Supplementary Text

1. Selection of proteins and interactions

Protein functional annotations were obtained from the Gene Ontology (GO) Consortium (Ashburner et al., 2000). We used both GO terms and GO-Slims in the Biological Process ontology. The complete yeast GO annotation as of 05/25/04 was obtained by FTP from the Gene Ontology Consortium at ftp://ftp.geneontology.org/pub/go/gene-associations/gene_association.sgd.gz. A mapping of yeast GO process terms to GO-Slim categories as of 05/18/04 was obtained by FTP

from the Saccharomyces Genome Database at ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/. At the time of our classification, there were 1,002 Cellular Process GO terms and 33 Cellular Process GO-Slims annotated to yeast. We used Perl scripts to import, parse and store the annotations in a Microsoft SQL Server database.

Each annotation term for a gene product is accompanied by one or more "evidence codes" that indicate the origin of the data supporting the annotation (definitions available on the GO website at http://www.geneontology.org/GO.evidence.shtml). Annotations evidenced solely by the evidence codes "Inferred from Physical Interaction" (IPI) and "Inferred from Genetic Interaction" (IGI) were treated separately in order to avoid circularity from overlap with interaction data input for this study.

The set of 1,002 terms in the GO yeast cellular process annotation was screened to identify 85 terms pertinent to gene expression. These were grouped into nine categories corresponding to sub-processes of gene expression (Suppl. Table 1a) using a keyword-identification SQL query to search GO term definitions and descriptions. Terms specific to PolI or PolIII transcription or to the processing of RNAs other than mRNA were excluded. Sub-process #1, transcription initiation, encompassed the most GO terms because it included any annotation related to the regulation of transcription (positive or negative) and chromatin modification. A tenth category was added for terms related to mRNA (such as "mRNA localization") not directly involved with any of the previously defined sub-processes, yielding 107 total terms. An eleventh category was made for terms corresponding to all other cellular roles. A twelfth category included the term "cellular_process_unknown."

The set of yeast proteins used in this study was selected to include only those with GO-annotated roles in gene expression or their direct interaction partners. First, 980 gene products annotated to one of the 85 select GO terms in the Categories #1-10 with an evidence code other than IPI/IGI were identified as the core network (Suppl. Table 1a). The most numerous group of proteins corresponded to Category #1, because it includes all basal transcription factors, transcriptional activators and repressors, and chromatin-modifying factors.

Second, the network was expanded to include all proteins in Categories #11 and #12 ("other cell roles" and "unknown") shown to interact with one of the core 980 proteins based on evidence in the interaction database at the Munich Information Center for Protein Sequences (MIPS) (Mewes et al., 2002). We did this in order to gain insight into the potential gene expression-associated roles of previously uncharacterized ORFs, and proteins currently annotated in GO to other cell functions. With similar aim, we include proteins regardless of annotated localization in order to enable identification of potential shuttling proteins that had not previously been reported in the nucleus.

Third, the 267 proteins annotated to Categories #1-10 only by the IPI/IGI evidence codes were grouped into a separate category, Category #13, and added to the final network to arrive at 2100 members.

Supplementary Table 1. Functional classification of proteins in the network. (a) Categories corresponding to sub-processes of gene expression. Shown are the number of GO terms mapped to the category and represented by at least one protein in the network, and the number of proteins included in the network based on their membership in the category. Protein counts total more than 2,100 because some proteins have multiple annotations. The GO-term count for Category #11 marked by (*) indicates "other," non-gene expression GO term annotations in the yeast proteome and represented among proteins in our network, respectively. Categories #1-10 are used for functional enrichment calculations shown in Fig. 2a. (b) GO-Slim categories. Shown are the number of proteins in the network annotated to each GO-Slim. GO-Slim categories were used to evaluate clusters formed when different values of *k*, RDQ scores, and CC calculations are used (see Suppl. 11).

2. Collection of protein interaction data

Complete yeast protein interaction data were obtained from sources described below in more detail. Data from each raw dataset was imported and stored in a SQL Server database as a binary list of interacting proteins. SQL queries were used to filter each dataset to include only pairwise interactions among our subset of 2,100 proteins. Among these 2,100 proteins, 42,666 pairs were linked by an interaction in at least one dataset.

High-throughput screens

Sources 1-3 and 8-13 were obtained from the Munich Protein Information Server (MIPS) interaction database (Mewes et al., 2002) by FTP at ftp://ftpmips.gsf.de/yeast/PPI/ on 5/24/04. Sources 1-3 were culled from other MIPS entries by PubMed ID.

1. Ito-core Y2H. The set of interactions demonstrated and triply verified by high-throughput yeast two-hybrid screen (Ito et al., 2001). PubMed ID: 11283351.
2. Ito-full Y2H. The set of interactions demonstrated by high-throughput yeast two-hybrid screen, excluding the triply-verified interactions already included in (1) (Ito et al., 2001). PubMed ID: 11283351.
3. Uetz Y2H. The set of interactions demonstrated by high-throughput yeast-two-hybrid screen (Uetz et al., 2000). PubMed ID: 10688190.
4. Complex. Data from bait-prey complex precipitation experiments (Gavin et al., 2002); (Ho et al., 2002) were downloaded by FTP from MIPS at ftp://ftpmips.gsf.de/yeast/catalogues/complexes/ on 5/24/04. SQL queries were used to transform the complex data into binary interaction data. Following the example of previous analysis (Bader and Hogue, 2002), data was interpreted according to the "spoke" model to include only links between bait proteins and each of their co-precipitated preys, rather than between each protein in a precipitated complex.

In-silico predictions

5. Rosetta fusion. This dataset assigns a link to any pair of proteins for which a single "fusion protein" gene exists in another organism on the assumption that a component of the protein machinery may evolve either as a single protein or as a pair of interacting proteins (Enright et al., 1999). Downloaded from http://predictome.bu.edu/ on 5/06/03.
6. Paralog. This dataset implies an interaction between gene products and paralogs of their known interactors (Deane et al., 2002). The set of yeast protein paralogs used in previous studies was obtained (L. Salwinski, personal communication, 4/23/2003) and used in conjunction with protein interaction data in the MIPS interaction set as of 5/24/04
7. Phylogenetic. This dataset assigns a link to any pair of proteins that are co-inherited across the proteomes of many different organisms (Deane et al., 2002; Enright et al., 1999; Marcotte et al., 1999; Wu et al., 2003). The set of yeast gene pairs with significantly similar phylogenetic profiles according to the Clusters of Orthologous Genes (COG) database provided by the National Center for Biotechnology Information (Tatusov et al., 2001) was downloaded from http://predictome.bu.edu/ on 5/05/03.

Hypothesis-driven experimental results
Sources 8-12 represent results from individual experiments deposited into the MIPS repository by individual research groups. SQL queries were used to categorize data according to the description of the experiment performed to demonstrate the interaction.
8. MIPS-affinity. Affinity column precipitation, affinity chromatography.
9. MIPS-co-precipitation. Co-immunoprecipitation, GST pull-down.
10. MIPS-co-purification. Co-sedimentation, in-vitro binding assay, gel shift assay, crosslink.
11. MIPS-synthetic. Synthetic lethal, suppression of mutant phenotype
12. MIPS-Y2H. Non-high-throughput yeast-two-hybrid experiments. The high-throughput screens already included in sources 1-3 were excluded from this category.
13. MIPS-other. All other hypothesis-driven experiments in MIPS. This category includes all interactions in MIPS that did not match the search terms used to select sources 8-12.

3. Adjacency matrix representation

Manipulations were performed using MATLAB 6.5 (Mathworks). Data from each dataset were imported from SQL Server databases as adjacency matrices S_i. For any pair of proteins x and y, $S_i(x, y) = n$, where n gives the number of independent interactions between x and y in dataset i. Self-interactions were removed from the adjacency matrix yielding a diagonal of zeros. The complete interaction map S integrating data from all datasets is a weighted sum of the thirteen source matrices: $S = \sum_{i=1}^{n} w_i S_i$, where w_i gives the weight assigned to all interactions in adjacency matrix S_i.

4. Hypergeometric *p*-value

The hypergeometric *p*-value (Tavazoie et al., 1999; Wu et al., 2002), used independently in several different parts of our method, is given by the formula

$$(1) \qquad p = 1 - \sum_{i=0}^{\min((k-1),(C-G+n))} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}$$

.

In general, the *p*-value quantifies the likelihood that out of *G* total elements divided into overlapping groups, a group Gr_C of size *C* will share *k* elements with a group Gr_n of size *n.* In other words, the *p*-value quantifies the extent to which Gr_C is enriched or dis-enriched in members of Gr_n, with *p*-values close to zero representing statistically significant enrichment or dis-enrichment. In our applications of *p*-value, we wished to reflect statistically significant enrichment only. Thus, for each pair of groups, we calculated the expected value *E* of the number of elements shared between the groups. When the actual number of elements shared between the groups was less than the expected value ($k<E$), the calculated *p*-value instead reflected the statistically significant lack of enrichment and was replaced by 1-*p* to correct for this fact. In addition, we omitted applying the Bonferroni correction for multiple independent categories because only relative *p*-values were needed in our applications.

5. Development and comparison of dataset quality calculation algorithms

In contrast to existing methods that attempt to calculate absolute reliability for each dataset based on an independent evaluative metric, we calculate a relative reliability for each set which allows the data itself to dictate the relative dataset weights, free of subjective bias introduced by external evaluation. We call this the relative dataset quality (RDQ) score. We present three novel computational methods, described below, to calculate the RDQ score for each of our thirteen protein interaction datasets.

Our automated RDQ scoring methods rely on quantitative comparison using the extent of pairwise overlap of each dataset with every other. It was desired that the influence of any test set *j* on a tested set *i* be appropriately weighted by its own calculated weight so that poor quality datasets would not penalize other sets for a lack of overlap with them. We approached this inductive requirement by applying a fixpoint solution. We accomplished this by solving the equation $\mathbf{M}\mathbf{X}_R = \boldsymbol{\lambda}_R\mathbf{X}_R$, where **M** is the matrix such that **M**(*g*, *h*) gives the extent of overlap between datasets *g* and *h*, and $\boldsymbol{\lambda}_R$ and $\mathbf{X}_R$ are the dominant eigenvalue and the corresponding (principal) eigenvector, respectively, of **M.** The existence of $\boldsymbol{\lambda}_R$ and $\mathbf{X}_R$, where $\boldsymbol{\lambda}_R$ has multiplicity one and $\mathbf{X}_R$ is non-negative, is guaranteed by the Perron-Frobenius theorem for all non-negative, irreducible square matrices **M**. The matrix **M** is irreducible if it is strongly connected (Meyer, 2000), i.e. if there exists a path of nonzero overlaps from any dataset *g* to any other dataset *h*. This was the case for our 13 datasets. The RDQ values thus obtained are given by the entries in the vector $\mathbf{X}_R$.

Three different measures of pairwise dataset overlap were developed, yielding three variants of the overlap matrix **R**. The first measure, called "percent-covering," defines $\mathbf{R}_{\text{percent-covering}}(g, h)$ as the percentage of links in dataset *h* included in, or "covered by," dataset *g*. Thus, it penalizes datasets *g* for false negatives. This approach assigns a

high quality for comprehensive datasets *g* that cover other datasets *h* extensively. However, it does not discriminate against large, noisy datasets that include data contained in other datasets but also include a large number of false positives. The second, "percent-covered" measure addresses this problem by defining $\mathbf{R}_{\text{percent-covered}}(g, h)$ as the percentage of dataset *g* covered by dataset *h*. Thus, it assigns the quality of *g* in terms of the extent to which its data is corroborated in other datasets, penalizing false positives. However, it does not penalize for false negatives. Note that $\mathbf{R}_{\text{percent-covered}}$ is simply the transpose of $\mathbf{R}_{\text{percent-covering}}$. The third, "*p*-value" measure attempts to address both false positives and negatives, defining $\mathbf{R}_{p\text{-value}} = -1/\log \mathbf{P}$, where $\mathbf{P}(g, h)$ gives the nonzero hypergeometric *p*-value of overlap between datasets *g* and *h*. If $\mathbf{P}=0$, we define $\mathbf{R}_{p\text{-value}}= 0$; if $\mathbf{P}=1$, we define $\mathbf{R}_{p\text{-value}}= 1$. Here, the *p*-value gives the likelihood that the overlap of two datasets, or the number of links appearing in both, is greater than that expected by chance. The *p*-value is obtained using Formula (1), with G = the number of possible pairwise protein interactions between all proteins in our network, C = the size of the dataset being tested (the number of distinct links it contains), n = the size of the dataset tested against, and k = the number of links they have in common.

The scores assigned to the datasets using each of the RDQ scoring methods are listed in Suppl. Table 2.

Supplementary Table 2: RDQ scores and properties of protein interaction datasets used in this study. Dataset (1) is a high-confidence dataset based on a high-throughput assay. (2-4) are lower-confidence high-throughput datasets. Datasets (5-7) are generated by *in silico* predictions. Datasets (8-13) include all the interactions verified by small-scale experiments and deposited in the MIPS database, categorized by experimental method. Columns C-E indicate calculated RDQ scores, scaled to sum to 1 over all datasets. Columns F-H indicate independent dataset properties described in Suppl. 7. Columns I-M indicate the number of distinct (i.e. without repeating the same pair of proteins within the same dataset) and reiterated (i.e. the number of times the dataset repeated a listed interaction, including swapping the order in which the proteins were reported) interactions in each dataset and the counts of total and essential proteins involved.

6. Comparing RDQ methods based on FP/FN rate prediction

To assess the ability of RDQ measurement methods to correctly penalize datasets with noisy data, a series of trials were performed to calculate the RDQ score of datasets in which a known proportion of interactions were corrupted. Data was corrupted in three ways. First, links were removed at random to simulate false negatives. Second, spurious links were added at random to simulate false positives. Third, the network was scrambled by rewiring one member of a pair of linked nodes to a random new neighbor. To maintain connectivity properties of each network upon corruption, links were added and removed with a probability proportional to node degree.

Results are shown in Suppl. Fig. 1. RDQ calculations using the "percent-covering" and the "percent-covered" overlap matrices both penalize rewired links. The former implementation also penalized false negatives, while the latter penalized false

positives. To our surprise, we found that RDQ scores generated using the "*p*-value" overlap matrix did not correlate with any type of dataset corruption.

7. Comparing RDQ methods based on independent dataset quality indicators

RDQ score calculations based on the three different dataset overlap matrices were compared against two independent dataset metrics. The functional conservation metric indicates the proportion of interactions in a dataset linking proteins annotated to the same GO process term (Suppl. Fig. 2a). The functional conservation of each dataset was plotted against the rank of that dataset when using each of the three RDQ calculations. While RDQ calculations using the "percent-covered" overlap matrix yielded rankings that correlated with functional conservation values, the "percent-covering" and "*p*-value" overlap matrices did not.

The second, network saturation dataset metric is calculated by the global clustering coefficient (CC) of the network. This value reflects the average number of binding partners per protein. High network saturation may indicate "sloppy" datasets with a high number of false positives that over-report the true amount of protein-protein interactions in yeast. Oversaturated networks, furthermore, may demonstrate excellent functional conservation statistics because of the artificially large number of links between proteins in local regions of the network. In support of our analysis, we found that neither functional conservation scores nor assigned RDQ scores among our datasets were biased by network saturation (Suppl. Fig. 2b).

We chose the RDQ method using the "percent-covered" overlap matrix for further analyses because of its ability to 1) accurately punish large, noisy datasets for false positives (Suppl. Fig. 1) and to 2) correlate functional conservation to assigned RDQ score in a manner unbiased by network saturation (Suppl. Fig. 2).

8. Pairwise clustering coefficient

Protein pairs in densely connected regions of networks are characterized by a high pair-wise clustering coefficient (CC), which quantifies the cohesiveness of a network in a local neighborhood (Goldberg and Roth, 2003). For a pair of proteins, the CC assesses relative connectivity to common and distinct neighbors. Reliably interacting pairs in a small-world network are more likely to share strong links to common neighbors and thus have a higher CC. Under the assumption that a functional protein complex consists of a set of strongly interconnected proteins that link more strongly to each other than to other proteins in the environment, the CC presents a powerful metric towards the identification of network regions that correspond to physical protein complexes.

Previous analysis (Goldberg and Roth, 2003) applied the CC to an interaction network derived from high-throughput yeast-two-hybrid screens in order to identify unlinked yet high-scoring pairs as putative false negatives in the datasets. These and other previous studies using the CC have been limited to unweighted graphs, though its application to weighted graphs has been suggested as a network centrality measure (Wuchty and Stadler, 2003). In the absence of precedent, we developed six variants of a

novel CC formula for use in weighted graphs (Suppl. Fig. 3) and compared them on the basis of their ability to predict common function of a pair of proteins.

Different CC formulas emphasize various aspects of local network density. All variants developed use the strength of the link between a pair of proteins as well the connectivity and strength of the links of each to other proteins in the environment. The existence of a direct link between a pair of proteins is not required to assign the pair a nonzero CC as long as the protein pair shares common neighbors. The CC for a pair of nodes is positively affected by the link between the pair and by links to mutual neighbors (which helps reconstruct links missed due to false negatives). The CC is negatively affected by links to other nodes in the network (to downweight links due to false positives). These positive and negative effects are reflected in the choice of numerator and denominator, respectively.

For a pair of nodes *N* and *N'*, *a* is the weight of the link between them, *b*i and *b*i' are the weights of the *i*-th links of *N* and *N'*, respectively, to common neighbors, and *c*i and *c*i' are the weights to uncommon neighbors. We then define the intermediate terms (see Suppl. Fig. 3 for explanation and illustration)

$T(N) = a + \sum b_i + \sum c_i$, the total weight of links from a single node *N*

$L(N) = a + \sum b_i$, the weight of links from node *N* to common neighbors of *N* and *N'*

$W_1(N, N') = a^2 + \sum (b_i * b_i')$, the dot product of weights to common neighbors, deemphasizing the mutual link (a<1)

$W_2(N, N') = a + \sum (b_i * b_i')$, the dot product of weights to common neighbors, emphasizing the mutual link

And define the CC formulas as follows:

$$CC_1(N,N') = \frac{L(N)+L(N')}{T(N)+T(N')}$$

$$CC_2(N,N') = \frac{W_1(N,N')}{T(N)+T(N')}$$

$$CC_3(N,N') = \frac{L(N)+L(N')}{T(N)*T(N')}$$

$$CC_4(N,N') = \frac{W_1(N,N')}{T(N)*T(N')}$$

$$CC_5(N,N') = \frac{W_2(N,N')}{T(N)+T(N')}$$

$$CC_6(N,N') = \frac{W_2(N,N')}{T(N)*T(N')}$$

The denominator of the CC accounts for all the links of a pair of proteins both in and outside their mutual neighborhood; it is meant to penalize a large number of interactions outside this neighborhood. The denominator of CC1, CC2 and CC5 is the sum of all interaction weights for both proteins in the pair; CC3, CC4 and CC6 use the product of the sum of interactions for each of the two proteins. Use of the sum is motivated by the Jaccard index and the product by the Meet/min coefficient used in previous definitions of the CC (Goldberg and Roth, 2003). Using the sum, however, carries the risk of being skewed by one member of the pair; we desire to punish a pair's CC score to a greater extent if both proteins have many links (are noisy binders). The product generates a smaller denominator, penalizing the CC less when only one of the proteins is a significantly more selective binder.

For each pair of proteins, the numerator of the CC is computed as follows. CC1 and CC3 use the sum of all weighted links to common neighbor proteins and twice the weight of their mutual link. This simple summation does not take into account the distribution of strong links to the same common neighbor. Formulas CC2 and CC4 sum the products of weights to each of their common neighbors and square the mutual link weight. Since the value of the mutual link is less than zero, taking the square diminishes its contribution. CC5 and CC6 use the same numerator as CC4 and CC5 but do not square the mutual link weight, thus emphasizing it.

9. Comparison of CC formulas based on prediction of common function

We selected a single one of our 6 CC formulas based on its ability to distinguish among links between functionally related and unrelated proteins, as determined by GO annotation. We were motivated by previous analysis of the interactome of *C. elegans* which notes that pairwise CC is a good predictor of functional relation, since it reflects sharing of common interacting partners (Li et al., 2004). In a previous study in *Drosophila*, the correlation with similar GO annotation of interacting proteins was used to analyze confidence scores in a protein interaction network (Giot et al., 2003).

For each CC formula, logistic regression was used to correlate the calculated CC between each pair of proteins with the probability that the two proteins were annotated to the same GO process term. Pairwise scores given by the raw, RDQ-weighted network were evaluated as well. Logistic regression analysis was performed using the glm function in the R statistics package version 1.9.1 to obtain correlation *z*-values (Suppl. Table 3). For a more detailed description of the implementation and usage of the glm function, readers are referred to documentation provided by The R Foundation for Statistical Computing. Of the seven scoring systems considered, CC1 displayed the best correlation and was chosen for further analysis.

	z-value
RDQ only	27.80
CC1	81.92
CC2	24.39
CC3	1.98
CC4	1.73
CC5	38.81
CC6	15.22

Supplementary Table 3: Comparison of protein pair scores in the raw RDQ-weighted network and scores calculated by six different CC formulas, based on ability to predict common function. We used logistic regression to relate scores in the RDQ-weighted network (“*RDQ only”*) and CC scores (*CC1-CC6*) to the probability that protein pair members share annotation to a common GO term. The *z*-value is shown; higher values indicate a stronger correlation. Since CC1 demonstrates the strongest correlation, this formula was chosen for subsequent analysis.

10. Cluster identification: *k*-means clustering algorithm

For a given number of clusters *k*, the *k*-means clustering algorithm (Hartigan, 1975) seeks to find the optimum partition of a network into *k* non-overlapping clusters based on the sum of cluster scores over the entire network. Scores assigned to each cluster are based on the members of the cluster and the identity of the node deterministically chosen as the cluster’s “centroid.” Any centroid-based metric that makes use of pairwise scores between proteins may be used as long as convergence to a local optimum is guaranteed.

The *k*-means algorithm is typically used to cluster large data sets in Euclidean space into *k* distinct regions, sometimes known as Voronoi cells, as follows. 1) *k* initial centroids are selected from among all the data points. 2) Holding centroids constant, each data point is assigned to the centroid closest to it by the Euclidean distance metric, forming *k* distinct clusters. 3) For each cluster, membership is held constant while one member is selected as the new centroid so as to minimize the cluster score, defined as the sum of squared distances of cluster members to the (new) centroid. Steps 2-3 in the procedure are iterated until no more nodes or centroids are re-assigned. The procedure always converges, and the final partition is a local minima for the overall sum of cluster scores.

We used an extension of the above method to cluster non-Euclidean data by using a weight metric as opposed to a distance metric and sought to maximize, rather than minimize, cluster scores. We used pairwise CC scores as the weight metric. Each node in the network was assigned to the centroid to which it had the greatest link weight, and each centroid was chosen to maximize the cluster score. The total network score was defined as the sum of cluster scores $\sum_{i=1}^{k}\sum_{j=1}^{r_i} CC(c_i, n_{i_j})$ where *k* is the number of clusters, r_i, c_i, and n_{i_j} are the size, centroid, and *j*-th member, respectively, of the *i*-th cluster, and *CC(x,y)* gives the pairwise CC between nodes *x* and *y*. Whenever a node has a CC score of zero with all of the current centroids, it is assigned to the centroid separated from it by the shortest path along links in the interaction network. Shortest path traversals were precomputed using Dijkstra's algorithm (Cormen, 2001). To increase the running speed of the algorithm, convergence of total network score to within four digits was used in place of absolute convergence.

Past approaches to protein interaction network clustering have used graphs with unweighted edges. In an investigation of the general yeast interaction network from a variety of diverse experimental sources, a previous study (Samanta and Liang, 2003) used an approach similar to our use of the CC to group proteins sharing a statistically significant number of protein node neighbors. While our method predetermines the number of clusters and samples over random cluster center seeds, however, their method iteratively aggregates proteins to clusters or to other single proteins and the final number of clusters is thus dynamically determined. Another study (Spirin and Mirny, 2003) clustered the yeast MIPS interaction network using three different techniques to find locally dense subgraphs corresponding to structural complexes or functional modules. First, clique identification simply searches for fully connected subgraphs. Superparamagnetic clustering, second, allows individual proteins to take on one of several allowable "spin" states that is affected by the spins of its protein interactors, leading to a dynamic equilibrium. The presumption behind the model is that identifying groups of proteins with aligned spins will indicate highly interconnected protein clusters. Monte Carlo optimization, third, is a randomly seeded method to find clusters corresponding to a given number of proteins with the maximal number of interactions between them. The "Molecular Complex Detection" (MCODE) algorithm (Bader and Hogue, 2003) likewise relies on a search for maximally connected subgraphs in the

network topology. MCODE uses a variant of the clustering coefficient to weigh nodes, then iteratively and conditionally adds nodes to highest-weighted center nodes; it also includes the ability to fine-tune individual clusters of interest. This method was used to cluster the yeast interactome using data from MIPS and from protein complex identification. Our use of weighted graphs represents a novel and significant addition to the existing body of protein interaction network clustering research.

11. Selection of clustering parameters based on ability to consistently generate clusters enriched in functionally related proteins

The total network score, used to select an optimum partition for a given number of centroids *k*, may not be used to determine an optimum value for *k* because the score function is biased towards smaller clusters in a non-trivial way that depends on the weight metric. We therefore performed the following analysis to choose a value for *k*; to confirm our choice of using calculated RDQ scores to define the network clustered; and to test the decision to cluster based on pairwise CC scores derived from this network. We evaluated the performance of clustering based on the consistent generation of biologically interpretable clusters, using a range of possible combinations of these parameters.

First, we have sampled a representative range of values for *k* which allow an average of 10, 20, or 30 proteins per cluster, corresponding to $k = 210$, 105, and 70, respectively. Second, we used the network with datasets integrated using RDQ scores calculated using the "percent-covered" overlap matrix, as well as two additional networks constructed using two control RDQ score sets, C1 and C2. C1 simply assigns an RDQ of 1 to each dataset. C2 reflects intuitive assessments of dataset quality by assigning an RDQ of 1 to MIPS-derived datasets 8-13 (Suppl. Table 2), an RDQ of 0.8 to the high-confidence high-throughput dataset 1, an RDQ of 0.6 to the standard high-throughput datasets 2-4, and an RDQ of 0.3 to the computationally-derived datasets 5-7. Third, we based our *k*-means clustering on either raw weights in the RDQ-derived protein interaction network, or on the pairwise CC scores derived from this network. Eighteen parameter sets (3 choices of *k*, 3 choices of RDQ, and 2 choices of CC or raw network scores) for the clustering program were thus considered. Each parameter set was used to cluster the network 70 times, with random initial centroid seedings each time, to generate 70 sets of *k* clusters.

We next quantified the extent to which use of each parameter set produced biologically interpretable cluster sets. For each parameter set, the hypergeometric *p*-value was used to quantify the enrichment of each of the 70*k* clusters in proteins annotated to each of the GO-Slim terms (Suppl. Table 1b). This test compensates for cluster size and is not inherently biased by *k.* The *p*-value is obtained using Formula (1), with G = the total number of proteins in the network, C = the number of proteins in the cluster, n = the number of proteins annotated to the tested GO-Slim category, and k = the number of proteins in the cluster annotated to that GO-Slim category. Smaller *p*-values indicate more significant functional enrichment. The best (smallest) *p*-value of enrichment in any GO-Slim category corresponds to the strength of annotation to its most likely biological function. For each parameter set, we thus found 70*k* best *p*-values.

We wished to determine which parameter set generated clusters more significantly enriched in function than random clusters of the same size. Each cluster set was randomized by randomly reassigning proteins to clusters while maintaining cluster size and the best *p*-value of enrichment in a GO-Slim category was calculated as well. To compare *p*-values obtained using real and random clusters, we used the Wilcoxon and Kolmogorov-Smirnov tests in MATLAB. For a more detailed description of the implementation and usage of the Wilcoxon (ranksum) and Kolmogorov-Smirnov (kstest) tests, readers are referred to documentation provided by Mathworks.

In detail, for each parameter set, the highly non-normal distributions of the 70*k* real and 70*k* randomly-derived *p*-values were compared by applying the Wilcoxon rank-sum test. Results indicated a highly shift toward more significant *p*-values of functional enrichment in real versus random cluster sets for all eighteen parameter sets, but results were indistinguishable among parameter sets *(data not shown)*. We thus applied the more sensitive Kolmogorov-Smirnov test, which was able to discriminate between parameter sets. Results are shown in Suppl. Table 4. The cluster sets that displayed the most significant enrichments over random clusters were generated when k=70, the network was weighted by RDQs calculated using the "percent-covered" overlap matrix, and clustering was based on CC scores derived from this network.

	RDQ	*k*=70	*k*=105	*k*=210
derived CC scores	"percent-covered"	**0.4097**	0.2126	0.1664
	C1 assigned	0.2237	0.2053	0.1637
	C2 assigned	0.235	0.2059	0.1582
RDQ-weighted network scores	"percent-covered"	0.2819	0.2303	0.1831
	C1 assigned	0.2649	0.2365	0.186
	C2 assigned	0.2811	0.2423	0.1898

Supplementary Table 4: Comparison of parameter sets for the clustering program based on generation of clusters significantly enriched in function as compared to random clusters. The nonparametric Kolmogorov-Smirnov test is used to measure the shift in distribution towards more significant *p*-values of enrichment of clusters generated in 70 independent clusterings, over random clusters of the same size. Shown are the Kolmogorov-Smirnov statistics reported by the test, with greater values representing greater enrichments over random. Columns indicate choices for the parameter *k*, the number of clusters. Rows indicate choices for the method of assigning RDQ scores, including the set of RDQ scores calculated using the "percent-covered" dataset overlap matrix as well as two control sets of assigned RDQ scores. The top panel indicates clustering based on network-derived CC scores, while the bottom panel indicates clustering based on the raw RDQ-weighted network. The most significant functional enrichment was observed when k=70, the clustered network was weighted by RDQs calculated the "percent-covered" overlap matrix, and clustering was based on CC scores derived from this network.

The parameter set chosen above includes k=70, which corresponds to an average cluster size of 30 proteins, out of average cluster sizes of 10, 20, and 30 surveyed. To ensure that larger clusters would not generate still better results, we also tested an average cluster size of 40, corresponding to k=52, while holding all other parameters constant ("percent-covered" RDQ and derived CC scores). Applying the Kolmogorov-Smirnov test as above, we obtained a value of 0.26238. This indicates that clusters of average size 30 generated by our method demonstrate more significant functional enrichment than either smaller (top row, Supp. Table 4) or larger clusters, and confirms the choice of k=70.

12. Cluster set

The k-means clustering algorithm with k=70 was used as described in (10), using CC scores derived from the "percent-covered" RDQ-weighted network, to generate a single, optimum set of 70 clusters for further analysis (Suppl. Tables 5-6).

The hypergeometric p-value was used to quantify the enrichment of each cluster in proteins annotated to each of the sub-processes of gene expression (Suppl. Table 1a, Fig. 2a(iii)). The p-value is obtained using Formula (1), with G = the total number of proteins in the network, C = the number of proteins in the cluster, n = the number of proteins annotated to the tested category, and k = the number of proteins in the cluster annotated to that sub-process. Smaller p-values indicate more significant functional enrichment.

Supplementary Table 5. Assignment of proteins to clusters. The number of proteins in the cluster is noted in parentheses. Proteins are hyperlinked to the corresponding "Protein Page" maintained by the Saccharomyces Genome Database (Dwight et al., 2004).

Supplementary Table 6. Characterization of clusters. The number of proteins in the cluster is noted in parentheses. The sub-processes of gene expression (Suppl. Fig. 1a, Fig. 2a(iii)) for which the cluster is significantly enriched (p<0.05) are listed in brackets in order of significance of enrichment.

13. Motif identification and ranking

We automated the identification and ranking of all instances of the three coupling motifs illustrated in Fig. 1c as follows.

First, only direct interaction links appearing in the raw weighted interaction network were candidates for coupling links. Second, only links between distinct complexes were considered. Pairs of clusters are sorted according to a separability score that estimates the degree to which the clusters may be biologically distinct. This consideration helps to identify true instances of coupling from among false "coupled" clusters that may result from over-clustering, i.e., from artificially forcing biologically coherent clusters to split into separate clusters. The cluster separability score was defined

as $\frac{\sum_{r=1}^{s_i} CC(c_i, n_{i_r}) + \sum_{r=1}^{s_j} CC(c_j, n_{i_j})}{\sum_{r=1}^{s_l} CC(c_l, n_{l_r})}$ for a pair of clusters *i* and *j*, where s_i, c_i, and n_{i_r} are the size, centroid, and *r*-th member, respectively, of the *i*-th cluster, and *l* is the cluster formed by merging clusters *i* and *j* and finding its new centroid. Simply, the cluster separability score is the sum of individual cluster scores divided by the score of the merged cluster and is greater for more separable clusters. We considered only cluster pairs ranking in the top 50% by cluster separability score for further analysis.

Direct coupling links are defined by a pair of proteins in different clusters linked by a nonzero-weighted link in the RDQ-weighted interaction network. Each of the coupling proteins in the pair must also be linked to its own cluster significantly more strongly than to the coupled cluster. We assess this by calculating the ratio of the sum of network link weights to all proteins in its own cluster to the sum of link weights to all proteins in the coupled cluster. To qualify as a direct coupling link, this ratio must be greater than 2 for both proteins in the pair. Direct coupling links are ranked by the strength of the network link weight between the coupling protein pair.

Cluster-mediated coupling links are defined by two clusters indirectly coupled by a mediating cluster. At least one direct coupling link must exist between the mediating cluster and each of the coupled clusters. Two additional requirements must be met. First, to ensure link reliability, the weight of these coupling links must be in the top 10% among all direct coupling links in the network. Second, to facilitate identification of clusters dedicated to a coupling role, the mediating cluster must be smaller than either of the coupled clusters (i.e. contain fewer proteins). Instances of cluster-mediated coupling motifs are ranked by topology, with higher-ranking instances minimizing the total number of 1) top-10% direct coupling links between the coupled clusters (to indicate that a mediating cluster is necessary), and 2) top-10% direct coupling links between the mediating cluster and other clusters not in the motif (to indicate the dedicated coupling role of the mediating cluster).

Adaptor-mediated coupling links are defined by an adaptor protein assigned to one cluster but linked significantly to another cluster as well. To assess this, we use the same ratio as that used to identify direct couplers. In this case, however, we require that linkage of the adaptor protein to the coupled cluster (the sum of network link weights from the adaptor protein to proteins in the coupled cluster) must be at least half as great as its linkage to its own cluster (the sum of network link weights from the adaptor protein to proteins in its own cluster). Adaptor-mediated coupling links are ranked by the linkage of the adaptor protein to the coupled cluster as defined above. In our definition of the adaptor-coupled motif, we were motivated by previous work implicating an important role for individual, low-degree proteins in the coupling of complexes defined by highly binding “hub” proteins (Maslov and Sneppen, 2002).

Ranked lists of the top 25 ranked motifs of each motif pattern in our network are provided in Suppl. Tables 7-9. A higher-order graph to visualize the protein composition and top 25-ranked coupling motifs between protein clusters is provided in Suppl. Fig 4.

Supplementary Table 7. Top 25 instances of the direct coupling motif. Listed are the clusters and specific proteins involved in each link, as well as the score used for ranking.

Supplementary Table 8. Top 25 instances of the cluster-mediated coupling motif. Listed are the clusters and specific proteins involved in each link, as well as the score used for ranking.

Supplementary Table 9. Top 25 instances of the adaptor-mediated coupling motif. Listed are the clusters and adaptor protein involved in each link, as well as the score used for ranking.

14. Robustness of motif identification and ranking in suboptimal cluster runs

We next investigated the sensitivity of coupling motif identification to the selection of a particular, optimum run of *k*-means clustering (with a particular set of random cluster centroids, compared by total graph score). We compared motifs found within the optimum cluster set to those found within each of the 69 inferior cluster sets generated using the chosen parameters {*k=70*, RDQ is "percent-covered," and applying the CC formula}. For each inferior cluster set we: 1.) found and ranked all direct-coupling motifs within each of these cluster sets, and 2.) generated 50 randomized cluster sets with corresponding cluster sizes, and found and ranked all direct-coupling motifs within these. Of motifs found in each inferior and randomized cluster set, we determined how many were also found in the optimal cluster set. We found that approximately 20% of motifs found in sub-optimal cluster sets were conserved within the optimal cluster set, on average, a value over 400 times higher than that obtained using suboptimal cluster sets clusterings (Suppl. Table 10, "Identification conserved"). Approximately 10-20% of motifs were furthermore found in the same top percentile of ranked motifs in the suboptimal and optimal cluster sets, a value 75 to over 2000 times higher than expected in random clusterings (Suppl. Table 10, "Identification and rank conserved"). While the comparison of cluster sets and motifs is an active area of research (Hart et al., 2005) and is not the main focus of this paper, this test allows a cursory assessment of robustness of the final results of our method.

		Percentile motif rank 1	5	10	25	50	75	100
a. Identification conserved	Percent conserved	16.89	17.40	16.23	19.31	20.23	20.23	20.20
	Enrichment over random	278.47	488.39	456.62	472.35	462.48	462.48	460.45
b. Identification and rank	Percent conserved	11.34	14.59	16.20	19.08	20.23	20.23	20.20

conserved								
	Enrichment over random	75.00	973.53	2228.60	1926.11	472.82	472.82	460.45

Supplementary Table 10. Conservation of direct coupling motif identification and ranking in sub-optimal cluster sets. 69 cluster sets with a lower total graph score than the optimal cluster set were analyzed. Direct coupling motifs within the indicated top percentiles (columns) were identified in each of the suboptimal cluster sets and in 50 randomized cluster sets with cluster size conserved for each of the suboptimal sets. **a)** The average percentage of motifs found in the suboptimal cluster set also identified in the optimal cluster set is shown ("*percent conserved*"). We also report the fold enrichment of this percentage over that obtained using the randomized cluster sets ("*enrichment over random*"). **b)** Analysis was done as in a), but conservation statistics were only reported for motifs further found in the same top percentile rank (columns) in the optimal cluster set as in the suboptimal and randomized cluster sets. (Further data available upon request from KM).

15. Experimental validation

A comprehensive physical interaction dataset that focused on soluble proteins, where proteins were purified using TAP affinity tags (calmodulin-binding peptide plus protein A) (Rigaut et al., 1999), was also used as validation. Following affinity purification on IgG and calmodulin columns, components of the resulting highly-purified protein complexes were identified by two methods in this study: 1) direct analysis of the purified material by trypsin digestion followed by shotgun sequencing using high performance capillary-scale liquid chromatography-tandem mass spectrometry (LC-MS/MS) (Cagney and Emili, 2002) and 2) excision of silver-stained gel bands after SDS-PAGE followed by trypsin digestion and MALDI-TOF mass spectrometry (Krogan et al., 2002).

Each peptide mass spectrum is subjected to subsequent computational analysis for accurate protein identification. A confidence score for each protein pair is derived for the entire set of purified complexes. Interactions in the LCMS dataset are quantified by a calculated percent reliability, while interactions in the MALDI dataset are quantified by a quality index ranging from 0 to above 30, with values above 0.5 considered to indicate plausible interactions and values above 1.0 considered to indicate highly likely interactions (Tikuisis et al., unpublished data). Data from the independent LCMS and MALDI datasets were downloaded to a SQL Server database. Data were ported to MATLAB as described in Suppl. 2, additionally imposing minimum cutoff thresholds of 50%, 70%, 90%, or 98% on the LCMS data and 0.5, 0.7, 0.9, or 1.0 on the MALDI data. This yielded eight different validation datasets.

For each interaction in each validation dataset, we determined whether the interaction corresponds to a 1) direct, 2) cluster-mediated, or 3) adaptor-mediated coupling motif link in the top *x*% of all such coupling links in the network, 4) occurs between proteins within a single cluster, 5) does not define a coupling motif link, but

occurs between clusters already linked by a coupling motif link, or 6) occurs between clusters significantly (p<0.05) annotated to the same sub-process of gene expression. For (2), interactions were said to define an adaptor-mediated coupling link if they occurred between an identified adaptor protein and any protein in the coupled cluster. We sampled values of x= {1, 2.5, 5, 10, 15, 20, 25, 30, 35, 50, 75, 100}. Since coupling motif instances with the same score (the score used to calculate rank, see Suppl. 13) will be assigned consecutive ranks, we define the subset of coupling motif instances ranking in the top x% as those with score better than or equal to the score of the motif instance ranking exactly $\frac{x*n}{100}$, where n= the total number of instances of that motif.

We reported the total counts of interactions in the validation datasets in each of the six categories (Suppl. Table 11a, "Real model").

For each interaction in each validation dataset, we similarly determined whether the interaction falls into categories (1-6) above in a randomized model. For each value of x, we randomized our model of coupled clusters by randomly renaming proteins, keeping clusters and topologies of coupling and other links the same. We compared validation datasets to random models 50 times for each random model, and reported the average counts of interactions in the validation dataset in each of the six categories over the 50 randomized models (Suppl. Table 11b, "Average over 50 randomized models").

The fold enrichment of interactions from the validation datasets in categories (1-6) using our network model versus randomized models was calculated by taking the ratio of the total counts obtained using the real model to the average counts obtained using the corresponding randomized models (Suppl. Table 11c, "Enrichments: real over random"). Category 1 enrichments are shown in Fig. 4. Category 2-4 enrichments are shown in Suppl. Fig. 6.

Supplementary Table 11. Localization of interactions from the independent, validation protein interaction datasets MALDI and LCMS within our network model of coupled, annotated protein clusters. Eight validation datasets (row categories, indicated in leftmost column) were obtained by applying four different cutoff thresholds to each of the datasets (50%, 70%, 90%, and 98% thresholds for LCMS, and 0.5, 0.7, 0.9, and 1.0 cutoffs for MALDI). Only interactions among proteins appearing in our model are considered The number of interactions in each of these eight validation datasets is indicated (leftmost column). For each interaction in each validation dataset, we determined whether the interaction corresponds to a 1) direct, 2) cluster-mediated, or 3) adaptor-mediated coupling motif link in the top x% of all such coupling links in the network, 4) occurs between proteins within a single cluster, or does not define a coupling motif link, but occurs between clusters 5) already linked by a coupling motif link, or 6) significantly (p<0.05) annotated to the same sub-process of gene expression. The second column from left indicates x. **a)** The number of interactions from validation datasets located in categories (1-6) in our model. **b)** The number of interactions from validation datasets located in categories (1-6) in a randomized model, averaged over 50 randomizations. **c)** Fold enrichment of interactions from the validation datasets in categories (1-6) using our

model versus randomized models. (Further data available upon request from KM and NK).

A summary of interactions from the combined source datasets and for each validation dataset is provided in Supplementary Table 12. Here, we provide the number of interactions in each of these datasets that occur between proteins annotated to each functional category within each of the 70 clusters (involving exactly two proteins in the cluster) and crossing each cluster (involving exactly one protein in the cluster).

Supplementary Table 12. Functional annotation of proteins involved in interactions from the combined source datasets (a) and in each validation dataset (b-i). Data is presented for proteins in each of the 70 clusters, further specifying whether the links described are within the cluster (involve exactly two proteins in the cluster) or across the cluster (involve exactly one protein in the cluster). Indices of matrix rows and columns indicate functional categories of proteins involved, as indicated in the Legend. For interactions within clusters, all matrices are upper-triangular. For interactions across clusters, row indices indicate proteins within the cluster while column indices indicate proteins outside the cluster.

References

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

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

Bader, G. D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics *4*.

Cagney, G., and Emili, A. (2002). De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. Nat Biotechnol *20*, 163-170.

Cormen, T. H. (2001). Introduction to algorithms, 2nd edn (Cambridge, Mass., MIT Press).

Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics *1*, 349-356.

Dwight, S. S., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dolinski, K., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J., Hong, E. L., *et al.* (2004). Saccharomyces genome database: underlying principles and organisation. Brief Bioinform *5*, 9-22.

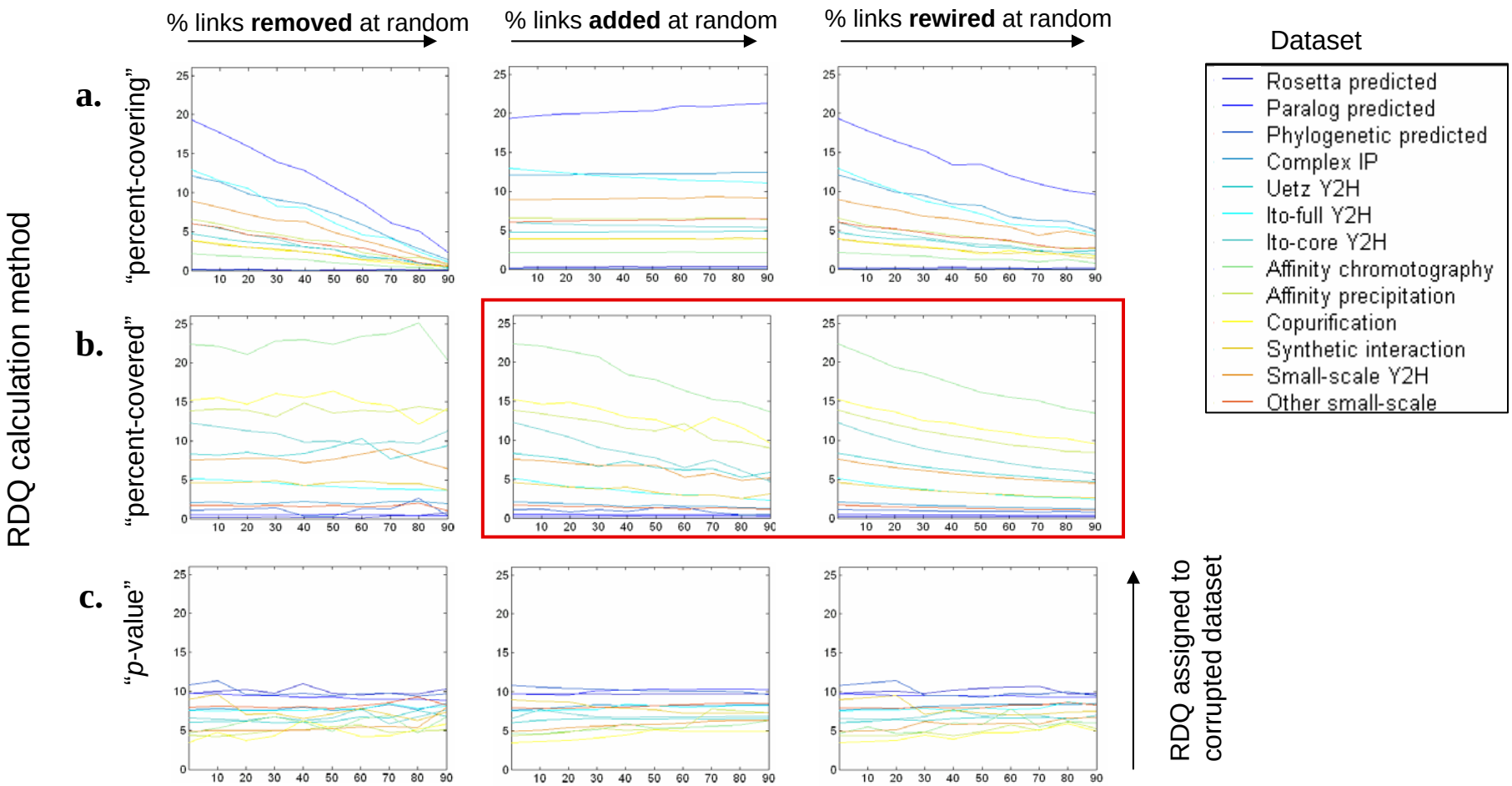
Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402*, 86-90.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141-147.

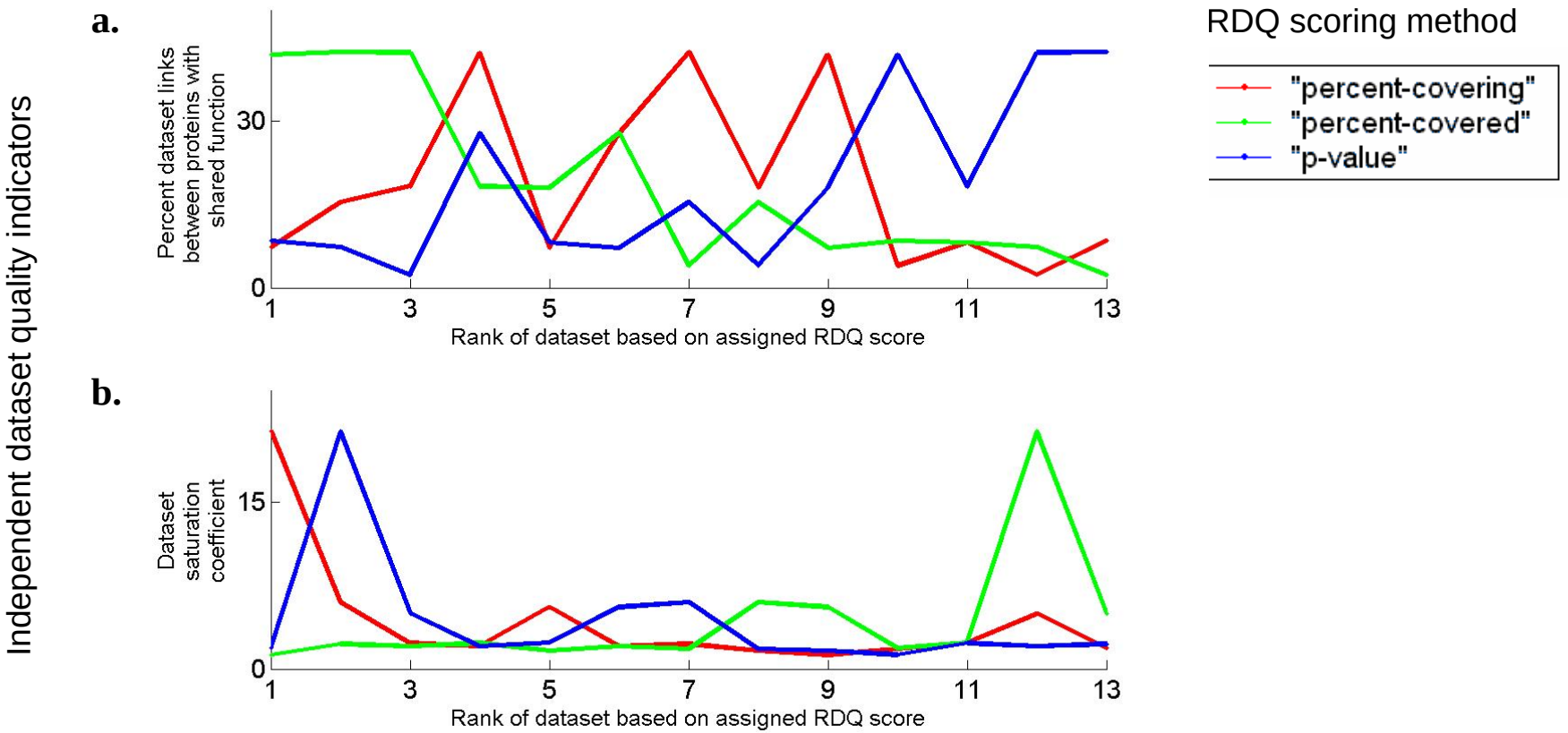
Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., *et al.* (2003). A protein interaction map of Drosophila melanogaster. Science *302*, 1727-1736.

Goldberg, D. S., and Roth, F. P. (2003). Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci U S A *100*, 4372-4376.
Hart, C. E., Sharenbroich, L., Bornstein, B. J., Trout, D., King, B., Mjolsness, E., and Wold, B. J. (2005). A mathematical and computational framework for quantitative comparison and integration of large-scale gene expression data. Nucleic Acids Res *33*, 2580-2594.
Hartigan, J. A. (1975). Clustering algorithms (New York,, Wiley).
Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., *et al.* (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180-183.
Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A *98*, 4569-4574.
Krogan, N. J., Kim, M., Ahn, S. H., Zhong, G., Kobor, M. S., Cagney, G., Emili, A., Shilatifard, A., Buratowski, S., and Greenblatt, J. F. (2002). RNA polymerase II elongation factors of Saccharomyces cerevisiae: a targeted proteomics approach. Mol Cell Biol *22*, 6979-6992.
Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., *et al.* (2004). A map of the interactome network of the metazoan C. elegans. Science *303*, 540-543.
Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science *285*, 751-753.
Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. Science *296*, 910-913.
Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002). MIPS: a database for genomes and protein sequences. Nucleic Acids Res *30*, 31-34.
Meyer, C. D. (2000). Matrix analysis and applied linear algebra (Philadelphia, Society for Industrial and Applied Mathematics).
Samanta, M.P. and Liang, S. (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci U S A*, **100**, 12579-12583.
Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, **100**, 12123-12128.
Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res *29*, 22-28.
Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol *17*, 1030-1032.
Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. Nat Genet *22*, 281-285.
Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-627.

Wu, J., Kasif, S., and DeLisi, C. (2003). Identification of functional links between genes using phylogenetic profiles. Bioinformatics *19*, 1524-1530.
Wu, L. F., Hughes, T. R., Davierwala, A. P., Robinson, M. D., Stoughton, R., and Altschuler, S. J. (2002). Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. Nat Genet *31*, 255-265.
Wuchty, S., and Stadler, P. F. (2003). Centers of complex networks. J Theor Biol *223*, 45-53.

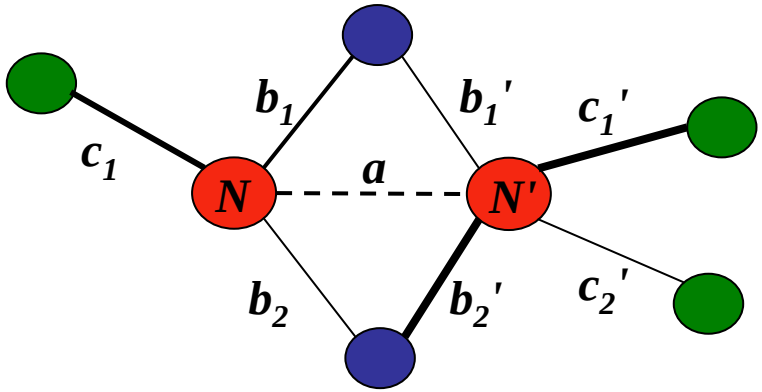


Supplementary Figure 1. Relative protein interaction dataset quality (RDQ) score calculation methods show sensitivity to different aspects of data integrity. For each trial, a single protein interaction dataset was replaced by a corrupted version by removing (false negatives, ***left***), adding (false positives, ***center***), or rewiring (noisy data, ***right***) a given fraction of interaction links at random. Shown are the RDQ scores assigned to the corrupted dataset (*y*-axis) as a function of the extent of its corruption (*x*-axis) for three different methods of RDQ calculation: **(a)** the "percent-covering" method, which accurately penalizes false negatives (***left***) and scrambled network links (***right***) but tolerates false positives (***center***); **(b)** the "percent-covered" method, which penalizes false positives (***center***) and rewired network links (***right***), but tolerates false negatives (***left***); and **(c)** the "*p*-value" method, which is unaffected by dataset corruption introduced by false negatives (***left***), false positives (***center***)**,** or rewired links (***right***). The "percent-covered" method was ultimately chosen because of its ability to discriminate against the types of data noise expected to occur in the datasets used in this study (***boxed***).



Supplementary Figure 2. Comparison of relative dataset quality (RDQ) score calculations with independent dataset integrity indicators. Shown are: **(a)** enrichment of the dataset in links between proteins annotated to the same Gene Ontology (GO) process term, and **(b)** the saturation coefficient of the network defined by the dataset. Only the "percent-covered" method (green) generates RDQ scores which give higher ranks (*a*) to datasets with greater enrichment in links between functionally related proteins. The correlation is shown to be unbiased by high network saturation (*b*), a metric to detect high rates of false positives which may be expected to occur among functionally related proteins.

a



b

$T(N) = a + \sum b_i + \sum c_i$ Total weight of links from a single node N

$L(N) = a + \sum b_i$ Weight of links from node N to common neighbors of N and N'

$W_1(N,N') = a^2 + \sum (b_i * b_i')$ Dot product of weights to common neighbors, deemphasized mutual link (a<1)

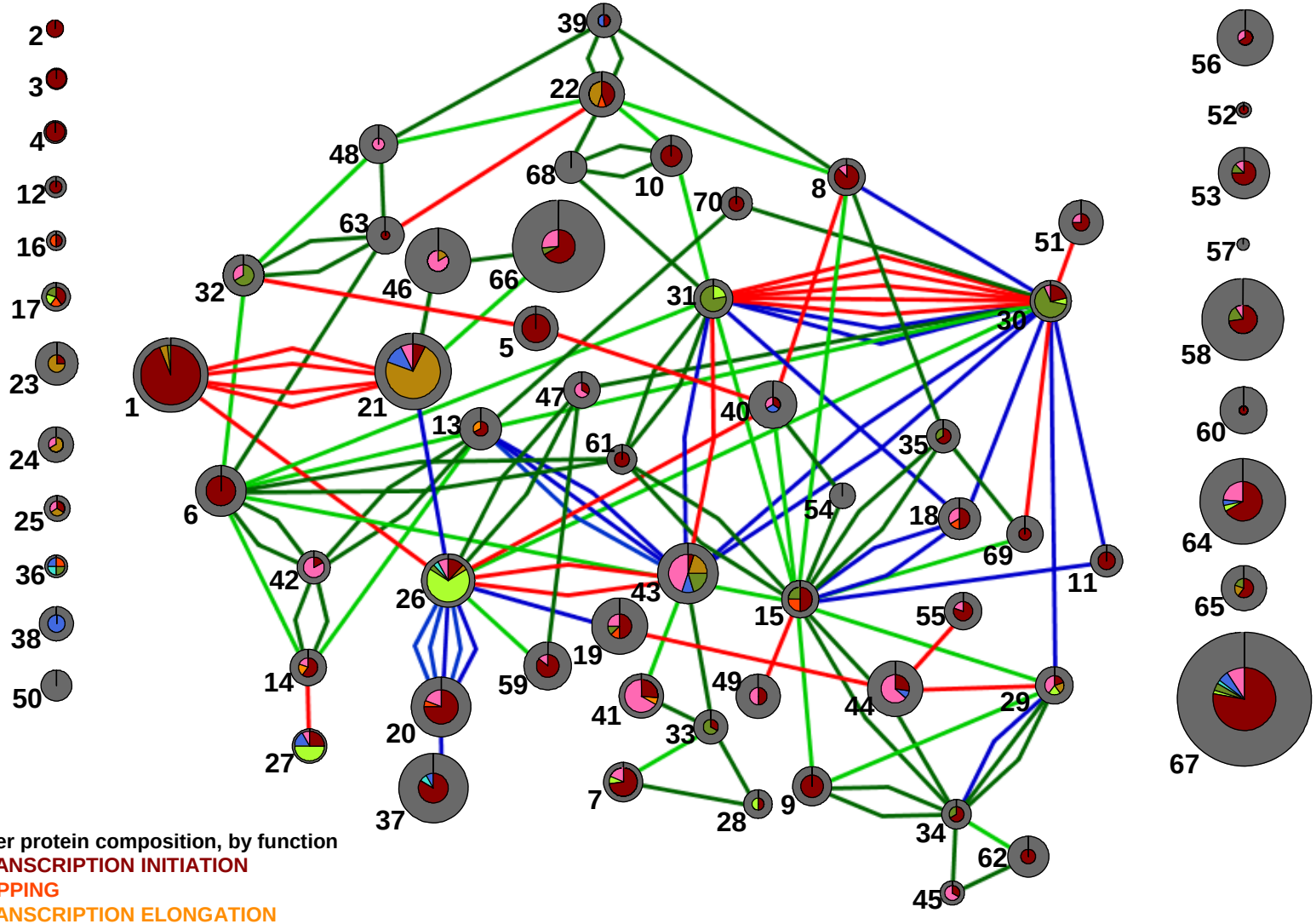
$W_2(N,N') = a + \sum (b_i * b_i')$ Dot product of weights to common neighbors, emphasized mutual link

c

CC formula interpretation

	Emphasizes common neighbors	Emphasizes mutual link	Rewards selectivity of single node
$CC_1(N,N') = \frac{L(N)+L(N')}{T(N)+T(N')}$			
$CC_2(N,N') = \frac{W_1(N,N')}{T(N)+T(N')}$	+		
$CC_3(N,N') = \frac{L(N)+L(N')}{T(N)*T(N')}$			+
$CC_4(N,N') = \frac{W_1(N,N')}{T(N)*T(N')}$	+		+
$CC_5(N,N') = \frac{W_2(N,N')}{T(N)+T(N')}$	+	+	
$CC_6(N,N') = \frac{W_2(N,N')}{T(N)*T(N')}$	+	+	+

Supplementary Figure 3. The pairwise clustering coefficient (CC) is a measure of local cohesiveness in the network neighborhood of a pair of nodes. **(a)** Graphical depiction of a network neighborhood. For a pair of nodes N and N' marked in red, a is the weight of the link between them, b_i and b_i' are the weights of the i-th links of N and N', respectively, to common neighbors, and c_i and c_i' are the weights to uncommon neighbors. **(b)** Metrics of total connectivity of a single node (T) and weighted connectivity to shared neighbor nodes (L, $W_{1,2}$), weighted by link strength. **(c)** The six formulas for the nodewise CC between N and N' in terms of the variables defined in (b). Each metric places a different emphasis on the mutual link within the pair and on links to common and distinct neighbors. In each case, strong links to common neighbors increase the CC, while links to uncommon neighbors decrease it. Note that a direct link between N and N' is not required to obtain a nonzero CC, making it especially appropriate in analyzing networks derived from datasets expected to contain false negatives.



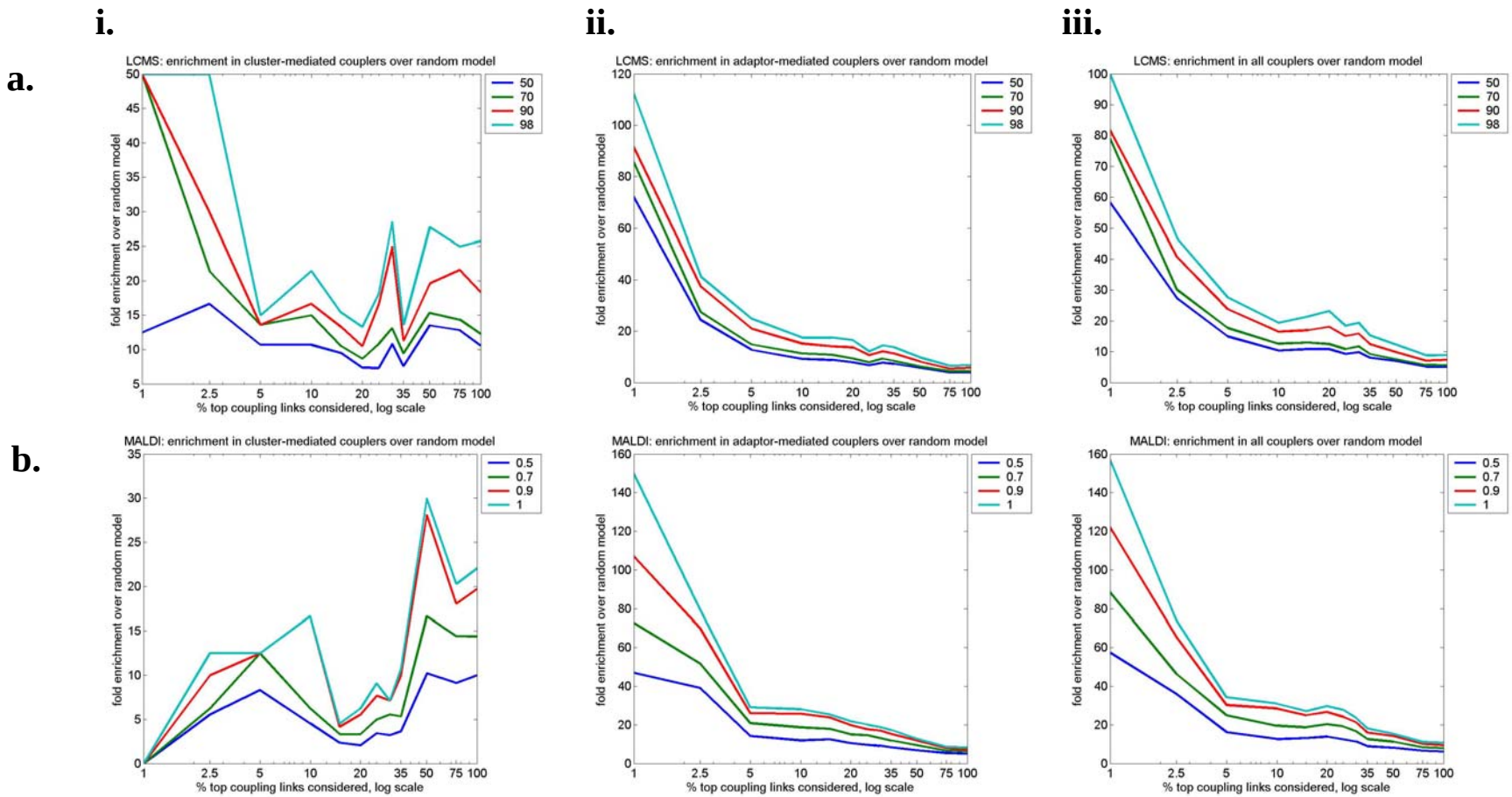
Cluster protein composition, by function

1. TRANSCRIPTION INITIATION
2. CAPPING
3. TRANSCRIPTION ELONGATION
4. SPLICING
5. TRANSCRIPTION TERMINATION AND POLY-A
6. EXPORT
7. NMD
8. PRE-MRNA DEGRADATION
9. TRANSLATION
10. OTHER MRNA-RELATED
11. OTHER AND UNKNOWN

Coupling motif patterns

- Direct
- Cluster-mediated
  - between coupled clusters
  - to mediating cluster
- Adaptor-mediated

Supplementary Figure 4. Protein clusters used in the analysis, along with the top 25-ranked coupling motifs among them. The area of each colored region in a cluster is proportional to the number of member proteins annotated to the corresponding functional category. Especially striking are the many motifs between cluster pairs (C1, C21), (C30, C31), and (C20, C26). The adaptor-mediated coupling motifs between C1 and C21, discussed in the text, suggest coupling of splicing machinery to chromatin modification. The direct coupling motifs between C30 and C31 correspond to potential mechanisms to link mRNA export to mRNA transcription and termination processes. Note that the many direct coupling motifs between C20 and C37 are suspected to be artifacts caused by the mutiple roles of actin protein, Act1p, in the cell (in C20, the actin monomerAct1p is a part of Swr transcription initiation complex, while C26 includes binding partners of polymerized actin).



Supplementary Figure 5. Fold enrichment of interactions in independent protein interaction datasets identified as **(i)** cluster-mediated, **(ii)** adaptor-mediated, or **(iii)** any coupling links in our model, as compared to randomized models. The independent, comprehensive protein interaction datasets were derived from systematic, previously unpublished complex precipitation studies using **(a)** LCMS and **(b)** MALDI-TOF mass spectrometry analysis. Shown are the fold enrichments of the number of interactions identified as couplers in the model used in this study, over the average number of interactions identified as direct couplers in 50 randomized models. The fold enrichment (*y*-axis) is shown as a function of the percentage of top-ranking coupling links considered (*x*-axis). Higher-ranking adaptor-mediated or total coupling links are more likely to appear in the independent datasets. Independent protein interaction datasets are subjected to thresholds at four different interaction confidence values (line colors). Higher-quality interaction data demonstrates greater enrichment in the model versus in random models.

Supplemental Table 1.

Category	Description	number of proteins included in network
1	conjugation	13
2	cytokinesis	31
3	carbohydrate metabolism	40
4	energy pathways	8
5	electron transport	3
6	DNA metabolism	170
7	transcription	257
8	protein biosynthesis	116
9	protein modification	123
10	amino acid and derivative metabolism	30
11	lipid metabolism	33
12	coenzyme and prosthetic group metabolism	13
13	vitamin metabolism	9
14	transport	153
15	response to stress	94
16	organelle organization and biogenesis	39
17	nuclear organization and biogenesis	110
18	cytoskeleton organization and biogenesis	50
19	cell wall organization and biogenesis	46
20	cell cycle	94
21	budding	17
22	pseudohyphal growth	23
23	meiosis	52
24	signal transduction	26
25	morphogenesis	9
26	membrane organization and biogenesis	11
27	RNA metabolism	174
28	vesicle-mediated transport	68
29	cell homeostasis	13
30	protein catabolism	39
31	sporulation	18
32	ribosome biogenesis and assembly	30
33	cellular respiration	17

Supplemental Table 2.

Data-set ID	Dataset name	RDQ score using "percent-covering" overlap matrix	RDQ score using "percent-covered" overlap matrix	RDQ score using "*p*-value" overlap matrix	# interactions between proteins annotated to the same GO process term	% interactions between proteins annotated to the same GO process term	global CC	# distinct links in network	# links reiterated in network	# distinct proteins in network involved	# products of essential genes involved	% products of essential genes involved
1	Ito-core Y2H	3.098	5.398	7.762	67	3.995	1.778	1675	2	225	58	25.778
2	Ito-full Y2H	2.788	1.160	9.270	68	8.134	2.323	400	436	721	158	21.914
3	Uetz Y2H	4.536	6.842	6.303	90	18.000	1.638	462	38	282	75	26.596
4	Complex	16.855	2.472	8.024	923	15.435	6.229	4753	1227	763	293	38.401
5	Rosetta	0.340	0.262	10.270	43	2.376	5.621	905	905	161	40	24.845
6	Paralog	26.684	0.581	10.270	2413	7.359	21.231	32017	771	1508	324	21.485
7	Phylogenetic	0.150	1.477	11.333	12	8.451	3.381	71	71	21	9	42.857
8	MIPS-affinity	3.249	28.661	5.280	52	41.935	1.274	79	45	62	31	50.000
9	MIPS-co-precipitation	9.567	17.318	4.602	256	42.384	2.059	385	219	187	83	44.385
10	MIPS-co-purification	5.647	18.920	4.194	160	42.440	2.122	208	169	98	38	38.776
11	MIPS-synthetic	5.784	5.919	9.336	225	27.847	1.997	681	127	341	103	30.205
12	MIPS-Y2H	12.125	8.793	5.035	248	18.249	2.261	961	398	425	131	30.824
13	MIPS-other	9.176	2.196	8.321	254	7.157	5.474	2912	637	532	70	13.158

Supplemental Table 3.

	z-value
RDQ only	27.80
CC1	81.92
CC2	24.39
CC3	1.98
CC4	1.73
CC5	38.81
CC6	15.22

Supplemental Table 4.

	RDQ	*k*=70	*k*=105	*k*=210
derived CC scores	"percent-covered"	0.4097	0.2126	0.1664
	C1 assigned	0.2237	0.2053	0.1637
	C2 assigned	0.235	0.2059	0.1582
RDQ-weighted network scores	"percent-covered"	0.2819	0.2303	0.1831
	C1 assigned	0.2649	0.2365	0.186
	C2 assigned	0.2811	0.2423	0.1898

Supplementary Table 6. Characterizations of clusters.

Cluster C1 (76 proteins): [Transcr. Init.] SAGA, Swi/Snf, ISWI, RSC: Machinery for chromatin remodeling.

Cluster C2 (4 proteins): [Transcr. Init.] Transcription factors.

Cluster C3 (6 proteins): [Transcr. Init.] Chromatin remodeling in response to mating-type signals.

Cluster C4 (7 proteins): [Transcr. Init.] Signal integration, including mating.

Cluster C5 (27 proteins): [Transcr. Init.] Transcription initiation factors, especially cluster-domain-containing.

Cluster C6 (35 proteins): [Transcr. Init.] Chromatin remodeling, DNA replication.

Cluster C7 (21 proteins): [Transcr. Init., Transcr. Term/PolyA.] DNA-binding proteins, mRNA export proteins, membrane permeases.

Cluster C8 (19 proteins): [Transcr. Init.] TFIIA, TFIID.

Cluster C9 (21 proteins): [Transcr. Init.] Chromatin remodeling, GATA transcription factors.

Cluster C10 (23 proteins): [Transcr. Init.] Chromatin modification.

Cluster C11 (14 proteins): [Transcr. Init.] Telomere silencing and transcription factors.

Cluster C12 (6 proteins): [Transcr. Init.] Transcription factors.

Cluster C13 (24 proteins): [Capping] RNA PolII/III core, capping-elongation switch.

Cluster C14 (18 proteins): [Capping] pre-mRNA capping, signal sensors, ER/Golgi enzymes.

Cluster C15 (19 proteins): [Transcr. Elong., mRNA Export] THO complex.

Cluster C16 (5 proteins): [Transcr. Elong.] Elongation through chromatin.

Cluster C17 (11 proteins): [Transcr. Elong., Transcr. Term/PolyA., mRNA Export] Elongation, THO complex, export.

Cluster C18 (24 proteins): [Transcr. Elong.] DNA helicases, DNA repair, and meiosis/silencing at HML/R.

Cluster C19 (43 proteins): [Transcr. Elong.] THO complex, links to actin, cytoplasmic protein sorting and vesicle transport.

Cluster C20 (49 proteins): [Transcr. Elong., Transcr. Init.] Elongation through chromatin, actin-binding/modifying proteins, Swr complex.

Cluster C21 (80 proteins): [Splicing, mRNA Degrad.] Spliceosome, Lsm proteins.

Cluster C22 (28 proteins): [Splicing, Transcr. Elong.] Elongation and splicing machinery assembly.

Cluster C23 (25 proteins): [Splicing] Splicing, ER and plasma membrane.

Cluster C24 (17 proteins): [Splicing] Mitochondrial splicing, other non-nuclear proteins.

Cluster C25 (9 proteins): [Splicing] Diverse nuclear proteins.

Cluster C26 (39 proteins): [Transcr. Term/PolyA., NMD] Main termination and polyadenylation machinery.

Cluster C27 (16 proteins): [Transcr. Term/PolyA., mRNA Degrad.] Multifunctional CCR-NOT complex: represses transcription initiation, aids transcription elongation, and acts as a 3'->5' exoribonuclease for deadenylation-dependent mRNA decay.

Cluster C28 (11 proteins): [Transcr. Term/PolyA.] Cytoplasmic/membane proteins.

Cluster C29 (19 proteins): [Transcr. Term/PolyA., Translat.] Diverse nuclear proteins.

Cluster C30 (23 proteins): [mRNA Export, Transcr. Term/PolyA.] mRNA export, transcription termination, polyA-binding.

Cluster C31 (21 proteins): [mRNA Export, Transcr. Term/PolyA.] Export of polyadenylated mRNA.

Cluster C32 (23 proteins): [mRNA Export] mRNA export, heme-activated transcription initiation.

Cluster C33 (16 proteins): [mRNA Export] Nuclear export, mitosis/meiosis proteins, nuclear membrane disintegration.

Cluster C34 (12 proteins): [mRNA Export] Chromosome maintenance proteins, interaction partners of GTP-exchange factors.

Cluster C35 (15 proteins): [mRNA Export] Diverse proteins.

Cluster C36 (7 proteins): [NMD, Transcr. Elong., mRNA Degrad., mRNA Export] THO complex, NMD machinery components.

Cluster C37 (66 proteins): [NMD] Interaction partners of GTP-binding proteins: involved in chromatin modification, NMD, polyA-binding, export, as well as exocytosis, stress response, and amino acid metabolism.

Cluster C38 (16 proteins): [mRNA Degrad.] Exosome complex.

Cluster C39 (16 proteins): [mRNA Degrad.] mRNA and protein degradation, nuclear phosphatase and kinase regulators.

Cluster C40 (31 proteins): [mRNA Degrad.] Budding/cell wall formation, possible polyA-actin link, mRNA degradation.

Cluster C41 (27 proteins): [Translat., Capping] Capping, translation initiation.

Cluster C42 (15 proteins): [Translat.] Cotranslational targeting of nascent polypeptides.

Cluster C43 (50 proteins): [Translat., mRNA Export, Splicing, mRNA Degrad.]

Cluster C44 (42 proteins): [Translat.] Translation initiation complexes eIF2A/B, eIF3, eIF5; drug response/transport, enzymes.

Cluster C45 (8 proteins): [Translat.] Translational elongation and ribosomal subunit biogenesis.

Cluster C46 (58 proteins): [Translat.] Chromatin modification, translation, ribosome, amino acid synthesis.

Cluster C47 (18 proteins): [Translat.] Ribosome components, diverse non-nuclear proteins.

Cluster C48 (20 proteins): [Translat.] Various enzymes from diverse cellular components, several interaction partners of Sin4p.

Cluster C49 (27 proteins): Chromosome structure and repair, DNA exonucleases.

Cluster C50 (13 proteins): Mitochondrial and unknown proteins.

Cluster C51 (27 proteins): Protein targeting: vacuolar and secretion proteins; gene silencing.

Cluster C52 (3 proteins): Plasma membrane synthesis.

Cluster C53 (36 proteins): Previously unannotated transcription elongation proteins, cell cycle and actin structure-related proteins.

Cluster C54 (9 proteins): Interaction partners of Srb2p (by yeast-two-hybrid).

Cluster C55 (19 proteins): APC complex.

Cluster C56 (43 proteins): Mitochondrial and ribosomal proteins, PolII core.

Cluster C57 (2 proteins): Upstream activating factor for PolI and PIP3-phosphatase.

Cluster C58 (91 proteins): Diverse proteins, most interact with nuclear pore proteins.

Cluster C59 (31 proteins): Mitotic/meiotic checkpoint module, DNA double-strand break repair; proteasome.

Cluster C60 (30 proteins): Stress response, protein degradation.

Cluster C61 (12 proteins): RNA Pol I/III, RSC complex.

Cluster C62 (23 proteins): Diverse proteins.

Cluster C63 (19 proteins): Transcription Mediator, proteins assembling on DNA and chromatin.

Cluster C64 (100 proteins): Signaling transcription factors, other interactors of Sua7p by yeast-two-hybrid.

Cluster C65 (28 proteins): Cell signal responses.

Cluster C66 (116 proteins): Kinases, DNA replication.

Cluster C67 (245 proteins): DNA-binding proteins: PolII, transcriptional activation and repression, chromatin modifications, and cell cycle mechanisms; mRNA catabolism; enzymes and GTP-binding proteins.

Cluster C68 (14 proteins): Interaction partners of karyopherin Crm1.

Cluster C69 (18 proteins): Transcription Mediator, PolII holoenzyme.

Cluster C70 (14 proteins): mRNA export; histone deacetylation; ribosomal proteins; interaction partners of telomere-maintenance complex.

Supplemental Table 7.

Motif rank	Motif score	Protein 1	Cluster 1	Protein 2	Cluster 2
1	0.027927	SIR4	30	SIR2	18
2	0.027008	CDC33	21	TIF4632	26
3	0.020943	NPL3	30	MTR10	31
4	0.017478	SIR4	30	RAP1	11
5	0.017346	ACT1	26	PFY1	19
6	0.017004	PBP1	31	LSM12	43
7	0.015999	ACT1	26	SRV2	20
8	0.015932	YAP1	30	CRM1	43
9	0.015026	YRB2	30	CRM1	43
10	0.014328	ARC1	30	MES1	8
11	0.013414	ACT1	26	COF1	20
12	0.01306	ACT1	26	TWF1	20
13	0.011572	HRP1	30	NAB2	29
14	0.011402	TRI1	15	TOP1	18
15	0.011119	RIF2	15	RAP1	11
16	0.0109	NUP159	30	NMD5	31
17	0.01056	ACT1	26	TPM2	20
18	0.010273	DDI1	29	YJR141W	34
19	0.0094945	TOF2	15	TOP1	18
20	0.0094945	GLE1	31	YEL024W	18
21	0.0088657	RSP5	13	BUL1	43
22	0.0086723	BIT61	37	YGR071C	20
23	0.0086723	RSP5	13	YOR385W	43
24	0.0086723	RSP5	13	DIA1	43
25	0.0085688	ACT1	26	MYO4	20

Supplemental Table 8.

Coupled Cluster 1	Coupling protein(s) in Cluster 1	Coupling protein(s) in Mediating Cluster linked to Cluster 1	Mediating cluster	Coupling protein(s) in Mediating Cluster linked to Cluster 2	Coupling protein(s) in Cluster 2	Coupled Cluster 2
15	RIF2	YJL015C	54	YDR154C	RVS161	40
13	RPB9	SPR6	70	GluRS	ARC1	30
7	DSE3	PXL1	28	YKR021W	YOL014W	33
34	MUM2	KAR4	45	KAR4	YNL196C	62
21	LSM1	RIA1	46	HCH1, YMR293C, RSM23, DOA4, AHC2	JNM1, JSN1, BZZ1, BZZ1, ATG17	66
13	UFD1	YPL222W	42	SSA4	ALG2	14
9	EAF6	VPS68	34	AIR1	TRF4	15
9	EAF6	VPS68	34	YJR141W	DDI1	29
6	APQ12	SSA2	42	YPL222W	UFD1	13
15	RIF2	GIS3	35	LTV1	SRB4	69
6	APQ12	SSA2	42	SSA4	ALG2	14
41	SUA7	LRE1	33	PRO1	SUP35	43
15	RIF2	RPC40	61	RPC25	LOS1	31
10	NST1	IBD2	68	MPC54	NMD5	31
10	NST1	IBD2	68	YFL068W	ATG16	22
26	ACT1, RNA15	RPP2B, GRX3	47	WTM2	PSE1	30
26	ACT1, RNA15	RPP2B, GRX3	47	RFA3	RAD52	59
15	TRF4	AIR1	34	YJR141W	DDI1	29
8	ARO9	YIM1	35	GIS3	RIF2	15
8	BRF1	TFC4	39	TRK2	TFB1	22
6	YIR014W	HOM3	63	YKL137W	RRN11	32
6	SDS3	RSC6	61	RPC40	RIF2	15
6	SDS3	RSC6	61	RPC25	LOS1	31
32	RRN11	YKL137W	63	HMO1	FHL1	48
22	TFB1	TRK2	39	SKI8	COX4	48

Supplemental Table 9.

Motif rank	Motif score	Adaptor protein	Cluster containing adaptor protein	Coupled cluster
1	0.10942	HAP2	5	32
2	0.093934	PRP4	1	21
3	0.073389	NSP1	30	31
4	0.063381	STD1	40	5
5	0.057796	PRP46	1	21
6	0.044698	ACT1	26	40
7	0.039833	PAB1	43	26
8	0.036457	KAP95	51	30
9	0.032408	NUP57	31	30
10	0.025553	HSH49	1	21
11	0.020811	SRB2	1	21
12	0.020244	NAB2	29	44
13	0.018787	TFC7	14	27
14	0.017403	NUP49	31	30
15	0.017174	SPC29	55	44
16	0.014692	KAP104	31	30
17	0.014576	RPT5	8	40
18	0.014566	GSP1	30	31
19	0.014525	TIF4632	26	43
20	0.014516	FAR3	19	44
21	0.012929	CRM1	43	31
22	0.012771	YAP1	30	69
23	0.012548	HHF2	26	1
24	0.01246	IES5	22	63
25	0.011735	TRF4	15	49