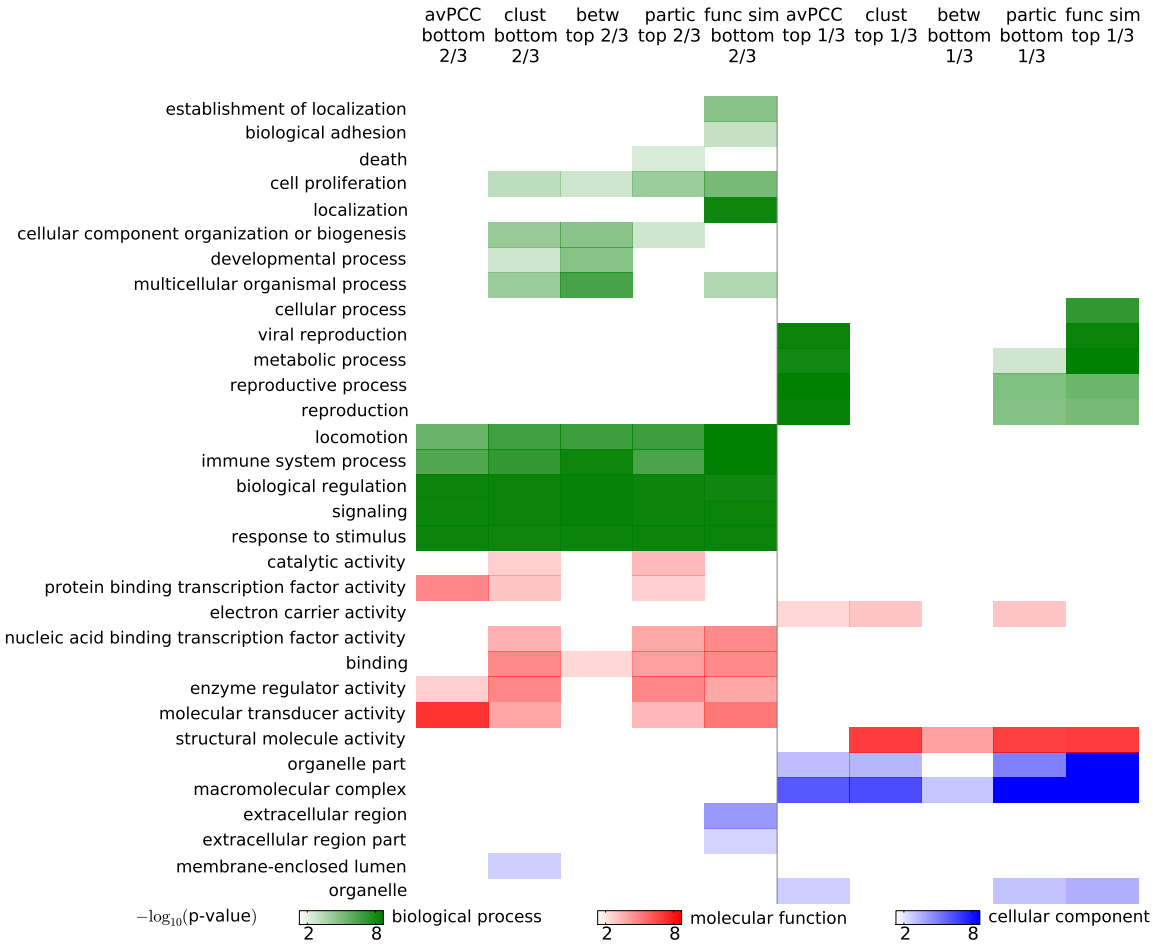

Figure A.17: GO annotation enrichment analysis of hubs in **Human-all**. GO annotation enrichment analysis of hubs divided in a 2-to-1 proportion by avPCC, clustering, betweenness, participation and functional similarity scores in **Human-all**. See Fig. 2.3 for details.

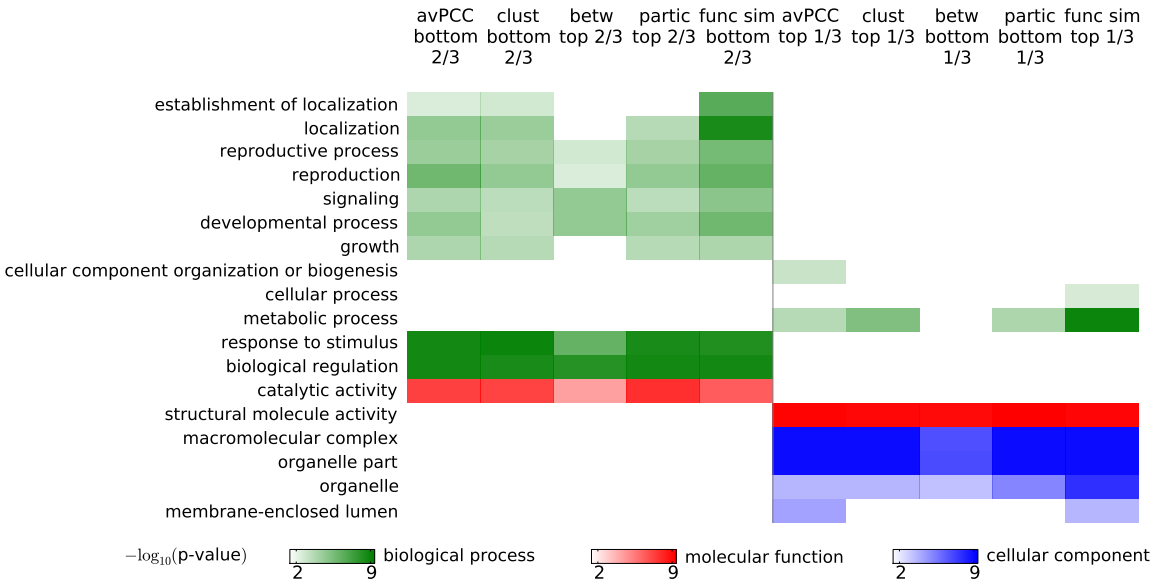

Figure A.18: GO annotation enrichment analysis of hubs in **Yeast-all**.
GO annotation enrichment analysis of hubs divided in a 2-to-1 proportion by avPCC, clustering, betweenness, participation and functional similarity scores in **Yeast-all**. See Fig. 2.3 for details.

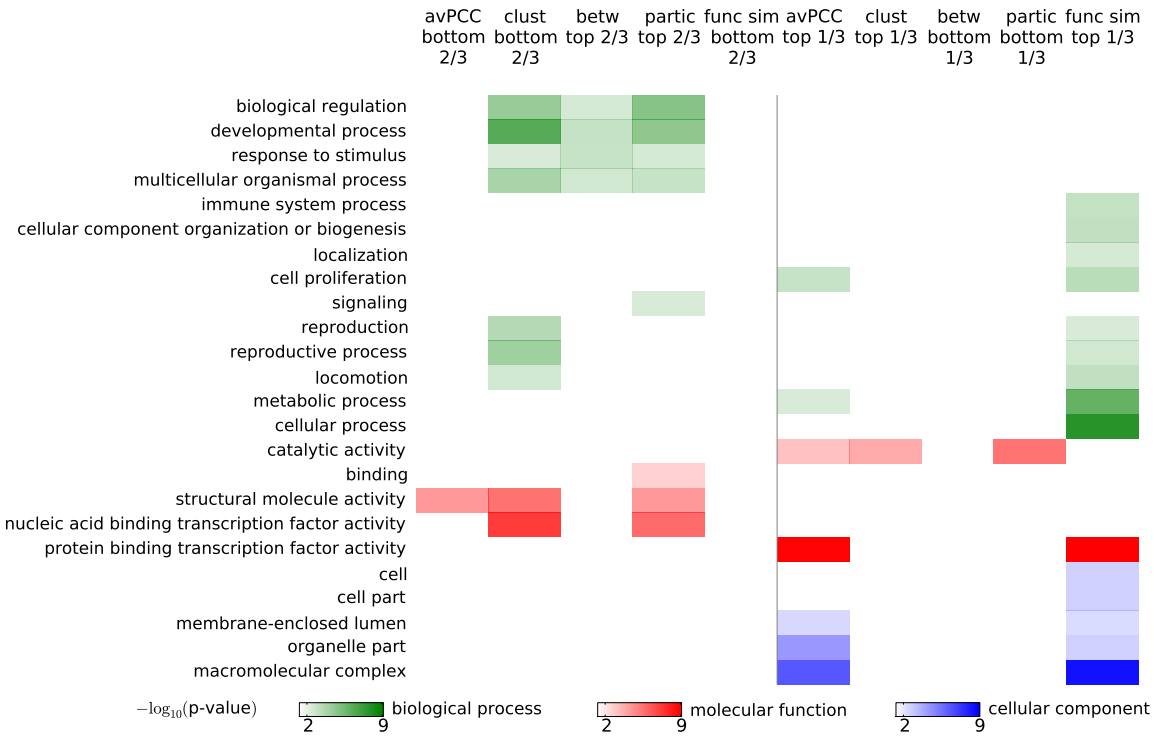

Figure A.19: GO annotation enrichment analysis of hubs in **Fly**.
GO annotation enrichment analysis of hubs divided in a 2-to-1 proportion by avPCC, clustering, betweenness, participation and functional similarity scores in **Fly**. See Fig. 2.3 for details.

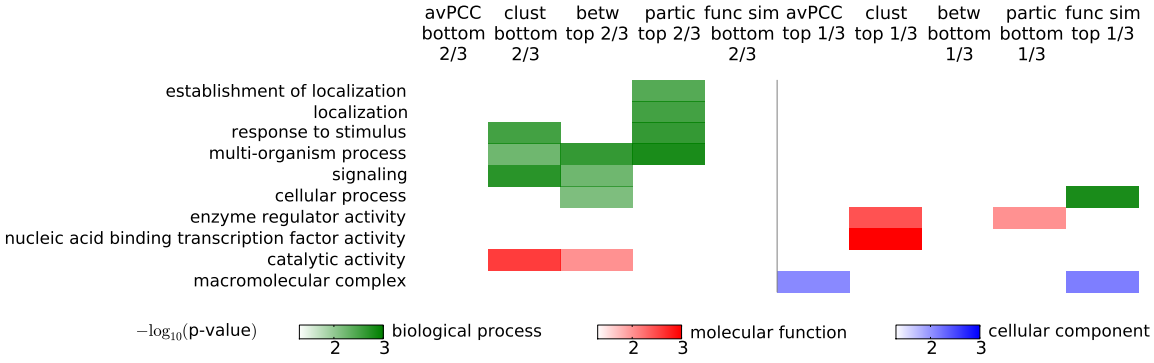

Figure A.20: GO annotation enrichment analysis of hubs in **Athal**.
GO annotation enrichment analysis of hubs divided in a 2-to-1 proportion by avPCC, clustering, betweenness, participation and functional similarity scores in **Athal**. See Fig. 2.3 for details.

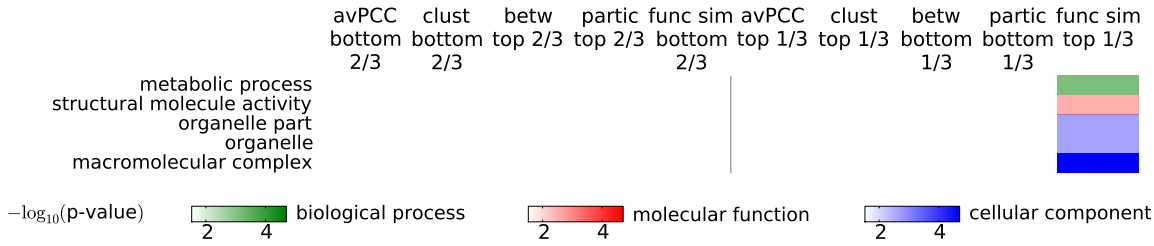

Figure A.21: GO annotation enrichment analysis of hubs in **Ecoli**.
GO annotation enrichment analysis of hubs divided in a 2-to-1 proportion by avPCC, clustering, betweenness, participation and functional similarity scores in **Ecoli**. See Fig. 2.3 for details.

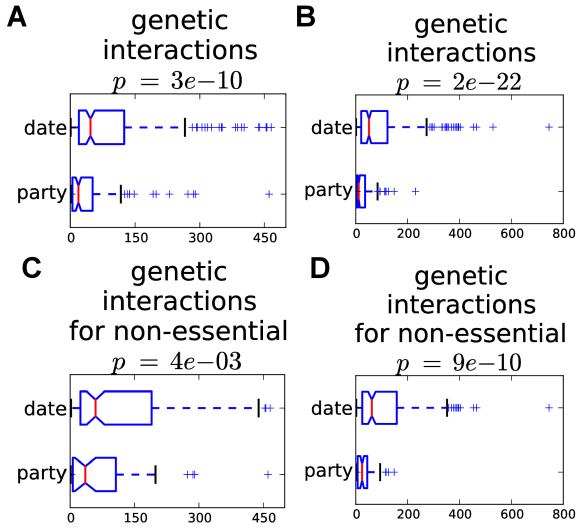

Figure A.22: Genetic interactions for date and party hubs in yeast.
Date hubs participate in significantly larger number of genetic interactions than party hubs, when date and party hubs are defined from yeast networks (A) **Yeast-hq** (B) **Yeast-all** (Mann–Whitney U). Even when all essential genes are removed from consideration, the same trend is observed for both networks (C) **Yeast-hq** and (D) **Yeast-all**.

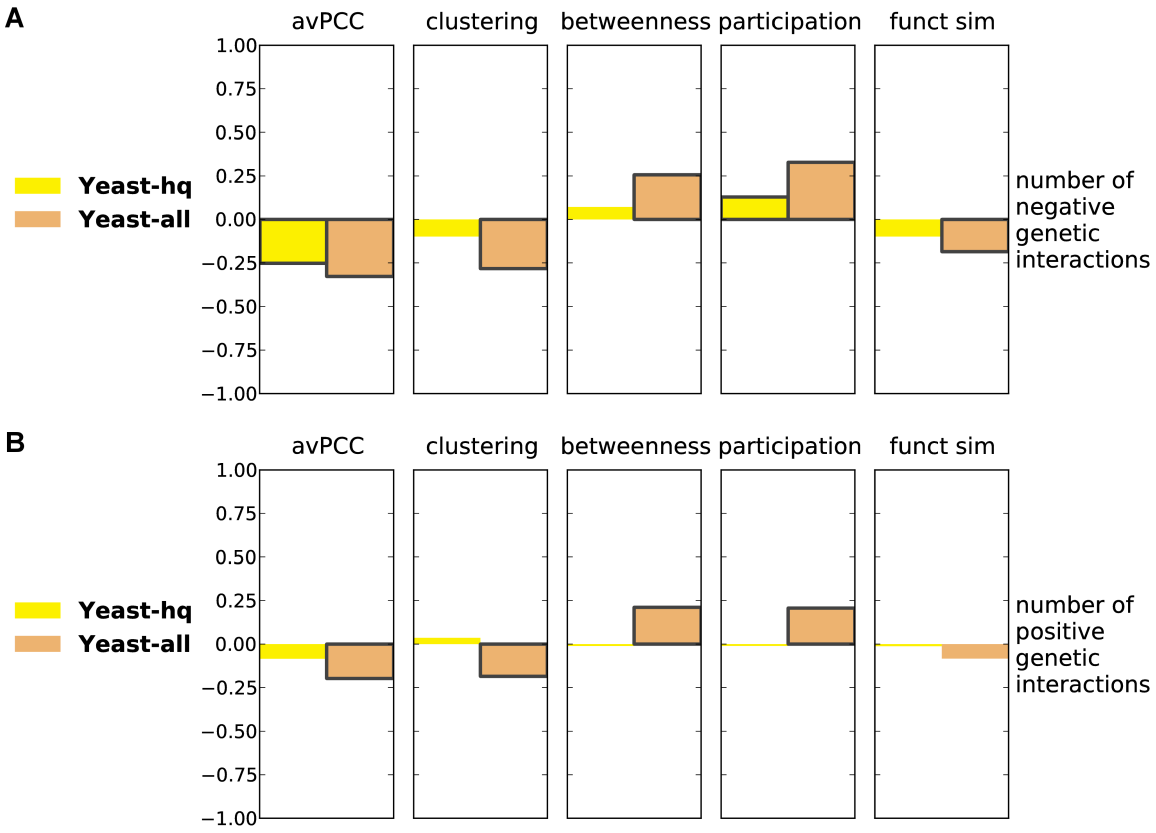

Figure A.23: Spearman correlation of hub characteristics with the number of negative and positive genetic interactions.
(A) Every bar represents a Spearman correlation between a hub characteristic and the number of negative genetic interactions for hubs in one of the physical interaction networks for yeast. (B) Every bar represents a Spearman correlation between a hub characteristic and the number of positive genetic interactions for hubs in one of the physical interaction networks for yeast. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges.

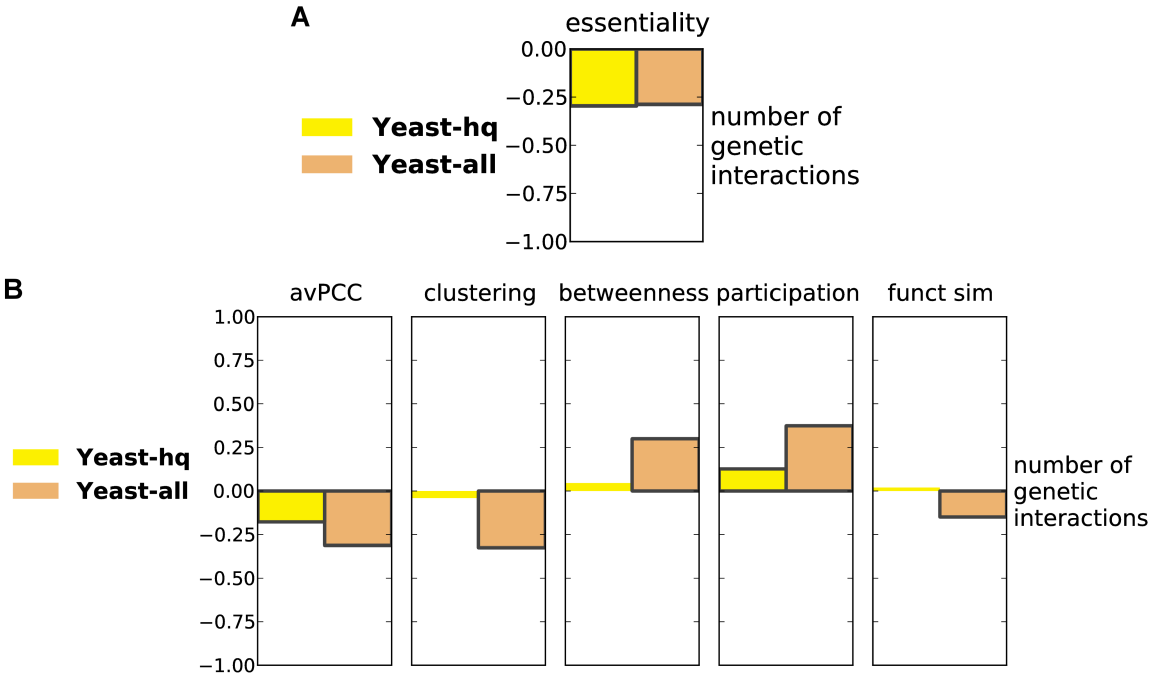

Figure A.24: Essentiality is not a confounding factor in the correlation analysis of genetic degree with hub characteristics in yeast physical interaction networks.
(A) Every bar represents a Spearman correlation between essentiality (1 if essential, 0 otherwise) and the number of genetic interactions for hubs in one of the physical interaction networks for yeast. (B) Every bar represents a partial Spearman correlation between a hub characteristic and the number of genetic interactions corrected for essentiality for hubs in one of the physical interaction networks for yeast. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges.

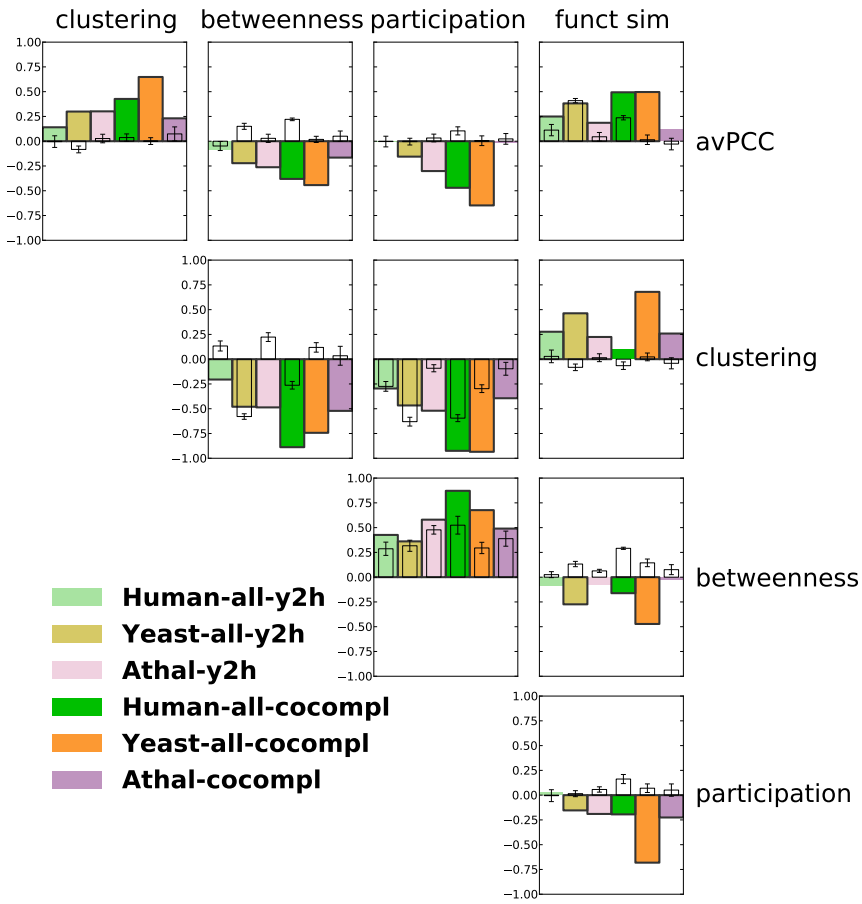

Figure A.25: Spearman correlation of hub characteristics in yeast two-hybrid and co-complex interaction networks.
Every bar represents a Spearman correlation between two characteristics of hubs in one of the networks. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges. Smaller uncolored bars show average correlation (with error bars for standard deviations) in 20 random networks on the same genes with the same number of interactions for each.

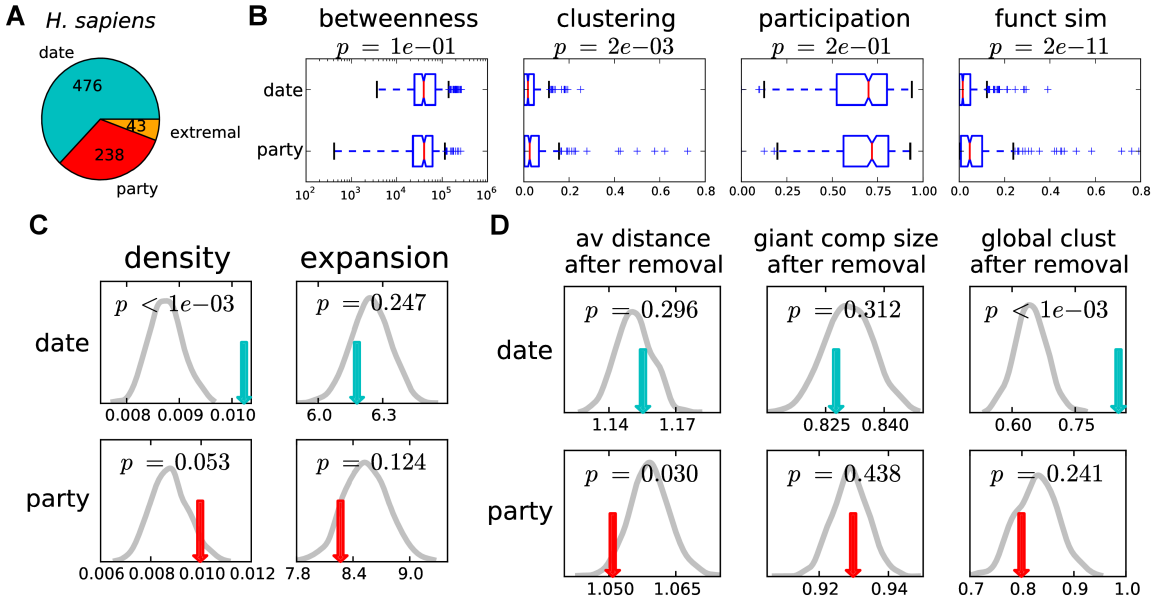

Figure A.26: Date and party hub classification analysis in human network of all known interactions from yeast two-hybrid experiments (**Human-all-y2h**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.08; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

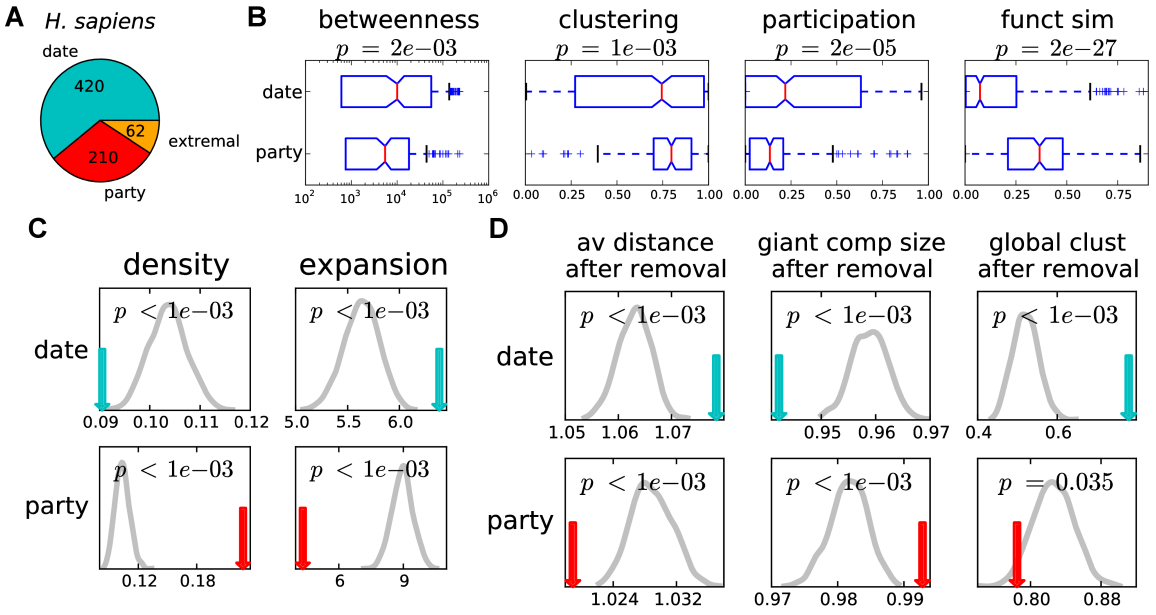

Figure A.27: Date and party hub classification analysis in human network of all known interactions derived from complexes (**Human-all-cocompl**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.30; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

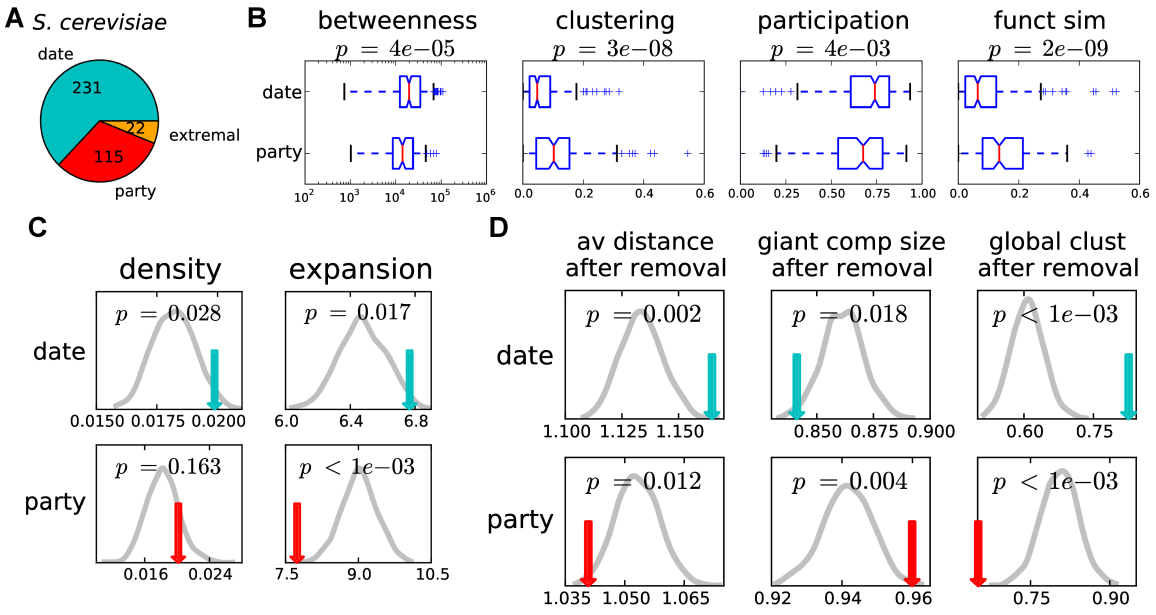

Figure A.28: Date and party hub classification analysis in yeast network of all known interactions from yeast two-hybrid experiments (**Yeast-all-y2h**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.08; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

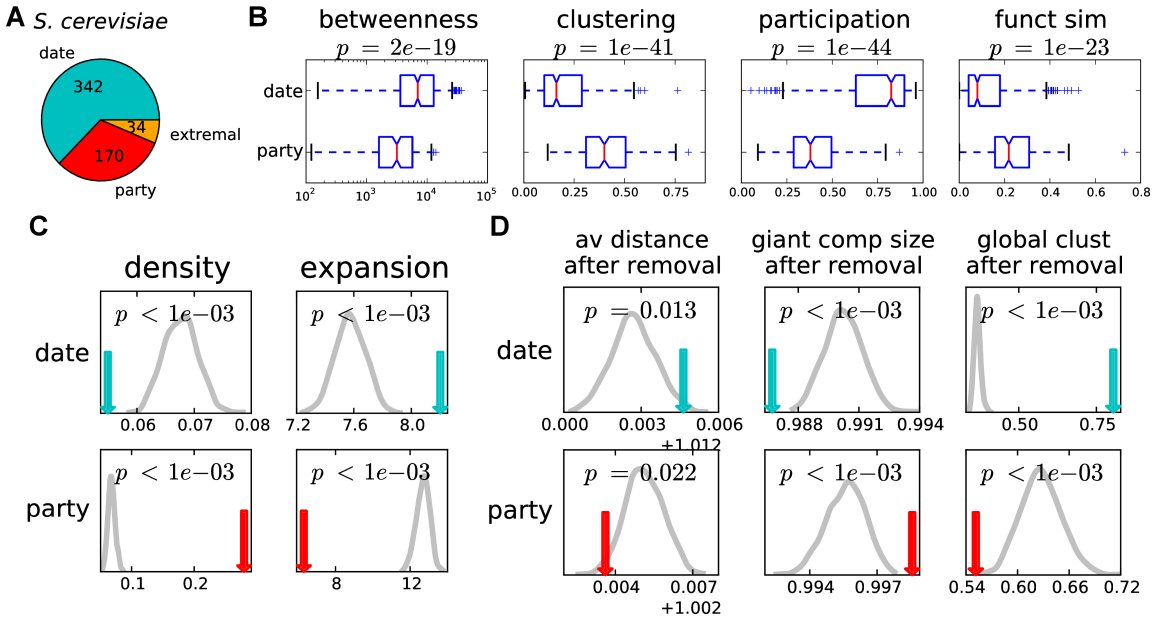

Figure A.29: Date and party hub classification analysis in the yeast network of all known interactions derived from complexes (**Yeast-all-cocompl**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.26; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

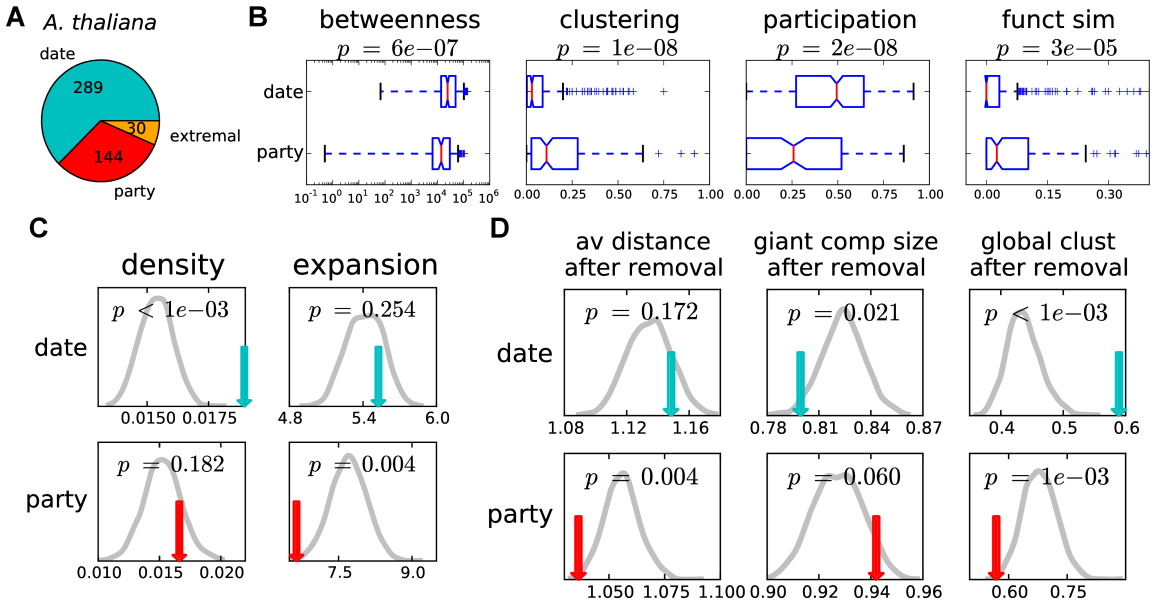

Figure A.30: Date and party hub classification analysis in the Arabidopsis network of all known interactions from yeast two-hybrid experiments (**Athal-y2h**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.12; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

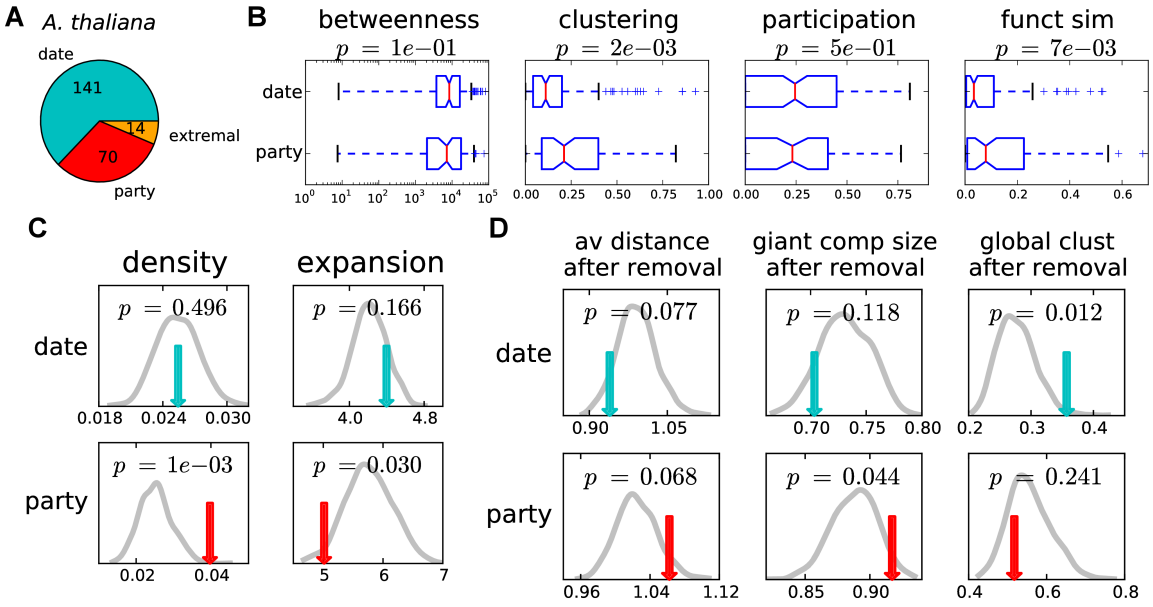

Figure A.31: Date and party hub classification analysis in Arabidopsis network of all known interactions derived from complexes (**Athal-cocompl**).
(A) Number of hubs in each class. Party hubs in this network have avPCC ≥ 0.30; this threshold corresponds to the top third of avPCC values for all hubs categorized as either party or date. (B) Betweenness, clustering coefficient, participation coefficient and functional similarity for date and party hubs. (C) Density and expansion of date and party hubs. (D) Effect of hub removal for party and date when considering the average path distance, the size of the largest connected component, and the global clustering coefficient. See caption of Fig. 2.1 for details.

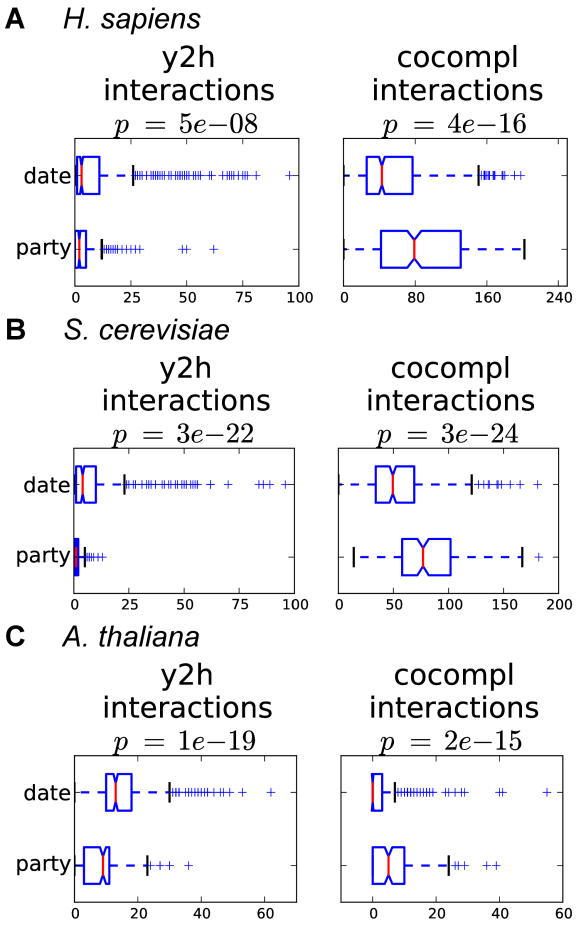

Figure A.32: Yeast two-hybrid and co-complex interactions of date and party hubs. Date hubs have significantly many more binary (yeast two-hybrid, y2h) interactions, while party hubs participate in significantly larger number of interactions derived from complexes (co-complex, cocompl) in networks (A) **Human-all** (B) **Yeast-all** (C) **Athal** (Mann–Whitney U).

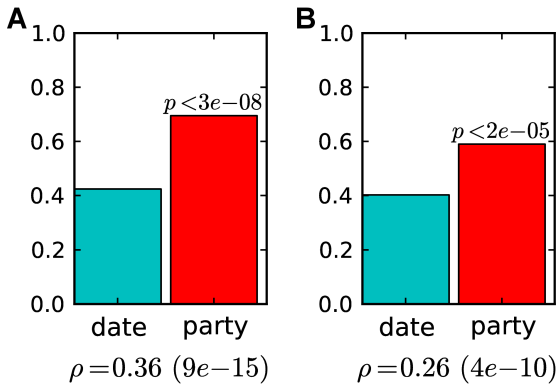

Figure A.33: Party hubs are more likely to be essential than date hubs. Fraction of date and party hubs that are essential in (A) **Yeast-hq** (B) **Yeast-all**. Party hubs are significantly enriched with essential genes (hypergeometric test). Spearman correlation (with p-value) of essentiality indicator vector (1 if essential, 0 otherwise) and avPCC shown on bottom is significantly positive.

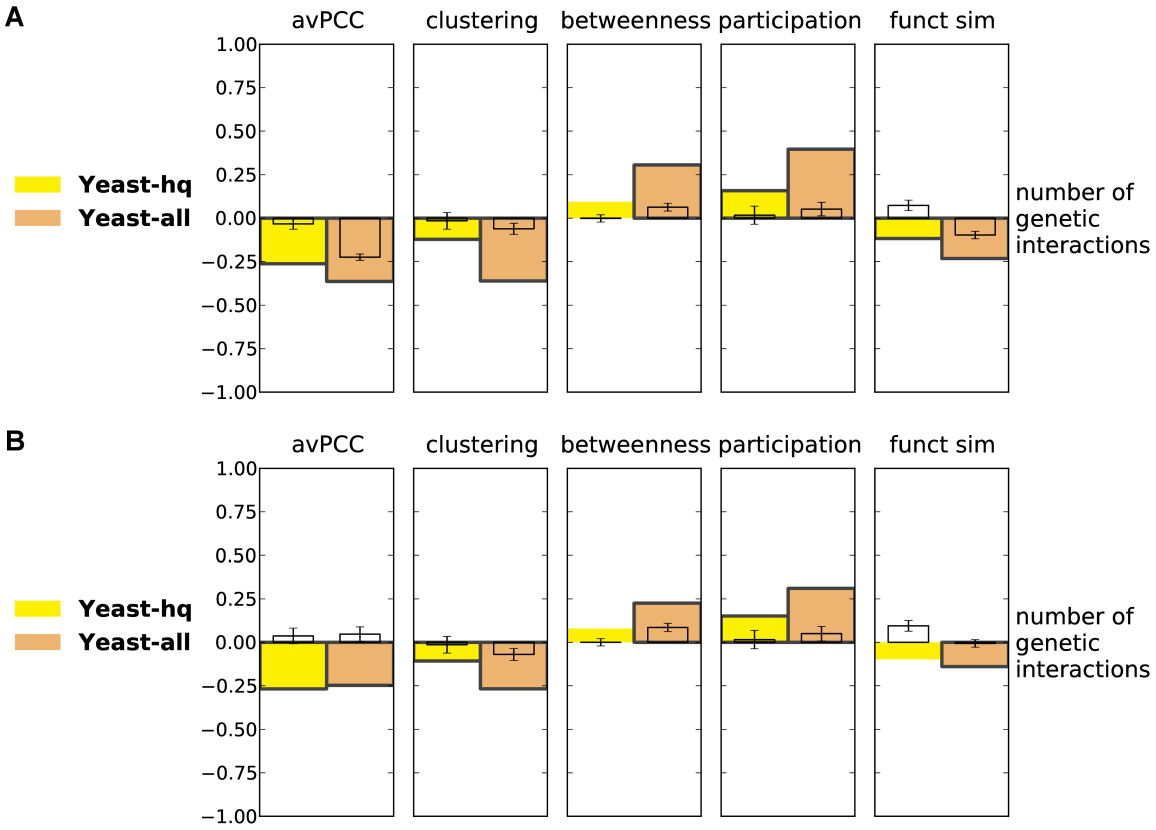

Figure A.34: avPCC-rand is not a confounding factor in the correlation analysis of hub characteristics and genetic degree in yeast physical interaction networks.
(A) Every bar represents a Spearman correlation between a hub characteristic and the number of genetic interactions for hubs in one of the yeast networks. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges. (B) Every bar represents a partial Spearman correlation between a hub characteristic and the number of genetic interactions corrected for avPCC-rand for hubs in one of the yeast networks. Smaller uncolored bars show average correlation (with error bars for standard deviations) in 100 random networks on the same genes with the same number of interactions for each. Random networks used for the plot are different from those used for the calculation of avPCC-rand.

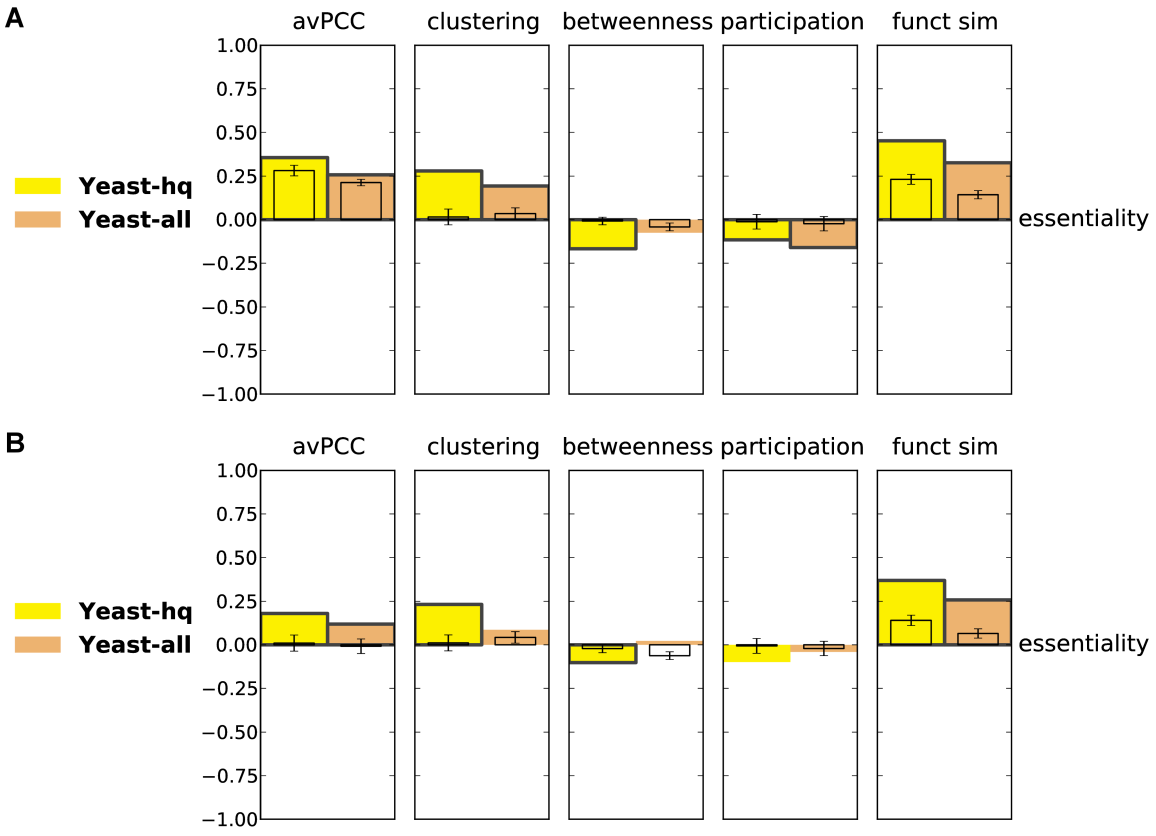

Figure A.35: avPCC-rand is not a confounding factor in the correlation analysis of hub characteristics and essentiality in yeast physical interaction networks.
(A) Every bar represents a Spearman correlation between a hub characteristic and essentiality (1 if essential, 0 otherwise) for hubs in one of the yeast networks. (B) Every bar represents a partial Spearman correlation between a hub characteristic and essentiality (1 if essential, 0 otherwise) corrected for avPCC-rand for hubs in one of the yeast networks. Bars of significant correlations (absolute value > 0.1, p-value < 0.05) have black edges. Smaller uncolored bars show average correlation (with error bars for standard deviations) in 100 random networks on the same genes with the same number of interactions for each. Random networks used for the plot are different from those used for the calculation of avPCC-rand.

A.4 Supplementary tables

	clustering	betweenness	participation	func. similarity
Human-hq	**0.61** ($7e{-}48$)	**−0.54** ($1e{-}36$)	**−0.59** ($1e{-}45$)	**0.59** ($1e{-}44$)
Yeast-hq	**0.39** ($5e{-}18$)	**−0.36** ($6e{-}15$)	**−0.30** ($2e{-}10$)	**0.42** ($2e{-}20$)
Fly	**0.51** ($5e{-}59$)	**−0.28** ($3e{-}17$)	**−0.42** ($2e{-}37$)	**0.50** ($4e{-}55$)
Athal	**0.31** ($5e{-}13$)	**−0.20** ($5e{-}06$)	**−0.26** ($2e{-}09$)	**0.25** ($1e{-}08$)
Ecoli	**0.30** ($6e{-}08$)	0.03 ($6e{-}01$)	**−0.30** ($4e{-}08$)	**0.30** ($4e{-}08$)
Human-all	**0.56** ($4e{-}79$)	**−0.42** ($2e{-}40$)	**−0.58** ($2e{-}83$)	**0.46** ($4e{-}49$)
Yeast-all	**0.72** ($1e{-}91$)	**−0.55** ($2e{-}45$)	**−0.69** ($6e{-}82$)	**0.56** ($3e{-}49$)

Table A.1: Spearman correlation of avPCC with clustering, betweenness, participation and functional similarity of hubs in the network.
All correlations except one are significant ($p < 0.05$) and are shown in bold. avPCC is positively correlated with clustering coefficient and functional similarity, while negatively correlated with betweenness centrality and participation coefficient. See also Tables A.2, A.3, and A.4.

	betweenness	participation	func. similarity
Human-hq	**−0.84** ($4e{-}131$)	**−0.85** ($5e{-}135$)	**0.71** ($6e{-}75$)
Yeast-hq	**−0.78** ($1e{-}94$)	**−0.78** ($3e{-}93$)	**0.57** ($1e{-}39$)
Fly	**−0.59** ($1e{-}80$)	**−0.86** ($2e{-}250$)	**0.29** ($4e{-}18$)
Athal	**−0.57** ($6e{-}49$)	**−0.55** ($5e{-}46$)	**0.31** ($5e{-}14$)
Ecoli	**−0.52** ($5e{-}23$)	**−0.84** ($8e{-}85$)	0.03 ($6e{-}01$)
Human-all	**−0.85** ($5e{-}290$)	**−0.96** (0)	**0.26** ($6e{-}17$)
Yeast-all	**−0.78** ($9e{-}116$)	**−0.92** ($4e{-}234$)	**0.68** ($6e{-}78$)

Table A.2: Spearman correlation of clustering coefficient with betweenness, participation and functional similarity of hubs in the network.
All correlations except one are significant ($p < 0.05$) and are shown in bold. See also Tables A.1, A.3 and A.4.

	participation	func. similarity
Human-hq	**0.83** ($2e{-}123$)	**−0.62** ($6e{-}52$)
Yeast-hq	**0.60** ($1e{-}44$)	**−0.43** ($7e{-}22$)
Fly	**0.62** ($2e{-}92$)	**−0.09** ($5e{-}03$)
Athal	**0.51** ($1e{-}38$)	**−0.09** ($3e{-}02$)
Ecoli	**0.39** ($6e{-}13$)	0.09 ($1e{-}01$)
Human-all	**0.83** ($1e{-}263$)	**−0.19** ($2e{-}09$)
Yeast-all	**0.70** ($1e{-}86$)	**−0.54** ($9e{-}44$)

Table A.3: Spearman correlation of betweenness centrality with participation and functional similarity of hubs in the network.
All correlations except one are significant ($p < 0.05$) and are shown in bold. See also Tables A.1, A.2 and A.4.

	func. similarity
Human-hq	**−0.65** ($1e{-}58$)
Yeast-hq	**−0.27** ($4e{-}09$)
Fly	**−0.21** ($3e{-}10$)
Athal	**−0.21** ($4e{-}07$)
Ecoli	−0.03 ($6e{-}01$)
Human-all	**−0.27** ($6e{-}19$)
Yeast-all	**−0.59** ($3e{-}54$)

Table A.4: Spearman correlation of participation coefficient with functional similarity.
All correlations except one are significant ($p < 0.05$) and are shown in bold. See also Tables A.1, A.3 and A.2.

characteristic	ρ	p-val	empirical p-val
avPCC	**0.23**	0.005	0.001
clustering	**0.62**	$7e{-}17$	< 0.001
betweenness	**0.53**	$6e{-}12$	< 0.001
participation	**0.58**	$2e{-}14$	< 0.001
func. sim	**0.49**	$4e{-}10$	< 0.001

Table A.5: Spearman correlation for characteristics of orthologous hubs in **Yeast-hq** and **Human-hq**.
Five hub characteristics for all 149 orthologous pairs between 109 hubs in **Yeast-hq** and 124 hubs in **Human-hq** are significantly positively correlated, as measured by Spearman's rho (ρ) and the correspondingly determined p-values and empirical p-values for 1000 random permutations of hubs. See Section 2.4 for details.

networks 1 and 2	ρ	p-val	empirical p-val	# orthologs	# hubs org 1	# hubs org 2
Athal and **Fly**	**0.38**	$2e{-}06$	< 0.001	143	99	73
Yeast-all and **Athal**	**0.26**	0.005	0.004	115	48	67
Yeast-hq and **Fly**	**0.26**	0.002	0.001	136	100	117
Athal and **Human-all**	0.14	0.07	0.030	180	90	89
Athal and **Human-hq**	0.08	0.3	0.150	133	79	63
Fly and **Human-all**	0.02	0.7	0.362	316	201	235
Yeast-all and **Fly**	-0.06	0.4	0.215	182	139	134
Yeast-hq and **Athal**	-0.18	0.1	0.066	68	39	60

Table A.6: Spearman correlation of avPCC for orthologs between species.
avPCC correlation analysis for hubs in pairs of networks: Spearman's rho, corresponding p-value, empirical p-value for 1000 random permutations of avPCC among hubs, the number of orthologous pairs of hubs and the numbers of hubs in each organism involved into orthologs (only for hubs with assigned avPCC score). Correlations with absolute value above 0.1 and both p-values < 0.05 are shown in bold. See Section 2.4 for details.

networks 1 and 2	ρ	p-val	empirical p-val
Yeast-hq and **Athal**	**0.35**	0.003	< 0.001
Athal and **Human-hq**	**0.34**	$6e{-}05$	0.001
Fly and **Human-all**	**0.28**	$7e{-}07$	< 0.001
Yeast-hq and **Fly**	**0.25**	0.003	0.003
Athal and **Human-all**	**0.23**	0.002	0.001
Yeast-all and **Athal**	**0.22**	0.02	0.012
Yeast-all and **Fly**	**0.17**	0.02	0.009
Athal and **Fly**	-0.03	0.7	0.343

Table A.7: Spearman correlation of clustering coefficient for orthologs between species.
Clustering coefficient correlation analysis for hubs in pairs of networks: Spearman's rho, corresponding p-value, empirical p-value for 1000 random permutations of clustering coefficient values among hubs. Correlations with absolute value above 0.1 and both p-values < 0.05 are shown in bold. See Section 2.4 for details.

networks 1 and 2	ρ	p-val	empirical p-val
Yeast-all and **Athal**	**0.42**	$3e{-}06$	< 0.001
Athal and **Human-hq**	**0.38**	$7e{-}06$	< 0.001
Athal and **Human-all**	**0.33**	$6e{-}06$	< 0.001
Yeast-hq and **Athal**	**0.37**	0.002	0.001
Fly and **Human-all**	**0.27**	$8e{-}07$	< 0.001
Yeast-hq and **Fly**	**0.26**	0.002	0.001
Yeast-all and **Fly**	0.11	0.1	0.067
Athal and **Fly**	-0.00	1.0	0.493

Table A.8: Spearman correlation of betweenness centrality for orthologs between species.
Betweenness centrality correlation analysis for hubs in pairs of networks: Spearman's rho, corresponding p-value, empirical p-value for 1000 random permutations of betweenness centrality values among hubs. Correlations with absolute value above 0.1 and both p-values < 0.05 are shown in bold. See Section 2.4 for details.

networks 1 and 2	ρ	p-val	empirical p-val
Fly and **Human-all**	**0.19**	$9e{-}04$	< 0.001
Yeast-hq and **Athal**	0.18	0.2	0.090
Yeast-all and **Athal**	0.15	0.1	0.043
Athal and **Human-all**	0.09	0.2	0.105
Athal and **Human-hq**	0.11	0.2	0.093
Yeast-hq and **Fly**	0.03	0.8	0.391
Yeast-all and **Fly**	-0.03	0.8	0.395
Athal and **Fly**	0.02	0.9	0.436

Table A.9: Spearman correlation of participation coefficient for orthologs between species.
Participation coefficient correlation analysis for hubs in pairs of networks: Spearman's rho, corresponding p-value, empirical p-value for 1000 random permutations of participation coefficient values among hubs. Correlations with absolute value above 0.1 and both p-values < 0.05 are shown in bold. See Section 2.4 for details.

networks 1 and 2	ρ	p-val	empirical p-val
Yeast-hq and **Fly**	**0.37**	$2e{-}05$	< 0.001
Fly and **Human-all**	**0.25**	$8e{-}06$	< 0.001
Athal and **Fly**	**0.24**	$4e{-}03$	0.004
Yeast-all and **Athal**	**0.19**	0.04	0.022
Yeast-all and **Fly**	0.12	0.1	0.055
Athal and **Human-all**	0.05	0.5	0.211
Yeast-hq and **Athal**	-0.02	0.9	0.424
Athal and **Human-hq**	-0.06	0.5	0.260

Table A.10: Spearman correlation of functional similarity for orthologs between species.
Functional similarity correlation analysis for hubs in pairs of networks: Spearman's rho, corresponding p-value, empirical p-value for 1000 random permutations of functional similarity values among hubs. Correlations with absolute value above 0.1 and both p-values < 0.05 are shown in bold. See Section 2.4 for details.

network	BP	CC	MF	none
Human-hq	97	96	74	3
Yeast-hq	92	96	71	2
Fly-hq	68	63	66	21
Athal	71	58	69	13
Ecoli	48	39	49	35
Human-all	85	86	61	10
Yeast-all	94	98	79	0

Table A.11: Fraction of hubs annotated with GO terms in each network.
For each network and for each ontology (Biological process, Cellular component, Molecular function), we show the percentage of hubs annotated with at least some terms other than the root (the most general) term in this ontology (BP, CC, MF), and the percentage of hubs not annotated in any of the three ontologies (none).

	yeast two-hybrid	co-complex
BioGRID	Two-hybrid	Affinity Capture-Luminescence Affinity Capture-MS Affinity Capture-RNA Affinity Capture-Western Co-purification Reconstituted Complex
IntAct	two hybrid	tandem affinity purification pull down anti tag coimmunoprecipitation anti bait coimmunoprecipitation affinity chromatography technology
Bossi and Lehner 2009	two_hybrid -two_hybrid two_hybrid_test yeast two_hybridarray two_hybridpooling sd4-two_hybrid lacz4-two_hybrid two_h	affinitycapture_ms affin gst_pulldown affinity mass_spectrometry indirect_complex reconstitutedcomplex affinity_tag affinity_chromatography co_purification direct_complex copurification affinity_c affinity_chrom tandem_affinitypurification tap literature_annotated_complex affinity_techniques affinity_co_purification affinitycapture_western affinit pull_down

Table A.12: Interaction evidence types from different sources used for interaction annotation.
Note that evidence types from [77] are not organized in any formal vocabulary and were parsed from Column 3 of the table with interactions using symbols |, (,), and : as punctuation.

Table A.13: Datasets used in *S. cerevisiae* expression compendium.

GEO accession number	publication or submission information	# datapoints
GSE29894	Cell cycle and G1 cyclins. Public on Jun 11, 2011. Skotheim Lab, Stanford `http://web.me.com/skotheim/Site/People.html`	32
GSE22904	Lewis JA, Elkon IM, McGee MA, Higbee AJ et al. Exploiting natural variation in Saccharomyces cerevisiae to identify genes for increased ethanol resistance. Genetics 2010 Dec;186(4):1197-205. PMID: 20855568	18
GSE23204	The Role of the Rad4-Rad23 Complex and Rad4 Ubiquitination in UV-Responsive Transcription. Public on Aug 02, 2010. Humphryes N, Reed S. Cardiff University School of Medicine	12
GSE22458	Bermejo C, Garca R, Straede A, Rodrguez-Pea JM et al. Characterization of sensor-specific stress response by transcriptional profiling of wsc1 and mid2 deletion strains and chimeric sensors in Saccharomyces cerevisiae. OMICS 2010 Dec;14(6):679-88. PMID: 20958245	10
GSE15254	Staschke KA, Dey S, Zaborske JM, Palam LR et al. Integration of general amino acid control and target of rapamycin (TOR) regulatory pathways in nitrogen assimilation in yeast. J Biol Chem 2010 May 28;285(22):16893-911. PMID: 20233714	18
GSE15147	Eng KH, Kvitek DJ, Keles S, Gasch AP. Transient genotype-by-environment interactions following environmental shock provide a source of expression variation for essential genes. Genetics 2010 Feb;184(2):587-93. PMID: 19966067	34
GSE18121	Gene expression regulation in response to heat stress in different yeast strains. Public on Nov 09, 2009. Cowart LA, Lu X, Hannun Y. Medical University of South Carolina	21
GSE13653	Halbeisen RE, Gerber AP. Stress-Dependent Coordination of Transcriptome and Translatome in Yeast. PLoS Biol 2009 May 5;7(5):e105. PMID: 19419242	12
GSE8335	Berry DB, Gasch AP. Stress-activated genomic expression changes serve a preparative role for impending stress in yeast. Mol Biol Cell 2008 Nov;19(11):4580-7. PMID: 18753408	32
GSE7645	Expression data for Saccharomyces cerevisiae oxidative stress response. Public on Oct 24, 2007. Sha W, Martins A, Laubenbacher R, Mendes P, Shulaev V. Virginia Bioinformatics Institute	16

GSE7362	The contribution of different nutrients to spore germination in Saccharomyces cerevisiae. Joseph-Strauss D, Zenvirth D, Simchen G, Barkai N. Spore germination in Saccharomyces cerevisiae: global gene expression patterns and cell cycle landmarks. Genome Biol 2007;8(11):R241. PMID: 17999778	38
GSE7358	Spore Germination in Saccharomyces cerevisiae : transfer of wild type spores to rich (YPD) medium. Joseph-Strauss D, Zenvirth D, Simchen G, Barkai N. Spore germination in Saccharomyces cerevisiae: global gene expression patterns and cell cycle landmarks. Genome Biol 2007;8(11):R241. PMID: 17999778	31
GSE12270	Capaldi AP, Kaplan T, Liu Y, Habib N et al. Structure and function of a transcriptional network activated by the MAPK Hog1. Nat Genet 2008 Nov;40(11):1300-6. PMID: 18931682	29
GSE4987	Pramila T, Wu W, Miles S, Noble WS et al. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. Genes Dev 2006 Aug 15;20(16):2266-78. PMID: 16912276	25
GSE5376	Cell cycle of yeast deleted for yox1. Public on Sep 30, 2007. Pramila T, Breeden LL. Breeden Lab, FHCRC	25
GSE6302	Levy S, Ihmels J, Carmi M, Weinberger A et al. Strategy of transcription regulation in the budding yeast. PLoS One 2007 Feb 28;2(2):e250. PMID: 17327914	92
GSE8825	Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R et al. Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. Mol Biol Cell 2008 Jan;19(1):352-67. PMID: 17959824	36
GSE8982	Mating response — six alpha factor concentrations (0.06, 0.2, 0.6, 6, 60 and 600 nM). Public on Sep 11, 2007. Barkai Lab, Department of Molecular Genetics, Weizmann Institute of Science	33
GSE10521	Azzouz N, Panasenko OO, Deluen C, Hsieh J et al. Specific roles for the Ccr4-Not complex subunits in expression of the genome. RNA 2009 Mar;15(3):377-83. PMID: 19155328	14
GSE11397	Willis IM, Chua G, Tong AH, Brost RL et al. Genetic interactions of MAF1 identify a role for Med20 in transcriptional repression of ribosomal protein genes. PLoS Genet 2008 Jul 4;4(7):e1000112. PMID: 18604275	12

Table A.14: Datasets used in *D. melanogaster* expression compendium.

GEO accession number	publication or submission information	# datapoints
GSE7763	FlyAtlas `http://flyatlas.org`. Chintapalli VR, Wang J, Dow JA. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet 2007 Jun;39(6):715-20. PMID: 17534367	30
GSE6186	Whole Genome Drosophila Embryogenesis Time Course. Public on Nov 02, 2006. Hooper SD, Boue S, Krause R, Jensen LJ, Mason CE, Ghanim M, Furlong EE, White KP, Bork P. State Lab, Department of Genetics, Yale University	28
GSE5430	Qin X, Ahn S, Speed TP, Rubin GM. Global analyses of mRNA translational control during early Drosophila embryogenesis. Genome Biol 2007;8(4):R63. PMID: 17448252	39
GSE8892	Abundant genetic variation in transcript level during early Drosophila development. Public on Aug 30, 2007. Nuzhdin SV, Tufts DM, Hahn MW. Department of Biology, Indiana University	18
GSE13303	Gene Expression during the Egg Development of Drosophila melanogaster. Baker DA, Russell S. Department of Genetics, University of Cambridge	12
GSE22354	Pavlopoulos A, Akam M. Hox gene Ultrabithorax regulates distinct sets of target genes at successive stages of Drosophila haltere morphogenesis. Proc Natl Acad Sci U S A 2011 Feb 15;108(7):2855-60. PMID: 21282633	12
GSE20497	Menin links the stress response to genome stability in Drosophila melanogaster. Razak Z. Canadian Drosophila Microarray Centre `http://www.flyarrays.com`	12
GSE5147	Sørensen JG, Nielsen MM, Kruhøffer M, Justesen J et al. Full genome gene expression analysis of the heat stress response in Drosophila melanogaster. Cell Stress Chaperones 2005 Winter;10(4):312-28. PMID: 16333985	18
modENCODE	Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH; modENCODE Consortium. Unlocking the secrets of the genome. Nature. 2009 Jun 18;459(7249):927-30. PMID: 19536255. As analyzed and published by FlyBase `http://flybase.org` at FlyBase High Throughput Expression Pattern Data Beta Version, FBrf0212041 (2010.10.13) `http://flybase.org/reports/FBrf0212041.html`	30

Appendix B

Supplementary information for Chapter 3

B.1 Supplementary results

B.1.1 Length and number of domains in multifunctional and other annotated proteins

We observe that proteins encoded by multifunctional genes are significantly longer than proteins encoded by other annotated genes (see main text and Fig. 3.2). Equivalently, whether a gene is multifunctional is positively correlated with the length of a protein: Spearman's $\rho = 0.17$ ($p < 6e{-}40$) for *D. melanogaster*, $\rho = 0.06$ ($p < 2e{-}9$) for *H. sapiens*, and $\rho = 0.10$ ($p < 1e{-}11$) for *S. cerevisiae*. Also, proteins encoded by multifunctional genes have a significantly larger number of unique domains than proteins encoded by other annotated genes (see main text and Fig. 3.2). Equivalently, multifunctionality has a significant positive Spearman correlation with the number of unique protein domains: $\rho = 0.07$ ($p < 1e{-}7$) for *D. melanogaster*, $\rho = 0.07$ ($p < 4e{-}11$) for *H. sapiens*, and $\rho = 0.06$ ($p < 1e{-}4$) for *S. cerevisiae*.

However, one may expect that longer proteins have more domains, so the difference in length could explain the difference in the number of domains between multifunctional and other annotated genes. Indeed, we observe strong significant positive Spearman correlation between protein length and the number of unique domains in all three organisms: $\rho = 0.52$ for *D. melanogaster*, $\rho = 0.65$ for *H. sapiens*, and $\rho = 0.61$ for *S. cerevisiae*. We compute the partial Spearman correlation between multifunctionality and the number of domains with a correction for protein length and observe a significant (though small) value only for human: $\rho = 0.04$ ($p < 2e{-}4$). Therefore the difference in length between multifunctional and other annotated proteins may indeed explain the significant difference in the number of domains, or the difference in the number of domains between multifunctional and other annotated proteins may indeed explain significant difference in length. Further investigation is required.

B.2 Supplementary materials and methods

B.2.1 Comparison of multifunctional and other annotated genes with correction for a gene feature

Comparison of the set of multifunctional genes M and other annotated genes N with correction for a gene feature f—e.g., degree or the number of associated publications (one value for each gene)—is performed as follows. We sample with replacement $n = 1000$ times independently at random from the set of genes N, so that each sample s_i is the list of the same number of genes as M (potentially with repetitions) having the same distribution of the feature f as M. The procedure to select each sample s is as follows. Start with a list l of size $|M|$ of genes each of which is chosen from N uniformly at random. Then repeat the following: on each step try to swap a random element in l with a random gene from N, and accept the swap only if the distributions of f on l and M become more similar to each other, as measured by the

Mann–Whitney U statistic. Continue with these steps until the relative change in the U statistic has been less than 10^{-4} in the past 100 steps (in practice this requires $< 10^5$ steps). The samples s_i, $i = 1, \ldots, n$ produced with this method are used for the corrected comparison of M and N.

For a corrected comparison of M and N with respect to a certain gene property D—e.g., association with disease (given as a set of genes having the property)—compute the frequency x of D in M and frequencies y_i of D in samples s_i. Then, estimate the corrected frequency of D in N as the median of all y_i. Use the distribution of y_i to compute the 95% confidence interval (using 2.5% quantile cutoffs on both sides) and the empirical p-value (as the fraction of times x is higher than y_i). Similarly, for the comparison with respect to a gene feature B—e.g., betweenness centrality (one value for each gene)—compute the distribution of B in M, including the median b. Also, compute the distributions of B in each sample s_i, including the medians b_i. Then, estimate the corrected median of B in N as the median of all b_i. Also, use the distribution of b_i to compute the 95% confidence interval (using 2.5% quantile cutoffs on both sides) and the empirical p-value (as the fraction of times b is higher than b_i). Use the distributions of B in samples s_i all merged together to compute 25% and 75% quantiles.

B.3 Supplementary figures

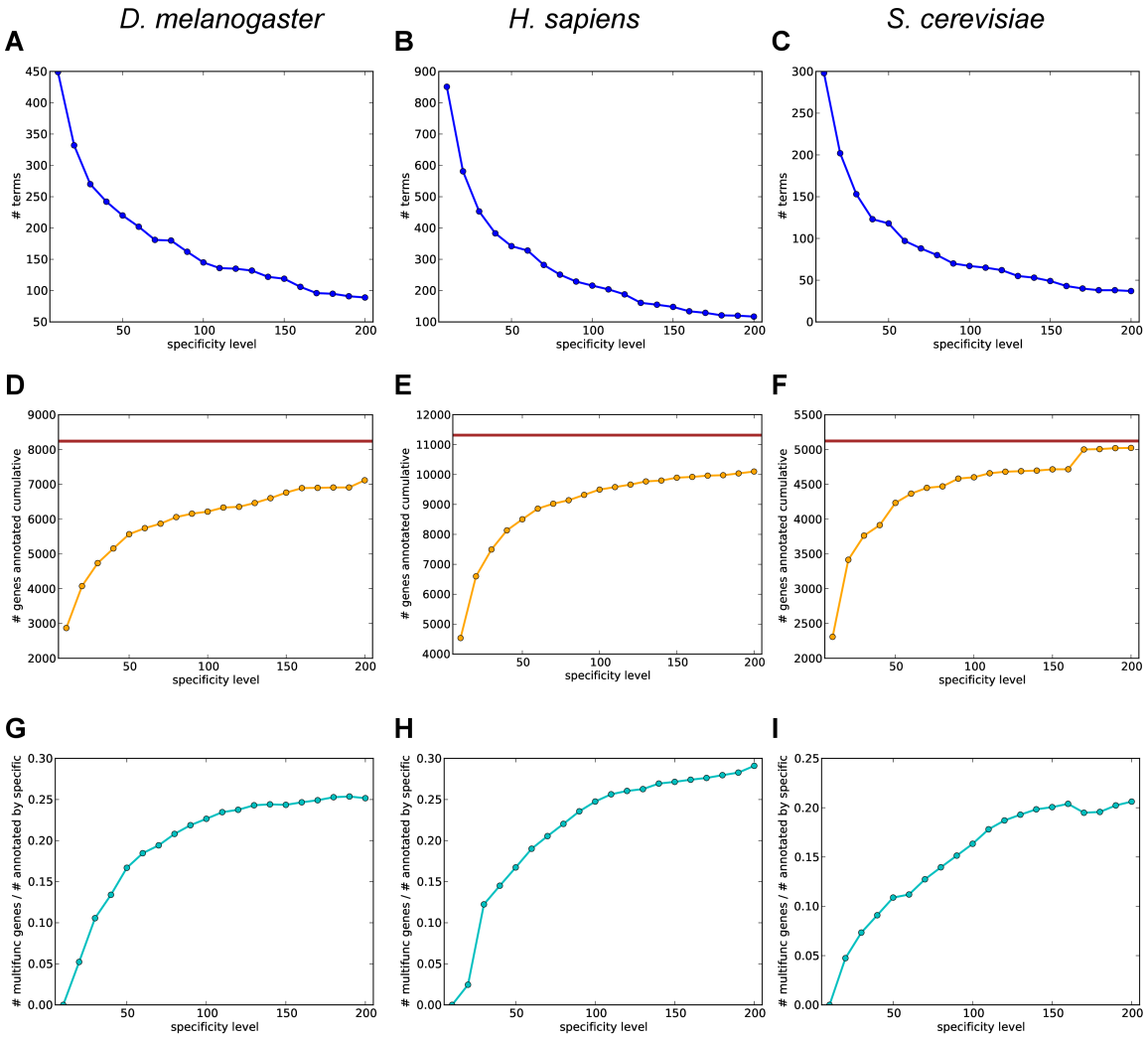

Figure B.1: Effect of varying parameters in the definition of multifunctional genes (A-C) Terms chosen at different specificity levels. The number of Biological Process (BP) Gene Ontology (GO) terms chosen is shown for each specificity threshold from 10 to 200 (increment of 10) for (A) fly, (B) human, (C) yeast. (D-F) Genes annotated with terms chosen at different specificity levels. For each M from 10 to 200 (increment of 10), the cumulative number of genes annotated with terms chosen for specificity thresholds $N \leq M$ is shown for (D) fly, (E) human, (F) yeast. Horizontal line shows the total number of genes annotated with any BP term. (G-I) Fraction of multifunctional genes in all annotated genes. For each specificity threshold, the fraction of the cumulative number of multifunctional genes to the total number of all genes annotated with terms chosen at this threshold is shown for (G) fly, (H) human, (I) yeast. See Section 3.4 for details.

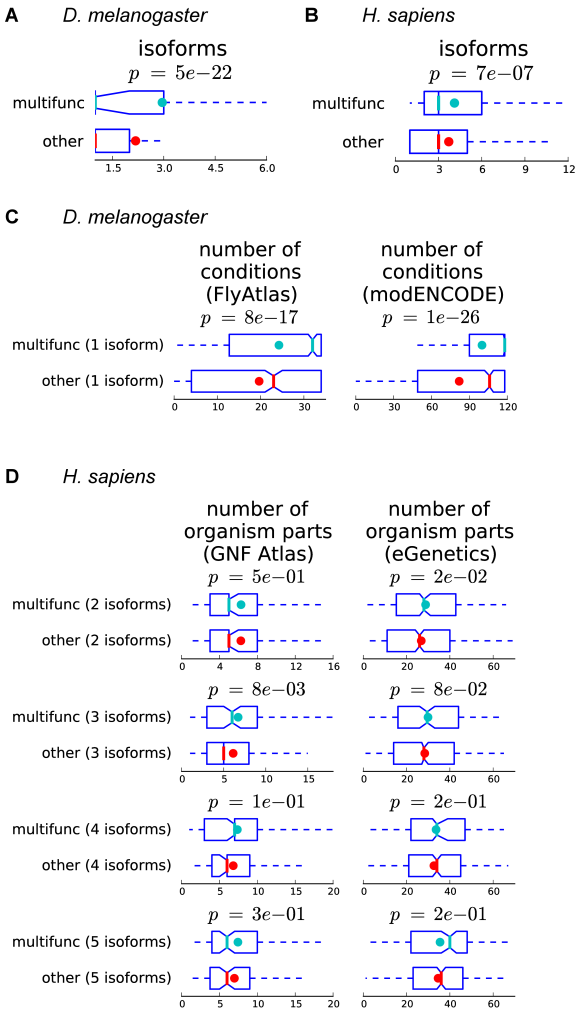

Figure B.2: Multifunctional genes have more isoforms in fly and human. Boxplots of the number of isoforms per gene for multifunctional and other annotated genes in (A) fly and (B) human. Multifunctional genes have significantly larger number of isoforms (Mann–Whitney U). (C) Boxplots of the number of conditions in which multifunctional and other annotated genes in fly are expressed, for genes with one isoform only (which constitute 49% of multifunctional genes and 59% of other annotated genes). (D) Boxplots of the number of organism parts in which multifunctional and other annotated genes in human are expressed, for genes with 2 to 5 isoforms (17% of multifunctional and 18% of other genes have 2 isoforms, 14% of multifunctional and 14% of other genes have 3 isoforms, 10% of multifunctional and 11% of other genes have 4 isoforms, 9% of multifunctional and 8% of other genes have 5 isoforms). See Fig. 3.3 for comparison across all genes.

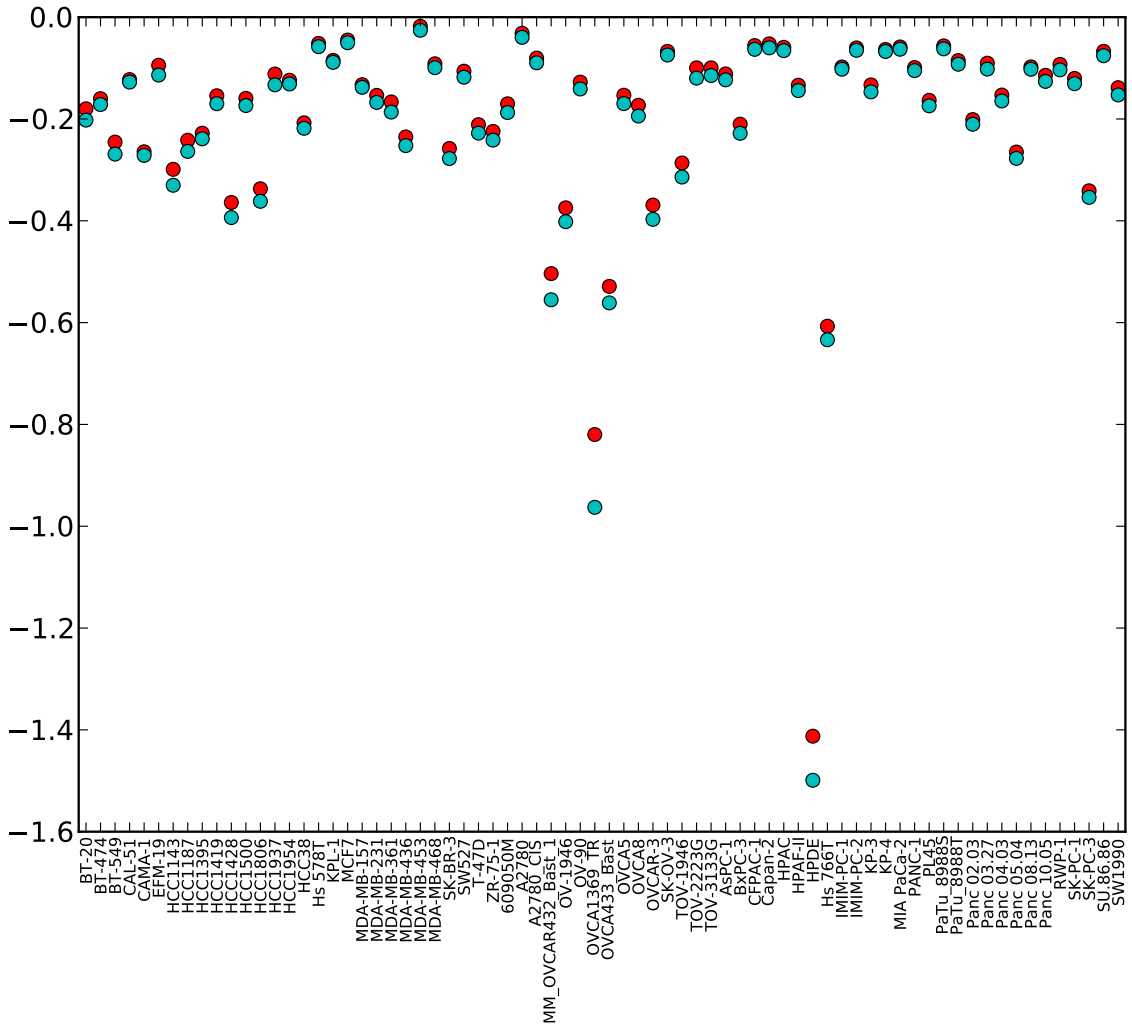

Figure B.3: Multifunctional genes are more essential in human cancer cell lines. For each of 72 human cancer cell lines (x-axis) in the COLT-Cancer database [106, 107] the median of GARP score of essentiality, as reported in the database, is shown for multifunctional (cyan) and all other annotated genes (red) on y-axis; lower GARP scores depict higher essentiality. Multifunctional genes are more essential than the other annotated genes in all 72 cell lines.

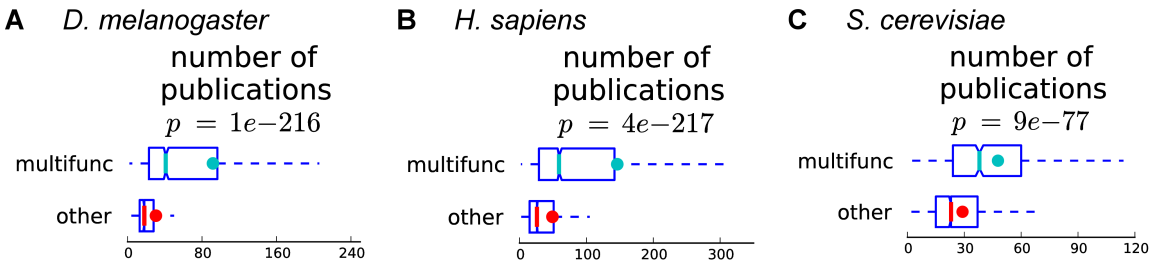

Figure B.4: Multifunctional genes have been more studied than other genes. Boxplots of the number of PubMed publications associated with multifunctional and other annotated genes are shown for (B) fly (C) human and (D) yeast. Multifunctional genes are associated with significantly larger number of publications (Mann–Whitney U test).

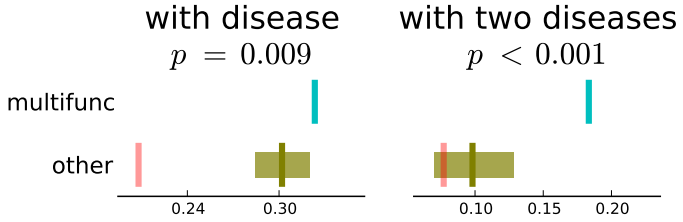

Figure B.5: Comparison of association of multifunctional and other human genes with diseases corrected for study bias.
Fractions of multifunctional (cyan) and non-multifunctional (red) genes associated with diseases are shown (same as in Fig. 3.7), as well as the estimated fraction in non-multifunctional genes after correction for study bias (olive, with the box for 95% confidence interval, and empirical p-value). The estimation is from 1000 independent random samples from the set of non-multifunctional genes, where the samples have the same distribution of the number of associated publications as multifunctional genes. Multifunctional genes are associated with significantly larger number of diseases even after correction for study bias. See Section B.2 for details.

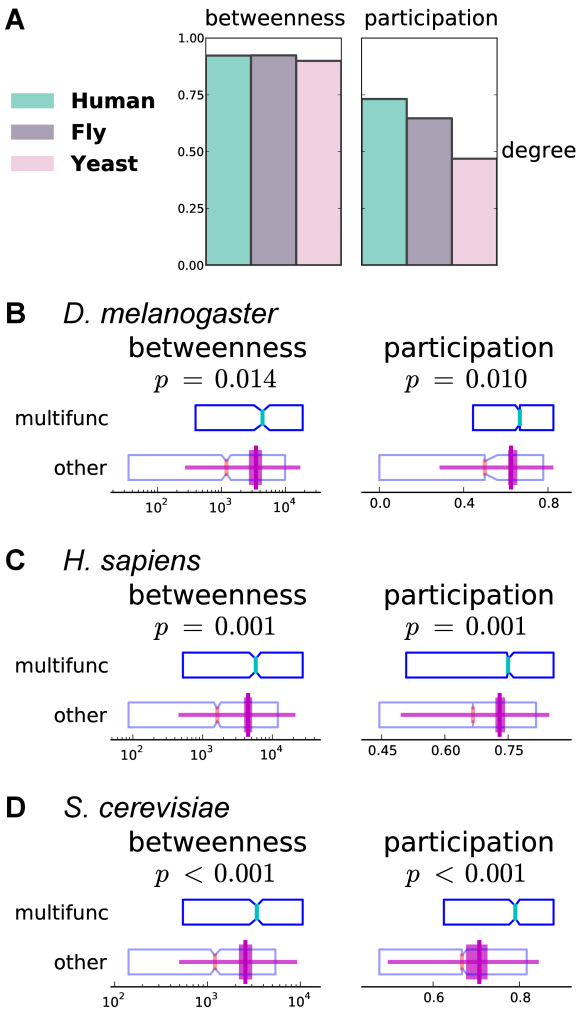

Figure B.6: Comparison of centrality in protein-protein physical interaction networks of multifunctional and other genes corrected for degree distribution.
(A) Barplot of Spearman correlation of degree with betweenness centrality and participation coefficient in physical protein-protein interaction networks. Degree is highly correlated with both measures. (B–D) Comparison of betweenness and participation of multifunctional and non-multifunctional genes with correction for degree. Boxplots show the distribution of betweenness or participation for multifunctional and non-multifunctional genes for (B) fly, (C) human, (D) yeast (same as in Fig. 3.8), while on top of that in magenta color the distribution of the same measure is shown for random samples from the set of non-multifunctional genes, where the samples have the same distribution of degree as multifunctional genes. Vertical line shows the estimated median, box shows 95% confidence interval around the median, and horizontal line shows 25%–75% quantile range. After correction for degree, betweenness and participation of multifunctional genes are significantly higher than for other annotated genes (as by empirical p-value for comparison between medians). See Section B.2 for details.

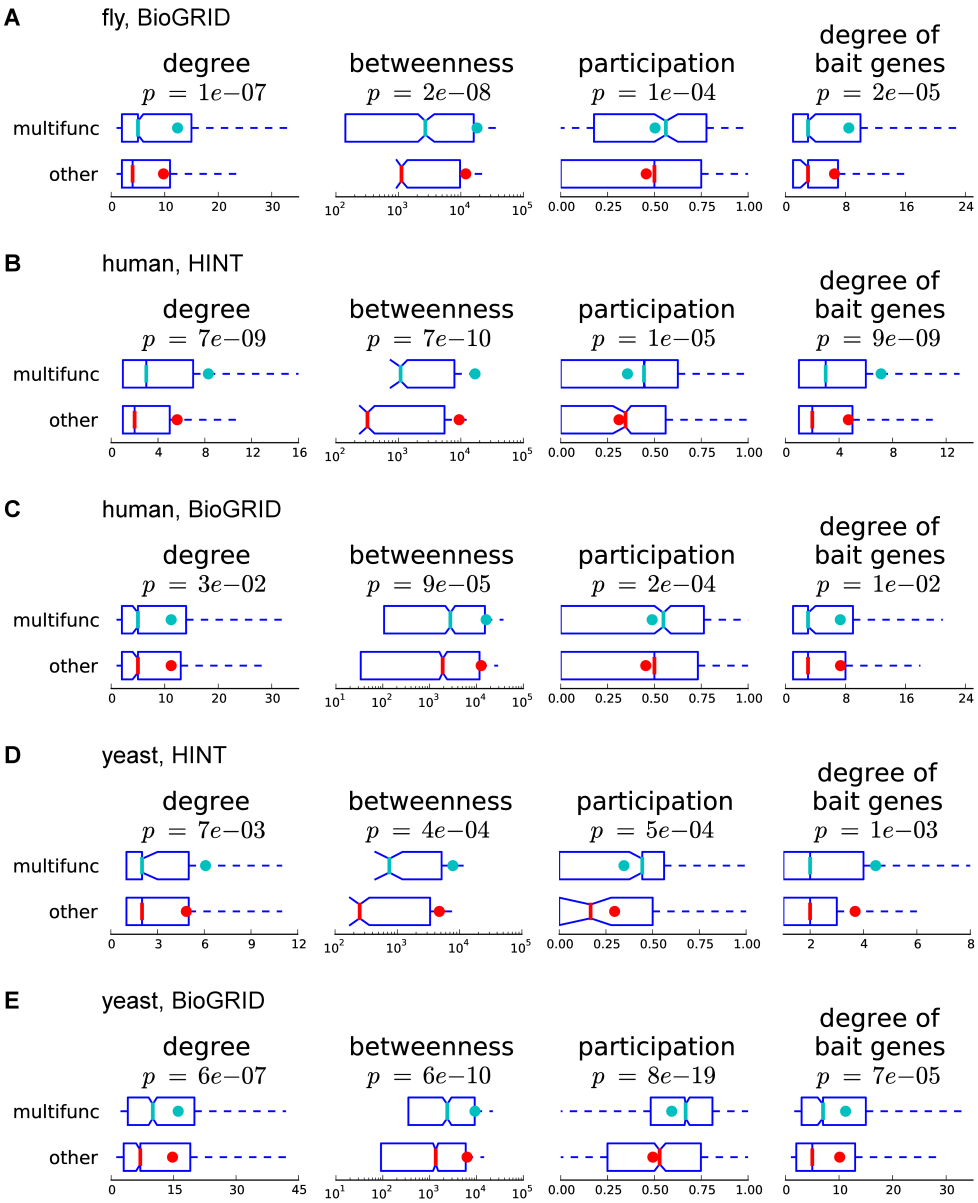

Figure B.7: Centrality of multifunctional genes in high-throughput protein physical interaction networks.

Boxplots with three measures of centrality—degree (number of interactions), betweenness centrality, participation coefficient—in high-throughput protein interaction networks of multifunctional and other annotated genes are shown for (A) fly (BioGRID), (B) human (HINT), (C) human (BioGRID), (D) yeast (HINT), (E) yeast (BioGRID). Multifunctional genes are significantly more central than other genes (Mann–Whitney U test) in high-throughput networks that are not prone to bias towards more studied genes. For an even stricter comparison, the degree of bait genes—i.e., number of interactions from bait to prey genes in these high-throughput experiments—is compared between multifunctional and all other annotated genes, and the trend is confirmed in all networks (A–E).

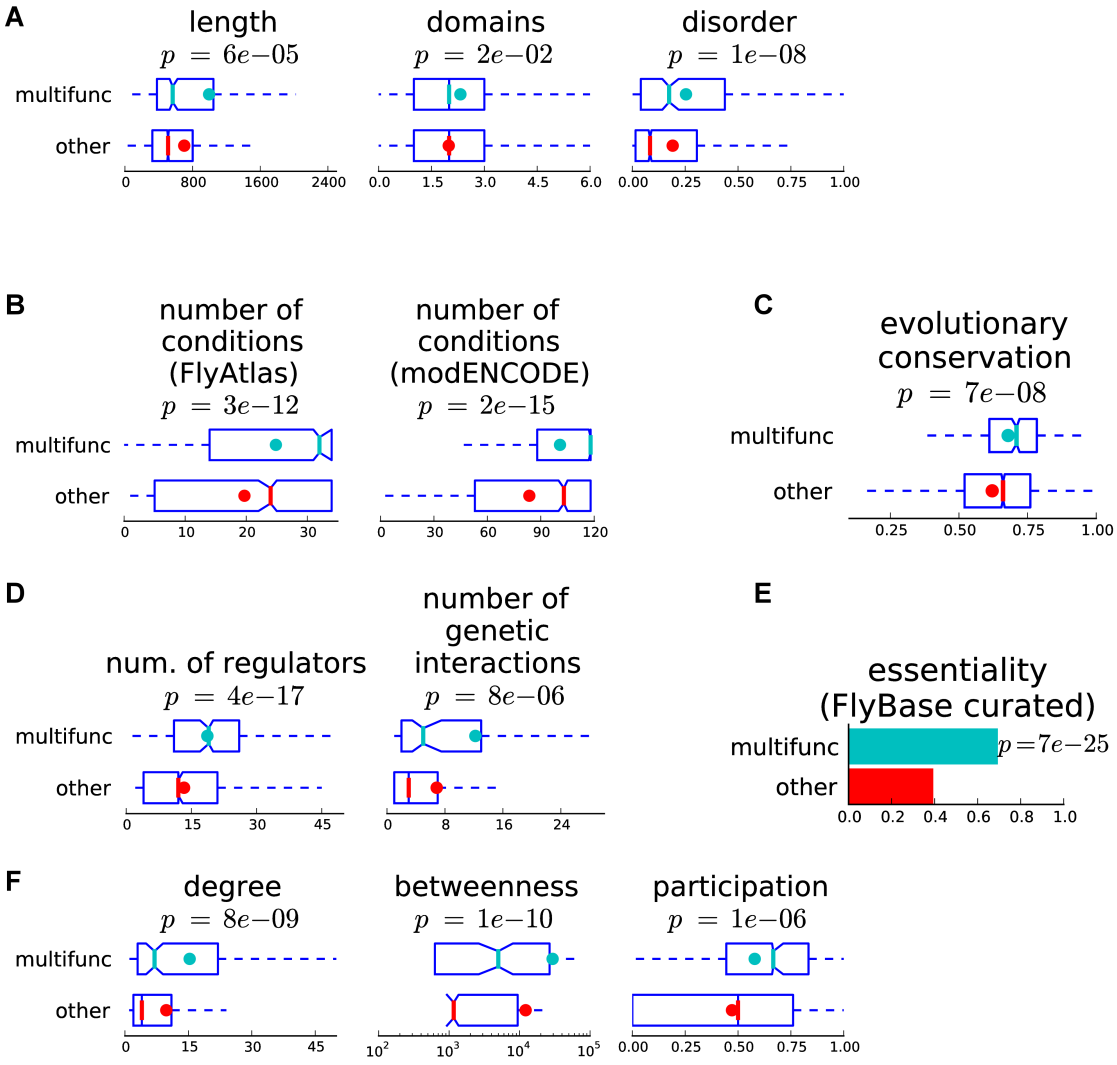

Figure B.8: Analysis of multifunctional genes in *D. melanogaster* obtained using the Molecular Function ontology.
Comparison of multifunctional and other annotated genes obtained from the Molecular Function Gene Ontology using our method (see Fig. 3.1 and Section B.2). (A) Physicochemical properties (compare with Fig. 3.2A). (B) Expression (compare with Fig. 3.3A). (C) Evolutionary conservation (compare with Fig. 3.4A). (D) Regulatory and genetic interactions (compare with Fig. 3.5A). (E) Essentiality (FlyBase curated; compare with Fig. 3.6A). (F) Centrality in protein-protein interaction networks (compare with Fig. 3.8A).

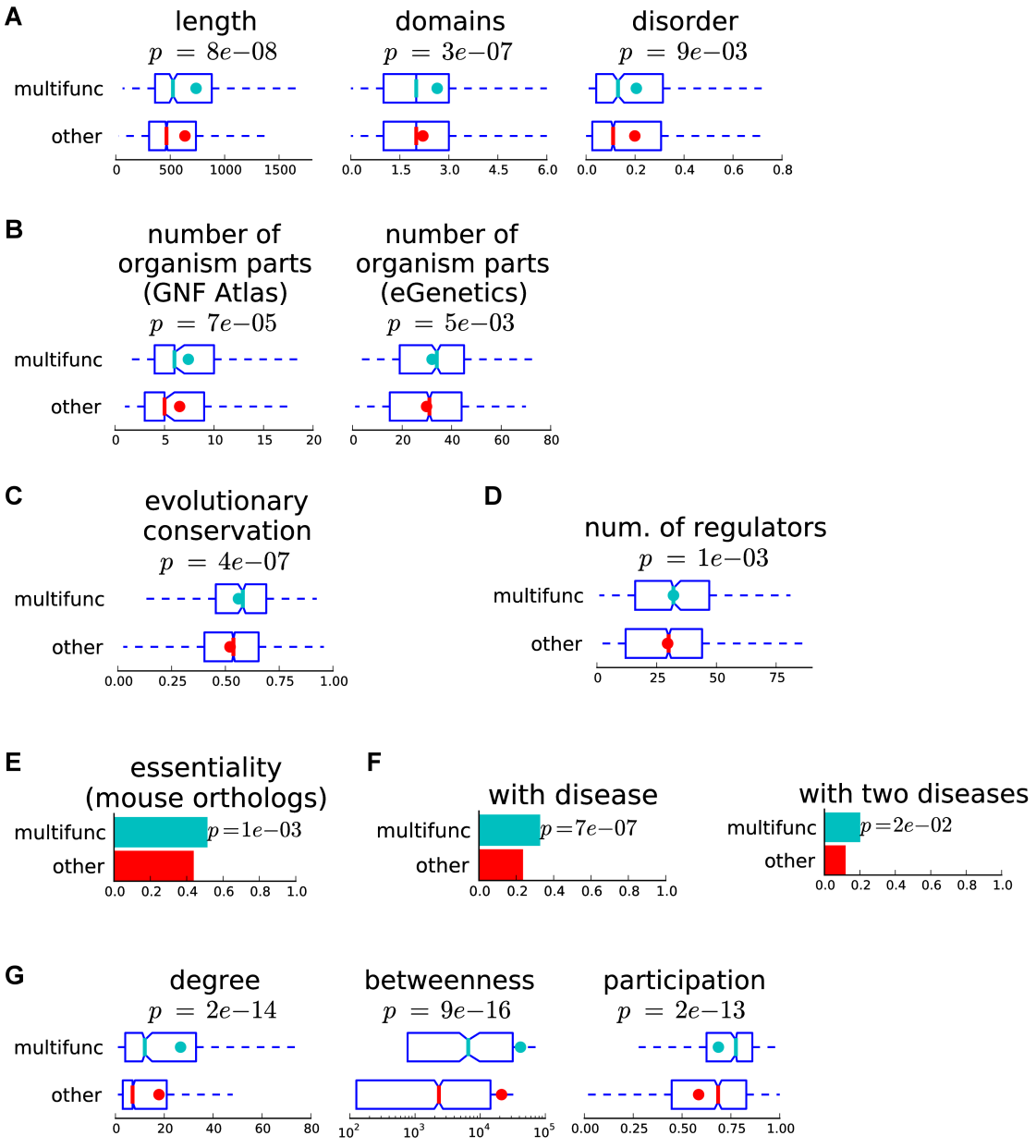

Figure B.9: Analysis of multifunctional genes in *H. sapiens* obtained using the Molecular Function ontology.
Comparison of multifunctional and other annotated genes obtained from the Molecular Function Gene Ontology using our method (see Fig. 3.1 and Section B.2). (A) Physicochemical properties (compare with Fig. 3.2B). (B) Expression (compare with Fig. 3.3B). (C) Evolutionary conservation (compare with Fig. 3.4B). (D) Regulatory interactions (compare with Fig. 3.5B). (E) Essentiality (mouse orthologs; compare with Fig. 3.6C). (F) Association with diseases (compare with Fig. 3.7). (G) Centrality in protein-protein interaction networks (compare with Fig. 3.8B).

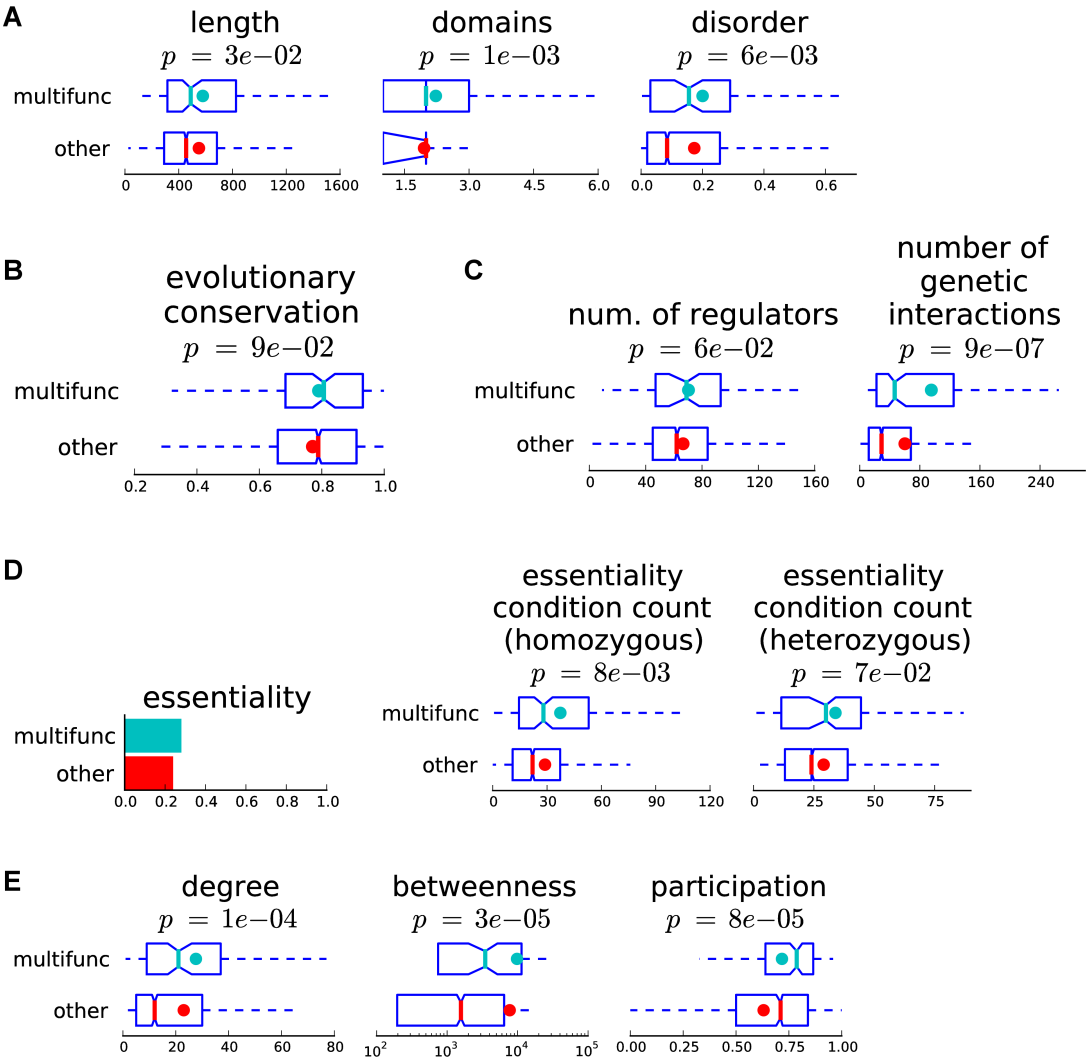

Figure B.10: Analysis of multifunctional genes in *S. cerevisiae* obtained using the Molecular Function ontology.
Comparison of multifunctional and other annotated genes obtained from the Molecular Function Gene Ontology using our method (see Fig. 3.1 and Section B.2). (A) Physicochemical properties (compare with Fig. 3.2C). (B) Evolutionary conservation (compare with Fig. 3.4C). (C) Regulatory and genetic interactions (compare with Fig. 3.5C). (D) Essentiality (compare with Fig. 3.6EF). (E) Centrality in protein-protein interaction networks (compare with Fig. 3.8C).

B.4 Supplementary tables

expression assay	Spearman's ρ
D. melanogaster	
FlyAtlas	0.18 ($p < 5e{-}61$)
modENCODE	0.20 ($p < 3e{-}80$)
H. sapiens	
GNF Atlas	0.12 ($p < 5e{-}34$)
eGenetics	0.33 ($p < 5e{-}238$)

Table B.1: Genes with more isoforms tend to be detected as more broadly expressed. Spearman correlation (with significance p-value) of the number of isoforms of a gene and the number of tissues or organism parts in which the gene is expressed, according to genome-wide assays in fly and human (see Section 3.4).

correlation with multifunctionality			
organism	degree	betweenness	participation
fly	0.13 ($p < 1e{-}15$)	0.15 ($p < 1e{-}19$)	0.12 ($p < 3e{-}13$)
human	0.15 ($p < 1e{-}36$)	0.17 ($p < 1e{-}50$)	0.13 ($p < 6e{-}31$)
yeast	0.15 ($p < 2e{-}21$)	0.16 ($p < 6e{-}26$)	0.18 ($p < 1e{-}31$)
partial correlation with multifunctionality corrected for degree			
organism		betweenness	participation
fly		0.07 ($p < 2e{-}5$)	0.05 ($p < 2e{-}3$)
human		0.10 ($p < 3e{-}17$)	0.05 ($p < 4e{-}5$)
yeast		0.07 ($p < 6e{-}6$)	0.13 ($p < 4e{-}18$)

Table B.2: Comparison of multifunctionality and centrality in protein-protein physical interaction networks.
In the first part of the table, we show Spearman correlations (with significance p-values) of gene multifunctionality (1 or 0 for a gene depending on whether it is multifunctional or not) with degree, betweenness, and participation in protein-protein interaction networks. All correlations are positive and significant (compare with Fig. 3.6). In the second part of the table, we show partial Spearman correlations of gene multifunctionality with betweenness and participation corrected for degree. All partial correlations are small but positive and are statistically significant (compare with Fig. B.6).

degree	number of multifunctional genes	number of intermodular multifunctional genes	number of intermodular multifunctional genes in random networks
D. melanogaster			
all	1075	267	4.3 ± 2.2
≥ 13 (top 20%)	337	183	4.3 ± 2.1
≥ 36 (top 5%)	115	88	3.7 ± 1.9
H. sapiens			
all	2160	828	27.2 ± 5.2
≥ 21 (top 20%)	728	545	26.6 ± 5.3
≥ 60 (top 5%)	249	222	24.6 ± 4.0
S. cerevisiae			
all	833	519	21.8 ± 4.7
≥ 32 (top 20%)	265	236	19.6 ± 4.5
≥ 76 (top 5%)	68	62	13.1 ± 3.0

Table B.3: Intermodularity of multifunctional genes in protein-protein interaction networks integrated with GO annotations. For each organism, we show the number of multifunctional genes in its protein-protein interaction network, the number of multifunctional genes that are detected as intermodular in this network when integrated with GO annotations, and the number of multifunctional genes detected as intermodular in degree- and annotation-preserving random networks. These numbers are shown for all genes in the network, for top 20% genes by degree, and for top 5% genes by degree. The number of multifunctional genes detected as intermodular is significantly higher than the same number in random networks. This observation is not explained by the tendency of multifunctional genes to have high degree, as it still holds when only genes of high degree are considered. See Sections 3.4 and 3.2.8 for more details.

organism	BP-multifunctional	BP-multifunctional annotated in MF by terms used to detect MF-multifunctionality	BP-multifunctional and MF-multifunctional	%	p-value
MF specificity upper bound 120					
D. melanogaster	1509	1210	223	18%	$9e{-}59$
H. sapiens	2517	1967	390	20%	$9e{-}62$
S. cerevisiae	876	682	81	12%	$1e{-}21$
MF specificity upper bound 500					
D. melanogaster	1509	1336	402	30%	$1e{-}87$
H. sapiens	2517	2168	760	35%	$1e{-}102$
S. cerevisiae	876	731	128	18%	$6e{-}18$

Table B.4: Comparison of BP-multifunctional to MF-multifunctional genes.
Analysis of multifunctional genes derived from the Biological Process ontology (BP-multifunctional) using the specificity parameter upper bound 120 (the same as used in the main analysis; see Fig. 3.2, Fig. 3.3, Fig. 3.4, Fig. 3.5, Fig. 3.6, Fig. 3.7, Fig. 3.8), when compared with multifunctional genes derived from the Molecular Function ontology (MF-multifunctional) using the specificity parameter upper bounds 120 (a more specific cut-off) and 500 (a more general cuf-off allowing more genes to be detected as MF-multifunctional). For each organism, shown is the number of BP-multifunctional genes (see Table 3.1), the number of them annotated with specific terms from MF, along with the number and percentage of such genes which are detected as MF-multifunctional, and the p-value from the hypergeometric test corresponding to the significance of intersection. A significant fraction of BP-multifunctional genes are also MF-multifunctional.

organism	MF-multifunctional	MF-multifunctional annotated in BP by terms used to detect BP-multifinctionality	MF-multifunctional and BP-multifunctional	%	p-value
BP specificity upper bound 120					
D. melanogaster	324	309	223	72%	$9e{-}59$
H. sapiens	607	584	390	67%	$9e{-}62$
S. cerevisiae	149	144	81	56%	$1e{-}21$
BP specificity upper bound 500					
D. melanogaster	324	316	248	78%	$2e{-}68$
H. sapiens	607	600	458	76%	$3e{-}61$
S. cerevisiae	149	145	101	70%	$2e{-}26$

Table B.5: Comparison of MF-multifunctional to BP-multifunctional genes.
Analysis of multifunctional genes derived from the Molecular Function ontology (MF-multifunctional) using the specificity parameter upper bound 120 (used in the analysis shown in Fig. B.8, Fig. B.9, Fig. B.10), when compared with multifunctional genes derived from the Biological Process ontology (BP-multifunctional) using the specificity parameter upper bounds 120 (a more specific cut-off) and 500 (a more general cuf-off allowing more genes to be detected as BP-multifunctional). For each organism, shown is the number of MF-multifunctional genes, the number of them annotated with specific terms from BP, along with the number and percentage of such genes which are detected as BP-multifunctional, and the p-value from the hypergeometric test corresponding to the significance of intersection. Most MF-multifunctional genes are also BP-multifunctional.

Bibliography

[1] Y Pritykin and M Singh. Simple topological features reflect dynamics and modularity in protein interaction networks. *PLoS Computational Biology*, 9(10):e1003243, 2013.

[2] P O Brown and D Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999.

[3] A Mortazavi, B A Williams, K McCue, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(621–628), 2008.

[4] J Seebacher and A C Gavin. SnapShot: Protein-Protein Interaction Networks. *Cell*, 144(6):1000, 2011.

[5] A Chatr-aryamontri, B-J Breitkreutz, S Heinicke, et al. The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D816–D823, 2013.

[6] T Barrett, S E Wilhite, P Ledoux, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[7] M Ashburner, C A Ball, J A Blake, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000.

[8] B Berger, J Peng, and M Singh. Computational solutions for omics data. *Nature Reviews Genetics*, 14:333–346, 2013.

[9] B Aranda, P Achuthan, Y Alam-Faruque, et al. The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, 38(suppl. 1):D525–D531, 2010.

[10] A-L Barabási and Z N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2):101–13, 2004.

[11] X Zhu, M Gerstein, and M Snyder. Getting connected: analysis and principles of biological networks. *Genes and Development*, 21:1010–1024, 2007.

[12] H Jeong, S P Mason, A-L Barabási, and Z N Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[13] J-D J Han, D Dupuy, N Bertin, et al. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnology*, 23(7):839–844, 2005.

[14] G Lima-Mendez and J van Helden. The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, 5(12):1482–93, 2009.

[15] H B Fraser. Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–752, 2002.

[16] D Ekman, S Light, ÅK Björklund, and A Elofsson. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome biology*, 7(6):R45, 2006.

[17] E Zotenko, J Mestre, D P O'Leary, and T M Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Computational Biology*, 4(8):e1000140, 2008.

[18] AD Fox, BJ Hescott, AC Blumer, and DK Slonim. Connectedness of ppi network neighborhoods identifies regulatory hub proteins. *Bioinformatics*, 27:1135–1142, 2011.

[19] L H Hartwell, J J Hopfield, S Leibler, and A W Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52, 1999.

[20] G D Bader and C W V Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003.

[21] C Brun, F Chevenet, D Martin, et al. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5:R6, 2003.

[22] P Jiang and M Singh. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*, 26(8):1105–1111, 2010.

[23] T M Przytycka, M Singh, and D K Slonim. Toward the dynamic interactome: it's about time. *Briefings in bioinformatics*, 11(1):15–29, 2010.

[24] J J Han, N Bertin, T Hao, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88–93, 2004.

[25] J Piatigorsky, W E O'Brien, B L Norman, et al. Gene sharing by delta-crystallin and argininosuccinate lyase. *Proc Natl Acad Sci U S A*, 85(10):3479–83, 1988.

[26] C J Jeffery. Moonlighting proteins. *Trends Biochem Sci*, 24(1):8–11, 1999.

[27] J Piatigorsky and G J Wistow. Enzyme/crystallins: gene sharing as an evolutionary strategy. *Cell*, 57(2):197–9, 1989.

[28] X He and J Zhang. Toward a molecular understanding of pleiotropy. *Genetics*, 173(4):1885–1891, 2006.

[29] P Tompa, C Szász, and L Buday. Structural disorder throws new light on moonlighting. *Trends in Biochemical Sciences*, 30(9):484–489, 2005.

[30] J A Capra and M Singh. Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, 24(13):1473–1480, 2008.

[31] P Uetz, L Giot, G Cagney, T A Mansfield, R S Judson, J R Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, and et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.

[32] T Ito, T Chiba, R Ozawa, M Yoshida, M Hattori, and Y Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–74, 2001.

[33] Y Ho, A Gruhler, A Heilbut, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[34] A-C Gavin, P Aloy, P Grandi, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

[35] N J Krogan, G Cagney, H Yu, et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440(7084):637–643, 2006.

[36] H Yu, P Braun, M A Yildirim, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[37] L Giot, J S Bader, C Brouwer, et al. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736, 2003.

[38] K G Guruharsha, J-F Rual, B Zhai, et al. A protein complex network of *Drosophila melanogaster*. *Cell*, 147(3):690–703, 2011.

[39] *Arabidopsis* Interactome Mapping Consortium. Evidence for Network Evolution in an *Arabidopsis* Interactome Map. *Science*, 333(6042):601–607, 2011.

[40] J-F Rual, K Venkatesan, T Hao, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[41] U Stelzl, U Worm, M Lalowski, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[42] R M Ewing, P Chu, F Elisma, et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology*, 3(89):89, 2007.

[43] V Spirin and L A Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100:12123–12128, 2003.

[44] E Zotenko, K Guimaraes, R Jothi, and T Przytycka. Decomposition of overlapping protein complexes: A graph theoretical method for analyzing static and dynamic protein associations. *Algorithms for Molecular Biology*, 1(1):7, 2006.

[45] I Ulitsky and R Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, 1(1):8, 2007.

[46] J Song and M Singh. How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics*, 25(23):3143–3150, 2009.

[47] H-M Kaltenbach and J Stelling. Modular analysis of biological networks. In Igor I. Goryanin and Andrew B. Goryachev, editors, *Advances in Systems Biology*, volume 736 of *Advances in Experimental Medicine and Biology*, pages 3–17. Springer New York, 2012.

[48] J Song and M Singh. From hub proteins to hub modules: The relationship between essentiality and centrality in the yeast interactome at different scales of organization. *PLoS Computational Biology*, 9(2):e1002910, 2013.

[49] S Wachi, K Yoneda, and R Wu. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 21(23):4205–4208, 2005.

[50] D Ghersi and M Singh. Disentangling function from topology to infer the network properties of disease genes. *BMC Systems Biology*, 7:5, 2013.

[51] H B Fraser. Modularity and evolutionary constraint on proteins. *Nature genetics*, 37(4):351–2, 2005.

[52] N Bertin, N Simonis, D Dupuy, et al. Confirmation of organized modularity in the yeast interactome. *PLoS Biology*, 5(6):e153, 2007.

[53] K Komurov and M White. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Molecular Systems Biology*, 3:110, 2007.

[54] I W Taylor, R Linding, D Wardey-Farley, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, 27(2):199–204, 2009.

[55] N Batada, T Reguly, A Breitkreutz, et al. Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biology*, 4(10):e317, 2006.

[56] N Batada, T Reguly, A Breitkreutz, et al. Still stratus not altocumulus: Further evidence against the date/party hub distinction. *PLoS Biology*, 5(6):e154, 2007.

[57] S Agarwal, C M Deane, M Porter, et al. Revisiting date and party hubs: Novel approches to role assignment in protein interaction networks. *PLoS Computational Biology*, 6(6):e1000817, 2010.

[58] T Barrett, D B Troup, S E Wilhite, et al. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Research*, 39(suppl 1):D1005–D1010, 2011.

[59] F Viger and M Latapy. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In *Computing and Combinatorics*, volume 3595 of *Lecture Notes in Computer Science*, pages 440–449. 2005.

[60] C Stark, B J Breitkreutz, A Chatr-aryamontri, et al. The BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 39(suppl. 1):D698–D704, 2011.

[61] J Das, J Mohammed, and H Yu. Genome-scale analysis of interaction dynamics reveals organization of biological networks. *Bioinformatics*, 28(14):1873–1878, 2012.

[62] S Heinicke, M S Livstone, C Lu, et al. The Princeton Protein Orthology Database (P-POD): A Comparative Genomics Analysis Tool for Biologists. *PLoS ONE*, 2(8):e766, 2007.

[63] T Ideker, O Ozier, B Schwikowski, and AF Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–S240, 2002.

[64] L Cabusora, E Sutton, A Fulmer, and CV Forst. Differential network expression during drug and stress response. *Bioinformatics*, 21:2898–2905, 2005.

[65] Y Park and JS Bader. How networks change with time. *Bioinformatics*, 28:i40–i48, 2012.

[66] M W Hahn and A D Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005.

[67] T K B Gandhi, J Zhong, S Mathivanan, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293, 2006.

[68] R Sharan, S Suthram, R M Kelley, T Kuhn, S McCuine, P Uetz, et al. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):1974–9, 2005.

[69] R Sharan and T Ideker. Modeling cellular machinery through biological network comparison. *Nature biotechnology*, 24(4):427–33, 2006.

[70] TA Gibson and DS Goldberg. Improving evolutionary models of protein interaction networks. *Bioinformatics*, 27:376–382, 2011.

[71] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[72] E Banks, E Nabieva, B Chazelle, and M Singh. Organization of physical interactomes as uncovered by network schemas. *PLoS Computational Biology*, 4(10):e1000203, 2008.

[73] S Bansal, S Khandelwal, and LA Meyers. Exploring biological network structure with clustered random networks. *BMC Bioinformatics*, 10:405, 2009.

[74] CM Schneider, L de Arcangelis, and HJ Herrmann. Modeling the topology of protein interaction networks. *Physical Review E*, 84:016112, Jul 2011.

[75] M Shao, Y Yang, J Guan, and S Zhou. Choosing appropriate models for protein-protein interaction networks: a comparison study. *Briefings in Bioinformatics*, pages first published online March 19, 2013, doi:10.1093/bib/bbt014, 2013.

[76] J Das and H Yu. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology*, 6(1):92, 2012.

[77] A Bossi and B Lehner. Tissue specificity and the human protein interaction network. *Molecular systems biology*, 5:260, 2009.

[78] T Murali, S Pacifico, J Yu, S Guest, G G Roberts, and R L Finley Jr. DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila. *Nucleic Acids Research*, 39(suppl 1):D736–D743, 2011.

[79] B Aranda, H Blankenburg, S Kerrien, F S L Brinkman, A Ceol, E Chautard, J M Dana, et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature Methods*, 8:528–529, 2011.

[80] R Guimerá and L A N Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.

[81] M E J Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[82] J van de Peppel and F C P Holstege. Multifunctional genes. *Molecular systems biology*, 1:1–2, 2005.

[83] C J Jeffery. Moonlighting proteins: old proteins learning new tricks. *Trends Genet*, 19(8):415–7, 2003.

[84] C J Jeffery. Moonlighting proteins–an update. *Mol Biosyst*, 5(4):345–50, 2009.

[85] D H Huberts and I J van der Klei. Moonlighting proteins: an intriguing mode of multitasking. *Biochim Biophys Acta*, 1803(4):520–5. Huberts, Daphne H E W.

[86] J L Payne and A Wagner. Constraint and contingency in multifunctional gene regulatory circuits. *PLoS Comput Biol*, 9(6):e1003071, 2013.

[87] A M Dudley, D M Janse, A Tanay, R Shamir, and G M Church. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol Syst Biol*, 1:2005.0001, 2005.

[88] M Salathé, M Ackermann, and S Bonhoeffer. The effect of multifunctionality on the rate of evolution in yeast. *Molecular Biology and Evolution*, 23(4):721–722, 2006.

[89] M Costanzo, A Baryshnikova, J Bellay, Y Kim, E D Spear, et al. The genetic landscape of a cell. *Science*, 327(5964):425–31, 2010.

[90] J Gillis and P Pavlidis. The impact of multifunctional genes on "guilt by association" analysis. *PLOS ONE*, 6(2):e17258, 2011.

[91] E Becker, B Robisson, CE Chapple, A Guénoche, and C Brun. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, 28(1):84–90, 2012.

[92] Z Dosztányi, V Csizmók, P Tompa, and I Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*, 347(4):827–39, 2005.

[93] Z Dosztányi, V Csizmók, P Tompa, and I Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–4, 2005.

[94] V R Chintapalli, J Wang, and J A T Dow. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nature Genetics*, 39:715–720, 2007.

[95] modENCODE Consortium, S Roy, J Ernst, P V Kharchenko, P Kheradpour, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330(6012):1787–97, 2010.

[96] S E. St. Pierre, L Ponting, R Stefancsik, et al. FlyBase 102—advanced approaches to interrogating FlyBase. *Nucleic Acids Research*, 42(D1):D780–D788, 2014.

[97] P Flicek, I Ahmed, M R Amode, et al. Ensembl 2013. *Nucleic Acids Research*, 41(D1):D48–D55, 2013.

[98] R Jovelin and PC Phillips. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biology*, 10:R35, 2009.

[99] A Siepel, G Bejerano, J S Pedersen, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

[100] M B Gerstein, A Kundaje, M Hariharan, S G Landt, K K Yan, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489:91–100, 2012.

[101] KD MacIsaac, T Wang, DB Gordon, DK Gifford, GD Stormo, and E Fraenkel. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7:113, 2006.

[102] B J Venters, S Wachi, T N Mavrich, et al. A Comprehensive Genomic Binding Map of Gene and Chromatin Regulatory Proteins in *Saccharomyces*. *Molecular Cell*, 41(4):480–492, 2011.

[103] M Boutros, A A. Kiger, S Armknecht, et al. Genome-wide RNAi analysis of growth and viability in drosophila cells. *Science*, 303(5659):832–835, 2004.

[104] S Köhler, S C Doelken, C J. Mungall, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1):D966–D974, 2014.

[105] Jose M. Silva, Krista Marran, Joel S. Parker, Javier Silva, Michael Golding, Michael R. Schlabach, Stephen J. Elledge, Gregory J. Hannon, and Kenneth Chang. Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science*, 319(5863):617–620, 2008.

[106] R Marcotte, K R Brown, F Suarez, et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discovery*, 2(2):172–189, 2012.

[107] Judice L. Y. Koh, Kevin R. Brown, Azin Sayad, Dahlia Kasimer, Troy Ketela, and Jason Moffat. COLT-Cancer: functional genetic screening resource for essential genes in human cancer cell lines. *Nucleic Acids Research*, 40(D1):D957–D963, 2012.

[108] M E Hillenmeyer, E Fung, J Wildenhain, et al. The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. *Science*, 320(5874):362–365, 2008.

[109] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh. Mckusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*, 37(Database issue):D793–6, 2009.

[110] L M Schriml, C Arze, S Nadendla, et al. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res*, 40(Database issue):D940–6, 2012.

[111] T R Hvidsten, J Komorowski, A K Sandvik, and A Laegreid. Predicting gene function from gene expressions and ontologies. In *Proceedings of Pacific Symposium on Biocomputing*, pages 299–310, 2001.

[112] X Zhou, M C J Kao, and W H Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12783–12788, 2002.

[113] Y Huang, H Li, H Hu, X Yan, M S Waterman, H Huang, and X J Zhou. Systematic discovery of functional modules and context-specific functional annotation of human genome. *Bioinformatics*, 23(13):i222–i229, 2007.

[114] A del Pozo, F Pazos, and A Valencia. Defining functional distances over gene ontology. *BMC Bioinformatics*, 9:50, 2008.

[115] M Punta, PC Coggill, RY Eberhardt, J Mistry, J Tate, C Boursnell, N Pang, K Forslund, G Ceric, J Clements, A Heger, L Holm, ELL Sonnhammer, SR Eddy, A Bateman, and RD Finn. The Pfam protein families database. *Nucleic Acids Research*, 40:D290–D301, 2012.

[116] A I Su, T Wiltshire, S Batalov, H Lapp, K A Ching, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, 2004.

[117] J Kelso, J Visagie, G Theiler, et al. eVOC: A controlled vocabulary for unifying gene expression data. *Genome Research*, 13(6a):1222–1230, 2003.

[118] Rama Balakrishnan, Julie Park, Kalpana Karra, Benjamin C. Hitz, Gail Binkley, Eurie L. Hong, Julie Sullivan, Gos Micklem, and J. Michael Cherry. YeastMine—an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit. *Database*, 2012, 2012.

[119] W-H Chen, P Minguez, M J Lercher, and P Bork. OGEE: an online gene essentiality database. *Nucleic Acids Research*, 40(D1):D901–D906, 2012.

[120] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk. Biomart–biological queries made easy. *BMC Genomics*, 10:22, 2009.

[121] S E Celniker, L A L Dillon, M B Gerstein, K C Gunsalus, S Henikoff, et al. Unlocking the secrets of the genome. *Nature*, 459:927–930, 2009.

[122] S Tweedie, H Ashburner, K Falls, others, and The FlyBase Consortium. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Research*, 37:D555–D559, 2009.

[123] W M Gelbart and D B Emmert. FlyBase High Throughput Expression Pattern Data Beta Version. FlyBase analysis FBrf0212041 (2010.10.13) at `http://flybase.org/reports/FBrf0212041.html`, 2010.

[124] M Schmid, T S Davison, S R Henz, et al. A gene expression map of Arabidopsis development. *Nature Genetics*, 37:501–506, 2005.

[125] J Kilian, D Whitehead, J Horak, et al. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant Journal*, 50:347–363, 2007.

[126] D P Sangurdekar, F Srienc, and A B Khodursky. A classification based framework for quantitative description of large-scale microarray data. *Genome Biology*, 7:R32, 2006.

[127] J J Faith, B Hayete, J T Thaden, I Mogno, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):e8, 2007.