TITLE

A computational framework for analysis of prey-prey associations in interaction proteomics identifies novel human protein-protein interactions and networks

AUTHORS AND AFFILIATIONS

Sudipto Saha[1], Jean-Eudes Dazard[1], Hua Xu[1]and Rob M. Ewing[1,2]*

[1]Center for Proteomics and Bioinformatics, [2]Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106, USA

*To whom correspondence should be addressed: Rob M. Ewing, Ph.D., Center for Proteomics and Bioinformatics, Case Western Reserve University,10900 Euclid Avenue, Cleveland, OH 44106; Tel: (216) 368-4380; Fax: (216) 368-6846, Email: rob.ewing@case.edu

Abstract

Large-scale protein-protein interaction datasets have been generated for several species including yeast and human and have enabled the identification, quantification and prediction of cellular molecular networks. Affinity purification-mass spectrometry (AP-MS) is the preeminent methodology for large-scale analysis of protein complexes, performed by immunopurifying a specific 'bait' protein and its associated 'prey' proteins. The analysis and interpretation of AP-MS datasets is however, not straightforward. In addition, although yeast AP-MS datasets are relatively comprehensive, current human AP-MS datasets only sparsely cover the human interactome. Here we develop a framework for analysis of AP-MS datasets that addresses the issues of noise, missing data and sparsity of coverage in the context of a current, real world human AP-MS dataset. Our goal is to extend and increase the density of the known human interactome by integrating bait-prey and co-complexed preys (prey-prey

associations) into networks. Our framework incorporates a score for each identified protein, as well as elements of signal processing to improve the confidence of identified protein-protein interactions. We identify many protein networks enriched in known biological processes and functions. In addition, we show that integrated bait-prey and prey-prey interactions can be used to refine network topology and extend known protein networks.

Keywords

protein-protein interaction network; affinity purification mass-spectrometry; interactome

Introduction

Interactions between proteins and the protein complexes and networks that these interactions form, are fundamental units of biological organization that mediate most cellular processes. Understanding the topology of protein interaction networks is therefore a primary goal of systems biology. The principal techniques for large-scale analyses of protein interactions are yeast two-hybrid and affinity-purification mass-spectrometry (AP-MS). These two approaches provide complementary views of the protein interactome[1]; the yeast-two-hybrid assay identifies binary interactions between pairs of proteins, whilst affinity-purification mass-spectrometry (AP-MS) identifies protein complexes associated with a given "bait" protein. AP-MS experiments have been conducted by purifying protein complexes using native antibodies or on a larger-scale by epitope tagging bait proteins and recovering associated protein complexes with antibodies directed against the epitope tag. Despite the power of AP-MS to map protein complexes, however, several technical hurdles exist, including the presence of non-specific interacting proteins in these large-scale datasets as well as lack of reproducibility and sparse coverage of the underlying networks.

The yeast protein interactome is by far the most well covered eukaryotic interactome with multiple studies and techniques contributing large-scale datasets[2-5]. With its larger size and greater complexity,

46 the human protein interactome has been mapped more selectively, with studies focusing on specific
47 complexes or classes of protein[6-9]. Diverse computational methods have been developed for analysis of
48 large-scale interaction proteomics datasets. For AP-MS datasets, the focus of these analysis methods
49 has been the assignment of confidence scores to observed interactions and to distinguish specific from
50 non-specific interactions [6-7,10-14]. In yeast, the socio-affinity index was developed to score a large-scale
51 yeast interactome dataset with high coverage, and reciprocal AP-MS experiments[10].
52 In human datasets, with less dense coverage of the underlying networks, computational methods and
53 scores have focused on distinguishing specific from non-specific interactions. For example, the
54 normalized spectral abundance factor (NSAF) was used as a measure of abundance for each protein
55 and the ratio of the vectors of counts for each protein between "control" and "bait" experiments used to
56 distinguish specific from non-specific interactors[6]. The D-score, a metric combining total spectral
57 count, reproducibility and overall frequency of prey proteins in AP-MS experiments into a single
58 confidence score for bait-prey associations was developed for the analysis of human AP-MS data[7]. The
59 D-score was applied to a large-scale study of human deubiquitinating enzymes and shown to out-
60 perform the socio-affinity index, the NSAF method and Z-score on the data in hand[7]. An alternative
61 approach, using mixture modeling and Bayesian statistics named SAINT (Significance Analysis of
62 Interactome) was developed[13] and applied to the analysis of phosphatase interaction networks[15]. The
63 Decontaminator uses a Bayesian approach to model false-positive protein-protein interactions (PPIs) by
64 comparing the score of a putative prey protein in induced vs control experiments[14].
65 Two broad models for interpreting AP-MS data in a network context have been proposed. The "spoke"
66 model holds that the bait protein interacts with each of the identified "prey" proteins (the bait
67 representing the center of a wheel with spokes connecting to each of the prey proteins) whereas the
68 "matrix" model assumes that each of the identified proteins (bait and prey) interacts with each of the
69 others in a given AP-MS experiment[16]. Although the matrix model will capture a higher proportion of

70 the underlying protein associations, this comes with the price of increased false positives. Therefore,
71 although both models have merits, it is likely that a combination of spoke and matrix models represent
72 underlying biological reality in most cases. Several scoring metrics for AP-MS data explicitly
73 implement these concepts. The socio-affinity index is a summary score including spoke and matrix
74 terms, as well as accounting for reciprocal bait-prey AP-MS experiments and computes the log ratio of
75 protein co-occurrence given their observed frequencies[10]. The hypergeometric distribution has been
76 used to calculate expected frequencies of co-occurrence for proteins using the matrix model in the
77 large-scale yeast AP-MS datasets and found to be an effective means of identifying protein-protein
78 interactions[17]. Other schemes that make use of co-occurrence frequencies have been proposed for
79 resolving protein complexes using large-scale yeast AP-MS datasets[18-19] but with added refinements
80 such as consideration of the variation in bait affinity when computing the results[18].
81 Probabilistic approaches using small to medium scale human AP-MS datasets have also been
82 developed[6,20]. These studies analyze datasets in which the coverage of given protein complexes by bait
83 proteins is relatively high. Notably, these studies[6,20] make use of the quantitative features of mass-
84 spectrometry data, rather than using binary co-occurrence frequencies. Measures of abundance such as
85 spectral counts and other mass-spectrometry based confidence measures are very informative features
86 of AP-MS data, since protein-protein affinities are likely to vary considerably. An emerging consensus
87 of these diverse computational methods is that both bait-prey and prey-prey associations in AP-MS
88 datasets can be used to identify protein-protein interactions. Utilization of prey-prey associations is
89 most obvious in AP-MS datasets with dense bait coverage, so that co-occurrence profiles are well
90 defined. However, it is less clear whether this concept may be applied to less dense AP-MS datasets
91 such as larger-scale human AP-MS datasets[7,12]. Our goal in this work is to test the utility of mining
92 prey-prey associations from large-scale, but less dense human AP-MS data.
93 We previously generated a large-scale human interaction proteomics dataset focused on 338 human bait

proteins, many of which are linked to human diseases[12]. This dataset is the largest human AP-MS dataset to date and mass-spectrometry experiments were performed in a uniform manner, making it a useful resource for development of data analysis techniques. In contrast to other human AP-MS datasets that focus on individual complexes or processes, our dataset represents a broad survey of human proteins and their interactions. In addition, the dataset represents several complexes with relatively high coverage (in terms of number of baits), such as the proteasome, and Eukaryotic Initiation Factors, and many other human protein complexes that are sparsely covered. Thus our motivation is to develop a method that can identify probable protein complexes from a heterogeneous (in terms of coverage or sampling of protein complexes) dataset. In our initial analysis of the dataset we focused on bait-prey interactions (spoke model) by using partial least squares to predict the reproducibility of each prey protein observation based upon a training set of highly-reproduced AP-MS experiments[12]. In the current work, we mine the dataset more comprehensively by identifying prey-prey associations based upon co-occurrence profiles. We show that these prey-prey interactions are a valuable source of protein interactions and that integration of bait-prey and prey-prey associations yields improved network models of protein complexes. Importantly, we show that prey-prey associations may be identified from less well covered interactomes, such as the current human interactome. Since comprehensive efforts to map the human protein interactome are still in their early states, methods to mine protein-protein interactions from incomplete AP-MS datasets will be important tools for the foreseeable future.

Our framework exploits the quantitative features of AP-MS data to assign confidence scores to each bait-prey observation. To improve signal-to-noise ratio, the sparse matrix of scores are then transformed to eigenvalues using singular value decomposition. A similarity measure between the profiles of each pair of prey proteins is then to detect potential prey-prey associations, which are the starting point for the clustering and reconstruction of protein complexes, which we validate by

reference to known complexes and annotations. Finally, we integrate bait-prey and prey-prey associations and show that these integrated networks have improved biological coherence. Importantly, we are able to directly compare bait-prey and prey-prey interactions in terms of their biological coherence and show that prey-prey interactions are a rich source of protein-protein associations. We illustrate our approach using selected specific protein networks and show how novel interactions can be identified.

The principal contribution of our work is to show how large-scale AP-MS datasets may be data-mined for identification of protein-protein associations. Our analysis pipeline is generically applicable to large-scale AP-MS datasets and we anticipate that as increasing volumes of AP-MS data are available, it can be applied and used to discover novel protein interactions and elucidate the topology of protein interaction networks.

Methods

Data set

Data used to develop our approach is principally derived from a previously described human AP-MS dataset, in which 832 single-step anti-FLAG immunoprecipitation experiments representing 384 human bait proteins (~50% of baits were replicated) identified 5269 distinct prey proteins[12]. The human bait proteins were selected based on association with diseases such as cancer and obesity. The data were generated as previously described[12], except that all spectra were re-searched against an IPI human protein sequence database (version 3.31) (92012 sequences) using the Mascot mass-spectrometry search engine (version 2.1, Matrix Science; fixed modification: Carbamidomethyl (C), variable modification: Oxidation (M); peptide mass tolerance 2Da; fragment mass tolerance 0.4Da; missed cleavages:2). Peptide and protein identifications were processed through Peptide and Protein Prophet[43] and imported to LabKey server[44] (version 10.3) for data management. For subsequent steps of the

analysis, corresponding gene symbols, where available, were used to represent baits and preys.

Protein Identification False Discovery Rates

To assess the protein identification false discovery rate, we used decoy database searches. Using the same Mascot search parameters as above, we searched all data against a concatenated decoy human IPI protein sequence database (version 3.31) and computed the mean false discovery rate across all searches (3.63%). We also searched the complete datasets using another search-engine, MassMatrix[42] (version 2.4.2, http://www.massmatrix.net) that provides the false discovery rate in terms of the % decoy hits when searched against concatenated decoy databases. The database and search parameters were the same as for the Mascot analysis, except for the following additional options: peptide length: 6-40 amino acid residues and score thresholds of 5.3 and 1.3 for the pp and pp_{tag} scores respectively. Proteins identified using MassMatrix were cross-referenced to those identified using Mascot. Of the set of 34383 bait-prey associations from the original Mascot searches, and for which a D-score was computed (see below), 82% were also identified with MassMatrix. The mean false discovery rate across this set is 4.52%.

Bait-prey confidence score

The D-score (Equation 1) is a previously described confidence score for AP-MS data that combines measures of abundance (spectral counts), specificity and reproducibility into a single score for each bait-prey observation[7].

$$D_{ij} = \sqrt{\left(\frac{k}{\sum_{i=1}^{i=k} f_{ij}} \right)^{p}} X_{ij} \text{ for all } i, j \qquad \text{(Eq. 1)}$$

Where, k= total number of unique bait proteins; X_{ij}= total spectral counts for prey i from bait j; f_{ij}={1,0}; p= number of replicates runs in which the prey protein is present. This generated a complete bait-prey D-score matrix (**D**) of dimensions (5269 proteins x 384 baits).

Contaminant identification

False-positive proteins may occur in AP-MS experiments for different reasons. First, there are many proteins that occur frequently in AP-MS experiments as a result of non-specific affinity. In the AP-MS dataset used in this study, a set of 200 control AP-MS experiments (using HEK293 cells with vector alone, i.e. no bait) provide a dataset for identifying 'control' proteins. Since the D-score accounts for protein frequency of occurrence, these proteins typically have low D-scores. D-scores for the top 100 highly frequent control proteins in this study have median value equal to 3 (Supplementary Figure S1). These were removed along with any protein with D-score <3 from the initial dataset (13688 bait-prey interactions with D-score <=3 were removed). A second type of contaminant results from cross-contamination at the experimental level between samples. Since the experimental protocol used to generate data used here[12] used gel-based separation of proteins prior to mass-spectrometry, we identified proteins with D-score > 10 co-occurring on the same gels or samples that were processed on the same date with different bait proteins as potential cross-contaminants and removed them from the dataset (1550 bait-prey interactions were removed using this criterion).

Latent Semantic Analysis (LSA)

The D-score matrix (**D**) is necessarily sparse (1.25% non-zero values), since the frequency of most proteins is low (identified with few or unique baits). To identify relationships between proteins in this matrix, we made use of the principles of latent semantic analysis (LSA) which uses Singular Value Decomposition (SVD) to reduce matrix sparseness[21]. SVD was applied to the rectangular bait-prey

183 matrix (Equation 2), and a dimensional reduction of the singular values was performed after identifying
184 the optimum *k*-value which determines the degree of reduction.
185
186 $$\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{T} \approx \mathbf{D}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{T} \quad \text{(Eq. 2)}$$
187
188 **D** denotes the D-score bait-prey matrix (*m* x *n*). **U** (*m* x *m* matrix) and **V** (*n* x *n* matrix) are orthogonal
189 matrices with unit-length columns (i.e., $\mathbf{U}\mathbf{U}^{T} = \mathbf{I}$ and $\mathbf{V}^{T}\mathbf{V} = \mathbf{I}$) and **Σ** is a diagonal matrix containing
190 the ordered singular values of rank $r = \min(m,n)$. $\mathbf{D}_k$ denotes the approximated latent semantic
191 representation of the bait-prey matrix **D** where *k* denotes the selected number of singular values used in
192 the approximation. This generated a final bait-prey D-score matrix ($\mathbf{D}_k$) of dimensions (2242 proteins x
193 384 baits). LSA computations were performed using the Perl Data Language (PDL, http://pdl.perl.org).
194

195 *Prey-prey similarity score*

196 For each bait protein, the bait vector is the vector of D-scores for all preys. Similarly, for each prey
197 protein, the prey vector is the vector of all D-scores for all baits. LSA projects the vectors of original
198 bait and prey vectors into lower dimensional semantic space. Transformed bait and prey vectors are the
199 vectors of bait or prey eigenvalues. Cosine similarity (Equation 3) was used to compute pairwise
200 similarities between prey-prey vectors from the bait-prey matrix. The similarity score for a pair of prey
201 proteins a, and b is then the cosine similarity of their associated D-score vectors (0 indicates minimal
202 similarity, 1 indicates maximal similarity). Hereafter the cosine similarity is referred to as the prey-prey
203 score (PPS).
204

205 $$sim(a,b) = \cos\alpha(a,b) = \frac{a.b}{\|a\|.\|b\|} = \frac{\left(\sum_{i=1}^{n} a_i b_i\right)}{\left(\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}\right)} \quad \text{(Eq. 3)}$$

206

207 *Analysis of known interactions*

208 Bait-prey and prey-prey interactions were compared to CORUM version 2.0[22] and BioGRID version
209 3.1.78[23] datasets. The complete set of CORUM mammalian protein complexes was used (2577 protein
210 complexes; 4304 distinct gene Ids). Human protein interactions were extracted from BioGRID (402127
211 total interactions; 37910 human interactions). Data originating from our previous publication[12] was
212 carefully removed from BioGRID prior to the current analysis. For each human protein interaction in
213 BioGRID, a 2-hop neighborhood network was computed. Thus, for two pairs of interacting proteins A-
214 B and B-C, the 2-hop network for protein A would include C. Bait-prey and prey-prey interactions in
215 the current study were compared to the 2-hop network for corresponding proteins in BioGRID and the
216 protein complexes in CORUM.

217 *Protein-protein interaction false discovery rates*

218 The mean False Discovery Rate (FDR) estimate was computed by generating a null distribution of
219 prey-prey cosine similarities from a reshuffled bait-prey D-score matrix (random permutation of rows
220 and columns independently). This was repeated 100 times to estimate the standard error of the mean
221 FDR. The number of prey-prey cosine similarity values greater than a given Prey-Prey Similarity (PPS)
222 threshold was used to estimate the number of false prey-prey interactions. The FDR estimate is then the
223 ratio of the number of false prey-prey interactions to the number of prey-prey interactions discovered.
224 This strategy is similar to the ones used to estimate the FDR of peptide identifications from Peptide
225 Spectrum Match with the help of decoy databases (reviewed in[47]), or to the re-sampling based methods

used by Lavallee-Adam et al[48] and Dazard et al.[49] to assess the FDR of Protein-Protein Interactions (PPI) from Affinity Purification Mass Spectrometry data with the help of a control set of experiments. A controlled FDR of 7.58% +/-0.02% was achieved for the chosen PPS threshold of PPS $\geq$ 0.75 (see Results section). Comparison plots of distributions of prey-prey cosine similarities in the original matrix and in the null matrix under the re-sampling scheme are shown in Supplementary Figure S5. We also looked at the FDR profile within a range of PPS, and found that the FDR was quite stable around 6-8% within the PPS range [0.5,0.75] and decreasing (as expected) within the PPS range [0.75,0.99].

Semantic Similarity

The semantic similarity is a metric for computing the similarity of Gene Ontology (GO) terms, their ancestors, or the descendants for two genes[24]. Semantic similarity uses information content (IC) as a measure of the specificity of a term and is quantified as the negative log likelihood,

$$IC = -\log p(c) \qquad \text{(Eq. 4)}$$

where $p(c)$ is the probability of occurrence of c in the GO structure. We used Resnik's max measure of similarity between two terms as the IC of their most informative common ancestor (MICA):

$$sim_{Res}(c_1,c_2) = IC(c_{MICA}) \qquad \text{(Eq. 5)}$$

This measure is effective in determining the information shared by the two terms[25]. Analysis was performed using the GO*Sim* R package[26].

Protein clustering and visualization

Assembly of protein clusters from the prey-prey similarity matrix was performed using Pearson

correlation and hierarchical clustering (using TIBCO Spotfire software). Enrichment analyses of clustered sets of proteins were perfomed using FuncAssociate 2[27]. FuncAssociate performs a Fisher's Exact Test analysis to identify GO terms and the results are corrected for multiple hypotheses via empirical re-sampling, and adjusted p-values computed for significance.Cytoscape (version 2.8) was used for biological network visualization[28].

Results and Discussion

Initial data processing and analysis

An overview of the data analysis workflow is shown in Figure 1. The workflow incorporates steps to address key issues in AP-MS datasets including the occurrence of non-specific prey proteins and missing data due to under-sampling of protein complexes. Scores for both bait-prey and prey-prey protein associations are computed and used to construct protein networks.

Abundance and specificity of prey protein peptides in AP-MS experiments are important features that allow the confidence of prey proteins to be assessed. The D-score[7] was used as the primary metric for each bait-prey observation, since it combines measures of abundance and specificity into a single score and gives weight to well-replicated and less frequent prey proteins.

A principal challenge in interpreting AP-MS data is the presence of non-specific prey proteins. Typically, these are identified by their frequency across the dataset or as proteins present in control experiments. Our previous study[12] analyzed ~200 empty vector control AP-MS experiments that were then used to define non-specific prey proteins. We set a defined frequency threshold and used this threshold to filter out frequently occurring prey proteins[12]. In this analysis, we first observed that D-score values are typically very low for highly frequent proteins (since the D-score negatively weights frequency). Highly frequent proteins in the initial dataset such as PRMT5 (present in 94% of the 995 AP-MS experiments in the dataset) have low median D-scores (the median D-score for PRMT5 is 1.8

and 95% of baits for which PRMT5 is identified have D-score < 10).The top 100 most frequent prey proteins in the dataset have a median D-score of ~3.0 (Supplementary Figure S1). We used this value as a threshold to remove any bait-prey interaction with D-score ≤ 3). Thus, 13688 bait-prey interactions with D-score ≤ 3 were removed from the dataset. The final matrix of D-scores used for subsequent analyses has dimensions of 384 baits x 5269 prey proteins, with ~1.25% (25344) non-zero values (Supplementary Table S1).

To better identify relations between associated proteins in the sparse bait-prey matrix (Figure 1A), we used singular value decomposition (SVD), to project the bait-prey matrix into a lower dimensional space, and assign eigenvalues to bait-prey pairs (Figure 1B). Application of SVD to the bait-prey matrix increases the overall similarity of prey protein vectors within the matrix as shown in Supplementary Figure S2. SVD in the form of latent semantic analysis was previously applied to detect analytical trends in the HUPO Plasma Proteome Project (HUPO PPP)[29]. Our goal in applying SVD to the bait-prey matrix is to increase the probability of detecting biologically relevant associations between proteins in the matrix. The extent to which the matrix dimension is reduced is determined by the *k*-value in SVD (Figure 1B). A high *k*-value corresponds to a small reduction of matrix dimensions with possible retention of too much noise, whereas a small *k*-value may retain too little information from the original matrix. We estimated an appropriate *k*-value by plotting the singular value at the k^{th} rank vs *k*-value (Supplementary Figure S3). A *k*-value of 150 was selected at the plateau in the graph so that the majority of significant information in the matrix is retained (singular values with rank < *k*). To identify relationships between prey proteins in the bait-prey matrix we then generated a prey-prey similarity matrix (Figure 1C) by computing the similarity between each pair of preys in the matrix (cosine similarity). Although a similar methodology could also be applied to the bait-bait matrix, we focused on the prey-prey matrix (5269 x 5269 proteins), because it is much larger than the bait-bait matrix (384 x 384 proteins), and we reasoned that it would represent a richer resource for discovery of

296 novel interactions and complexes. To identify similarities between prey proteins, we computed the
297 cosine distance between each pair of prey proteins for all prey proteins occurring with 2 or more baits
298 (2242 preys) (Figure 1C).

299 *Benchmarking bait-prey and prey-prey interactions*

300 To gauge the biological relevance of the bait-prey and prey-prey interactions and their associated
301 scores, we compared protein-protein pairs in our dataset to known datasets and measured the degree to
302 which protein pairs shared functional annotations. Since the bait-prey and prey-prey interactions
303 identified using our framework represent putative physical associations between proteins, we used two
304 complementary sources of protein interaction data as our benchmark sources. CORUM[22] is a map of
305 mammalian protein complexes curated from individual studies, whereas BioGRID[23] is a repository of
306 physical protein interactions including data from high-throughput interaction studies. CORUM groups
307 proteins according to protein complexes, whereas BioGRID represents protein interactions as protein-
308 protein pairs. To benchmark versus BioGRID, we first computed the 2-hop neighborhood of each
309 protein-protein pair in BioGRID. Since our data is derived from AP-MS experiments, representing
310 complexes of physically associated proteins that may or may not directly interact, we compared our
311 data to these neighborhoods of associated proteins rather than binary protein-protein pairs in BioGRID
312 to ensure a representative comparison. For each bait-prey or prey-prey protein pair, we determined
313 whether they co-occurred in a CORUM complex or within a BioGRID 2-hop network.
314 To compare bait-prey and prey-prey interactions to known interactions, we calculated log likelihood
315 scores for the relative enrichment of BioGRID and CORUM known interactions in our data, as
316 previously formulated[30]. Figure 2 shows the log likelihood of known interactions for bait-prey D-scores
317 (Figure 2A) and prey-prey scores (Figure 2B). Although the absolute numbers of interactions
318 overlapping between our data and BioGRID are higher than the overlap between our data and

CORUM, CORUM interactions represented in our bait-prey or prey-prey datasets have significantly higher D-score or prey-prey scores respectively. Ranked by prey-prey score, the 25th percentile CORUM prey-prey score is 0.44 whereas the 25th percentile BioGRID prey-prey score is 0.24. This is not unexpected, since CORUM is derived from manually curated protein complexes whereas BioGRID includes high-throughput protein interaction studies. This also demonstrates the value of comparisons of protein interaction data to multiple sources; CORUM provides higher specificity with low sensitivity whereas BioGRID improves the sensitivity at the cost of lower specificity. Figure 2 also shows that the relative enrichment of BioGRID or CORUM interactions decreases as bait-prey D-score or prey-prey score decrease, showing that both scores provide some discrimination of true positive interactions from false. In the case of the bait-prey interactions, BioGRID and CORUM interactions decrease very sharply below D-score ~ 20 (corresponding to ~ 95th percentile of all of the D-scores, as determined by Sowa et al.[7]). In the case of the prey-prey interactions, BioGRID and CORUM interactions decrease more evenly across the range of prey-prey scores. The large predicted size of the human protein interactome[30-31], the incompleteness of 'known' human protein interactions, as well as noise and context-sensitivity mean that intersections between experimental protein interaction datasets and known interactions tend to be small. For example, although BioGRID is a comprehensive source of available protein interaction data (~400,000 total interactions; ~38,000 human interactions)[23], 35% of the bait proteins used in our original AP-MS study[12] have 1 or fewer interacting protein in BioGRID (27% have no interactions at all). The overlap between bait-prey or prey-prey interactions and known interactions in the BioGRID set or in CORUM was 26.9% of bait-prey interactions (D-score>20) and 4.6% of prey-prey interactions (prey-prey score>0.75).

Biological coherence of bait-prey and prey-prey associations

To ascertain whether bait-prey and prey-prey protein pairs represent associations between proteins with

related functions, and to benchmark the bait-prey and prey-prey scores, we analyzed functional annotations of associated proteins using the Gene Ontology (GO)[32]. We first observed that computing co-annotation of GO terms for protein-protein pairs has low sensitivity, since many proteins, although biologically related may not be assigned the same term. We therefore used semantic similarity (SS) of GO terms which has proven to be robust measures of biological similarity for pairs or sets of genes[24]. Although there are multiple implementations of semantic similarity, here we use the Resnik max method[25], since it was previously shown to perform best in a comparative analysis of semantic similarity metrics using large-scale protein-protein interactions[33]. Semantic similarity vastly increases the sensitivity of analyzing co-annotations over simple analysis of co-annotated proteins pairs (Supplementary Tables S2 and S3).

For both bait-prey and prey-prey associations, we reasoned that gene-ontology semantic similarity scores should be higher for bait-prey or prey-prey pairs with higher D-scores and prey-prey scores respectively. GO semantic similarity scores were computed for each bait-prey or prey-prey protein pair (Supplementary Tables S2 and S3) and analyzed as follows. Bait-prey protein pairs were binned according to D-score: high (D-score>100), medium (100.00<D-score>20.00), low(20.0<D-Score>5.00), very low(D-score<5.00) and a randomly selected set (n=1000) of protein-protein pairs. Distributions of semantic similarity scores for bait-prey pairs are shown in Figure 3A for the molecular function GO ontology. The high and medium bins were found to be significantly higher than the random set (Wilcoxon test p-values : 1.5E-08, 3.3E-06, 0.59 for high, medium and low respectively). Similar trends were observed with the biological process and cellular compartment GO ontologies (Supplementary Figure S4). We used these analyses, along with the data shown in Figure 2A to calibrate the D-score threshold and therefore focused subsequent analyses on bait-prey pairs with D-score >=20.

Semantic similarity scores for all prey-prey associations were also computed and binned according to

366 prey-prey scores (Supplementary Table S3). Although not strictly monotonic, the log likelihood of
367 known interaction enrichment with respect to prey-prey score broadly increases as prey-prey score
368 increases (Figure 2B). We therefore grouped the prey-preys into four bins spanning the prey-prey
369 score(PPS) range high (PPS >=0.75), medium (0.75<PPS>0.50), low (0.50<PSS>0.25) and very low
370 (PSS<0.24) and compared with semantic similarity of GO molecular function as shown in Figure 3B.
371 A random set of 10000 prey-prey scores were generated to compute the significance test. Notably, the
372 high scoring bin (prey-prey score >=0.75) enriches for interactions with higher semantic similarity in
373 all 3 GO ontologies (Figure S3), and the difference between the high bin and the random set was
374 statistically significant (Wilcoxon Test p-values: 2.1E-12, 0.54 for high and medium respectively).
375 These results show that both the D-score and prey-prey scores can be used to define sets of bait-prey or
376 prey-prey pairs that are enriched for interactions with higher biological coherence, based upon their GO
377 annotations. These analyses also provide a guide for selected subsets of interactions for further
378 analysis, and as such we used bait-preys with D-score>=20 and prey-preys with score >=0.75 for
379 building network models of protein complexes. Approximately 5.6% (1900) of bait-prey interactions
380 and 7.7% (23,000) of prey-prey interactions meet these criteria, and were used in subsequent analyses.

381 *Identification of protein complexes*

382 Since AP-MS experiments identify co-complexed proteins, rather than binary pairs of interacting
383 proteins, we organized the data into more meaningful biological groupings by clustering sets of
384 proteins with significant prey-prey scores. In addition, since the sets of high-scoring bait-prey and prey-
385 prey interactions are large, and likely contain significant numbers of false positives, clustering provides
386 a means of focusing on higher-likelihood associations of proteins. The prey-prey similarity matrix was
387 hierarchically clustered as shown in Figure 4. We identified 107 clusters comprised of a total of 754
388 proteins (each cluster was required to have at least 3 proteins and all prey-prey associations >0.9) as

389 shown on the diagonal of the prey-prey matrix in Figure 4. For each of the 107 protein clusters, we
390 identified significantly enriched GO categories, and for 43 of the 107 protein clusters, one or more
391 significant ($p<0.05$) GO annotation terms were identified (Supplementary Table S4).

392 *Network models incorporating bait-prey and prey-prey interactions*

393 As demonstrated in the previous sections, the global biological coherence of prey-prey interactions is
394 similar to that of bait-prey interactions. Since matrix models of AP-MS data, in which all pairwise
395 interactions are assumed to occur, are prone to false positives[16] we sought to selectively combine high-
396 scoring bait-prey and prey-prey interactions into integrated network models. Protein clusters identified
397 in Figure 4 were used as network seeds and extended by addition of selected high scoring bait-prey (D-
398 score>20) and prey-prey (PPS>0.75) interactions. Four constructed network models, corresponding to
399 Eukaryotic Initiation Factor (EIF) complexes, G-protein signaling and regulation, chromatin assembly
400 factor complex and nucleosome regulation, and the proteasome are shown in Figure 5 (A-D
401 respectively) and were selected to illustrate the potential of combining bait-prey and prey-prey
402 interactions for delineation of network topology, and identification of new protein complex components
403 and interactions.
404 Figure 5A shows a network model constructed by integrating bait-prey and prey-prey interactions
405 corresponding to Eukaryotic Initiation Factor (EIF) complexes. Six bait proteins: EIF1B (also known as
406 GC20), EIF2B1, EIF3S10, EIF4A2, EIF4A1, EIF4EBP1 and their associated prey proteins were
407 integrated. Of particular interest is the separation of the cliques corresponding to EIF1/2/3 components
408 (red shaded nodes) and EIF4 components (green nodes). A large number of prey proteins (D-score>20)
409 were found for EIF1B and EIF2B1 baits, including many ribosomal proteins, in line with known
410 associations between these components and ribosomes[34]. In contrast, EIF4A2 and EIF4A1 baits had
411 relatively few high-scoring prey proteins as shown in Figure 5A. Of particular note, the LSM14A

412 protein was identified uniquely with EIF4A2 bait and with high prey-prey interactions with EIF4G1
413 and PDCD4. LSM14A is a component of P-bodies, cytoplasmic foci that govern mRNA storage and
414 degradation[35]. In line with previous findings, EIF4 components are present in P-bodies whilst other EIF
415 components are conspicuously absent[35]. PDCD4 is uniquely associated with EIF4A2 and EIF4A1 baits
416 in our data, and is known to interact directly with EIF4A2, EIF4A1 and EIF4G1[36].
417 Whilst EIF complexes are well represented in our dataset, we also analyzed protein complexes with
418 sparser coverage such as the networks in Figure 5B and Figure 5C. Figure 5B shows a network of
419 proteins involved in G-protein signaling. Members of the Rho family of GTPases (RHOB, RHOC) and
420 the Ras super-family of small GTP-binding proteins (RAC1, RAC2) are associated through their
421 common bait, ARHGDIA. All of these proteins function in vesicular transport and endosomal
422 signaling. An additional high-scoring interaction was observed with LYST, the lysosomal trafficking
423 regulator. Endosomes may ultimately fuse with lysosomes for degradation of constituent molecules. In
424 addition, LYST is associated with a rare lysosomal disorder, Chediak-Higashi syndrome, and thus the
425 association with endosome signaling may potentially shed additional light on the disease mechanisms.
426 We also observed a cluster of prey-prey interactions corresponding to the Chromatin Assembly Factor
427 (CAF-1) complex, and expanded this cluster of proteins into the network shown in Figure 5C. Several
428 protein complexes with known functions were identified. The CAF-1 complex and Polycomb
429 Recessive Complex 2 (PRC2) have related functions in chromatin metabolism and share components
430 such as RBBP4. The recently identified MMS22L-TONSL complex[37] that mediates recombination
431 mediated repair of replication forks and is comprised of MMS22L and TONSL (a.k.a NFKBIL2)
432 proteins was also identified. Other significant interactions between these proteins and the Tousled-like
433 kinases (TLK1, TLK2) were also observed. TLK1 and TLK2 heterodimerize and are also involved in
434 chromatin assembly[38]. Within this rich, overlapping network of complexes with related functions, we
435 searched for other high-scoring prey-prey interactions that might be novel components. Two other

proteins with significant prey-prey interactions with proteins in the chromatin modification network were observed. SMG6, a protein functioning in telomere regulation and nonsense-mediated mRNA decay was observed with high-scoring prey-prey interactions with several proteins (TLK1, 0.98; JARID2, 0.99). SMG6 is conserved (across eukaryotes) and has been found to physically interact with telomerase[39]. A second protein, UBN2, was recently identified as an ortholog of a yeast protein involved found in senescence associated chromatin foci[40]. The yeast ortholog of UBN2 interacts with the yeast ortholog of ASF1A, and both proteins function to create transcriptionally silent heterochromatin[41]. Thus, integration of bait-prey and prey-prey interactions serves to identify proteins linked to known complexes and functions.

Finally, we constructed a network based on coverage of the proteasome complex in our data (Figure 5D) . This network was constructed from four bait proteins and 16 prey proteins, highly enriched for components of the eukaryotic proteasome. Three of the baits used (PSMD6, PSMD10 AND PSMD13) function as 26S proteasome non-ATPase regulatory subunits of the proteasome and whereas one bait (PSMC4) is a 26S protease regulatory subunit. Although the structure of the eukaryotic proteasome is comparatively well understood, we note that the network model delineates sub components of the proteasome. For example, PSMC1, PSMC2 and PSMC5 cluster whilst PSMD subunits (PSMD8, PSMD12 and PSMD14) cluster separately. Although the coverage of protein-protein interactions is relatively sparse, we note that the alpha-type proteasome subunits, PSMA5 and PSMA6 cluster exclusively with PSMD subunits and not PSMC subunits, possibly indicating intermediate forms of the proteasome comprised of specific sets of components. In summary, integrated network models as shown by these examples have the potential to yield novel components of protein complexes as well as further delineating the topology and sub-structure of networks.

We next compared the biological coherence of integrated (bait-prey and prey-prey) and matrix models of equivalent complexes. Matrix models for selected baits consisted of all pairwise protein interactions

460 (bait-prey and prey-prey), whereas integrated models consisted of all bait-prey interactions with D-
461 score >20 and all prey-prey interactions with score>0.75. Semantic similarity distributions were used to
462 compare the integrated and matrix models of several protein complexes as shown in Figure 6. In
463 addition to the Eukaryotic Initiation Factor (EIF) and Proteasomal complexes, we also analyzed data
464 corresponding to four single baits (CTNNBIP1, VHL, REA and WDR8), the four most well replicated
465 baits in the original study. For all four single baits and in the EIF complex, semantic similarity scores
466 were significantly higher for integrated network models than for the matrix models. In only 1 case, the
467 proteasome, the integrated and matrix model showed no significant difference between the semantic
468 similarity distributions, suggesting that the biological coherence of integrated and matrix models of our
469 proteasomal data is similar. This may be explained by the fact that most of the prey proteins identified
470 by proteasome baits in our study are already known components of the proteasome, and so selection of
471 prey-prey interactions with high PPS score for incorporation in the network model does not improve
472 biological coherence over assuming that all preys interact with all other preys. In summary, these data
473 show that selective integration of high-scoring bait-prey and prey-prey interactions can be used to
474 generate protein network models with high biological coherence that can reveal new connections
475 between proteins as well as new components of protein complexes.

476 **Conclusions**

477 We present a data-driven framework for analysis of large-scale interaction proteomics data that uses
478 integrated computational techniques to derive additional value from these datasets in terms of novel
479 protein-protein interactions. We used this framework to analyze a unique, systematically generated
480 human AP-MS dataset, in which complexes were determined for 384 disease-relevant bait proteins[12].
481 Specifically, our analysis integrates associations between the baits and their identified proteins (bait-
482 prey), and associations that we detect amongst prey proteins (prey-prey), by analysis of the complete

data matrix. Analyzed globally, high-scoring bait-prey and prey-prey are enriched for known interactions and interactions between proteins with related biological function. In addition, by integrating prey-prey and bait-prey interactions into network models, we increase the biological coherence of those networks. We also show that integrated networks of bait-prey and prey-prey interactions provide a basis for delineation of network topology and identification of new protein complex members.

A major motivation for our work is to develop methods that enable novel protein-protein associations to be derived from large-scale datasets. Since mapping the complete human protein interactome experimentally is such an enormous undertaking, approaches that can be used to identify novel interactions from existing data, will continue to play a role in extending and refining the known human protein interactome. In addition, as previously observed[22], despite the rapid accumulation of large volumes of proteomics data, there has been surprisingly little re-analysis and evaluation of most proteomics datasets. The exceptions to this include unique datasets such as the yeast interaction proteomics datasets[2-5] that have fueled much of the development of scoring algorithms as well as analysis of interaction networks.

Other studies with similar intention to the current work have primarily focused on the relatively complete yeast AP-MS datasets, or on AP-MS datasets that although not complete, are focused on specific complexes or sets of complexes[6,20]. In addition, most methods that have been proposed for AP-MS data analysis are not appropriate to all datasets, for reasons of data size or completeness. The heterogeneity of current AP-MS datasets in terms of biological and analytical methodology (epitope tag, expression construct, organism, cell-type etc) along with the challenges associated with the data itself (missing data, noise, contaminant proteins) have hindered the development of widely applicable analysis tools. Exceptions to this include the SAINT algorithm where the explicit goal is to develop a widely applicable AP-MS analysis method[13].

With these challenges in mind, we have constructed a computational framework for analysis of large-scale human AP-MS data. Noteworthy features of our framework include representation of protein observations using the D-score that takes into account spectral counts, overall frequency, and replication for each protein observation. Although other studies with similar intent have made use of present/absent calls, particularly in analysis of the yeast AP-MS datasets, quantitative values derived via label-free analysis provide an additional dimension for ranking and discrimination of true positive interactions[20]. In the latter study, Choi *et al* used nested bi-clustering to group sets of baits or preys with similar quantitative spectral count levels. In our study, the dataset is more heterogeneous both in terms of the actual baits used as well as the number of replicate AP-MS experiments per bait. For this reason, rather than use the spectral counts directly for each protein, we used the D-score as the starting point for the analysis, so that frequency and number of replicates as well as spectral counts are taken into account. Second, we address the issue of sparseness of the primary bait-prey matrix by first transforming the matrix of bait-prey scores through singular value decomposition. Singular value decomposition in the form of latent semantic analysis has previously been used to large-scale expression proteomics albeit with a different goal[29]. In the latter study, SVD was used to analyze heterogeneous plasma proteomics datasets acquired using different technologies in different laboratories. The method used by Klie et al represented protein observations as binary measures, presumably so that data from different instruments and different laboratories could be appropriately integrated. Regardless of whether quantitative values are used however, tools such as SVD, that address the sparseness and missing data challenges of proteomics datasets are essential. SVD was also previously applied to the assembly of protein interaction networks from more focused human AP-MS data, although in this case, SVD was applied to the problem of identifying clusters of proteins[6].

To mitigate the problems of false positives in studying the large volume of prey-prey associations, we use semantic similarity measures of annotation to first calibrate our protein-protein scores (bait-prey

and prey-prey) and then to select small high-scoring subsets of protein-protein interactions to study. A principal challenge of data-driven inference of human protein-protein interactions remains the lack of 'gold-standard' protein interactions and complexes. Although curated annotations of mammalian protein complexes exist[22], these represent only a fraction of the total interactome. In the absence of gold-standard datasets, the semantic similarity measures provide a means of benchmarking protein-protein scores as shown here. Our method increases the amount of information in terms of protein-protein associations that may be gleaned from large-scale AP-MS studies. However, increasing the density of coverage (in terms of bait proteins used) will be the method of choice for truly defining protein complexes in the cell. Studies that iterate through a network, testing all proteins as baits in AP-MS experiments may provide the most detailed representations of protein complexes[8]. The complexity of mapping the protein interactome, in terms of distinguishing complexes that share components was highlighted in an AP-MS study of chromatin remodeling complexes[6]. This study reiterates the point that although computational and statistical analyses may allow for the prediction of protein-protein interactions and network topology, a detailed map of overlapping protein complexes may ultimately only be achieved through high density experimental analyses.

Integration of other data may further help refine of protein network topology. Although protein quantifications are only loosely correlated with protein-protein interactions, integration of protein abundance values may be used to refine prediction of protein interaction. The recently developed PaxDb[45] is a cross-species database of protein quantifications, thus providing an independent source ofproteome-wide abundance. In PaxDb, protein-protein interactions are used as a metric of consistency for quantitative proteomics studies, with the rationale that interacting proteins tend to have more similar levels of expression that randomly selected or non-interacting proteins[45]. Proteome-wide protein abundance values may also prove useful for normalizing the abundance values of proteins identified in AP-MS experiments. For example, a recent study used PaxDb values to account for the differing

555 cellular abundance of proteins identified in AP-MS analysis of chromatin remodeling complexes[46].
556 Future work might therefore utilize the approach that we have described in conjunction with other
557 proteome-wide information for further refinement of protein networks and protein interaction
558 discovery.

559 **Supplementary Material**

560
561 **Supplementary Figure S1** The density plot of top 100 frequent control proteins D-scores in the bait-
562 prey dataset.The top 100 frequent control proteins were selected from the 200 control experiments
563 (without bait), and the median D-scores of these 100 frequent control proteins in this study was close to
564 3.
565 **Supplementary Figure S2** The two-dimensional plot of prey-prey similarities scores using SVD in the
566 bait-prey matrix versus without using SVD. The application of SVD to the bait-prey matrix increases
567 the overall similarity of prey protein vectors within the matrix.
568
569 **Supplementary Figure S3** Selection of *K*-value for singular value decomposition of bait prey score
570 matrix.Singular values vs.*K* values was used to identify optimum *K* value=150. The value of K
571 determines the degree of reduction, a high K value corresponds to small reduction (minimal filtering of
572 noise), while a small K value correspond to strong reduction (too little information retained)
573
574 **Supplementary Figure S4** GO molecular function (MF), biological process (BP) and cellular
575 compartment (CC) semantic similarities comparison with bait-prey D-scores and prey-prey similarity
576 scores. A. Box plots represent distributions of semantic similarity scores for bait-prey protein pairs
577 binned according to D-scores (4 bins) for gene ontologies. Bin 1 (D-score>100) and bin 2 (100<D-
578 score>20) have statistically significantly higher semantic similarity than random set (R) of D-scores in

579 MF, BP and CC (Wilcoxon Test P values are 1.5E-8, 7.9E-6 and 1.7E-7 respectively in bin 1). B. Box
580 plots represent distributions of semantic similarity scores for prey-prey protein pairs binned according
581 to prey-prey scores (4 bins) for gene ontologies. Bin 1 prey-prey scores (>0.75) bin has significantly
582 higher semantic similarity than random set (R) of prey-prey similarity scores in MF, BP and CC (
583 Wilcoxon Test P values are 2.1E-12, 7.2E-12 and 5.0E-4 respectively) . Median of each distribution is
584 represented by horizontal bar in each box plot.
585
586 **Supplementary Figure S5** Exploratory Data Analysis plot of the distributions of prey-prey cosine
587 similarity values. Histograms represent the distribution of prey-prey cosine similarity values in the
588 original matrix (left) and in the null matrix under the resampling scheme (middle). The quantile-
589 quantile plot of prey-prey cosine similarities (right) shows the strong deviation of the two distributions
590 in the two situations.
591 **Supplementary Table S1** Bait-prey D-score table.Bait protein purifications name, prey protein
592 identified, and D-score.
593 **Supplementary Table S2** Comparison Bait-Prey (*A* and *B*) D-score with Molecular Function (MF),
594 Biological Process(BP) and Cellular Compartment(CC) GO terms semantic similarity.The column
595 names of this table are:Gene symbol *A*, Entrez gene ID *A*, Gene symbol *B*, Entrez gene ID *B*, Bait *A*-
596 Prey *B* D-score, Rank (based on Bait-Prey D-score, where 1=High (D-score>100); 2=medium (100<D-
597 score>20); 3=Low (20<D-score>5); 4=very Low (D-score<5)), Bait *A*- Prey *B* MF semantic similarity
598 score, Bait *A*- Prey *B* BP semantic similarity score, Bait *A*- Prey *B* CC semantic similarity score.
599 **Supplementary Table S3** Comparison Prey-Prey (*A* and *B*) similarity score with Molecular Function
600 (MF), Biological Process(BP) and Cellular Compartment(CC) GO semantic similarity. The column
601 names of this table are:Gene symbol *A*, Entrez gene ID *A*, Gene symbol *B*, Entrez gene ID *B*, Prey *A*-
602 Prey *B* similarity score, Rank (based on Prey-Prey similarity score(PPSS), where 1=High(PPSS>0.75);

2=medium(0.75<PPSS>0.5); 3=Low (0.5<PPSS>0.25); 4=very low (PPSS<0.25)), MF semantic similarity score, BP semantic similarity score, CC semantic similarity score.

Supplementary Table S4 Protein complexes identified.Modules name describing its components along with common bait proteins purification names, enriched GO terms, GO IDs , significant P-values.

Acknowledgments

We thank Joseph Abraham and Parminder Kaur for productive scientific discussions. R. M. E. acknowledges funds from the Cleveland Foundation used in part to fund this study.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

References

1. Saha S, Kaur P, Ewing RM. The bait compatibility index: computational bait selection for interaction proteomics experiments. J Proteome Res. 2010 Oct 1;9(10):4972-81.
2. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sørensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. 2002 Jan 10;415(6868):180-3.
3. GavinAC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002 Jan 10;415(6868):141-7.
4. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006 Mar 30;440(7084):637-43
5. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N,

Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabási AL, Tavernier J, Hill DE, Vidal M. High-quality binary protein interaction map of the yeast interactome network. Science. 2008 Oct 3;322(5898):104-10.

6. Sardiu ME, Cai Y, Jin J, Swanson SK, Conaway RC, Conaway JW, Florens L, Washburn MP. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. Proc Natl Acad Sci U S A. 2008 Feb 5;105(5):1454-9.

7. Sowa ME, Bennett EJ, Gygi SP, Harper JW. Defining the human deubiquitinating enzyme interaction landscape. Cell. 2009 Jul 23;138(2):389-403.

8. Goudreault M, D'Ambrosio LM, Kean MJ, Mullin MJ, Larsen BG, Sanchez A, Chaudhry S, Chen GI, Sicheri F, Nesvizhskii AI, Aebersold R, Raught B, Gingras AC. A PP2A phosphatase high density interaction network identifies a novel striatin-interacting phosphatase and kinase complex linked to the cerebral cavernous malformation 3 (CCM3) protein. Mol Cell Proteomics. 2009 Jan;8(1):157-71.

9. Glatter T, Wepf A, Aebersold R, Gstaiger M. An integrated workflow for charting the human interaction proteome: insights into the PP2A system. Mol Syst Biol. 2009;5:237.

10. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006 Mar 30;440(7084):631-6.

11. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics. 2007 Mar;6(3):439-50.

12. Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, Taylor R, Dharsee M, Ho Y, Heilbut A, Moore L, Zhang S, Ornatsky O, Bukhman YV, Ethier M, Sheng Y, Vasilescu J, Abu-Farha M, Lambert JP, Duewel HS, Stewart II, Kuehl B, Hogue K, Colwill K, Gladwish K, Muskat B, Kinach R, Adams SL, Moran MF, Morin GB, Topaloglou T, Figeys D. Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol. 2007;3:89.

13. Choi H, Larsen B, Lin ZY, Breitkreutz A, Mellacheruvu D, Fermin D, Qin ZS, Tyers M, Gingras AC, Nesvizhskii AI. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. Nat Methods. 2011 Jan;8(1):70-3.

14. Lavallée-Adam M, Cloutier P, Coulombe B, Blanchette M. Modeling contaminants in AP-MS/MS experiments. J Proteome Res. 2011 Feb 4;10(2):886-95.

15. Skarra DV, Goudreault M, Choi H, Mullin M, Nesvizhskii AI, Gingras AC, Honkanen RE. Label-free quantitative proteomics and SAINT analysis enable interactome mapping for the human Ser/Thr protein phosphatase 5. Proteomics. 2011 Apr;11(8):1508-16. doi: 10.1002/pmic.201000770.

16. BaderGD, Hogue CW. Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol. 2002 Oct;20(10):991-7.

17. Hart GT, Lee I, Marcotte ER. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics. 2007 Jul 2;8:236.

18. Yu X, Ivanic J, Wallqvist A, Reifman J. A novel scoring approach for protein co-purification data reveals high interaction specificity. PLoS Comput Biol. 2009 Sep;5(9):e1000515.

19. Geva G, Sharan R. Identification of protein complexes from co-immunoprecipitation data. Bioinformatics. 2011 Jan 1;27(1):111-7.

20. Choi H, Kim S, Gingras AC, Nesvizhskii AI. Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data. Mol Syst Biol. 2010 Jun 22;6:385

21. Deerwester S, Dumais St, Furnas GW, Landauer TK and Harshman R. Indexing by Latent Semantic Analysis. J. Am. Soc. Inf. Sci. 1990, 41(6), 391-407.

22. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res. 2010 Jan;38(Database issue):D497-501.

23. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2011 Jan;39(Database issue):D698-704.

24. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009 Jul;5(7):e1000443.

25. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: Proc. Of the 14th International Jpint Conference on Artificial Intelligence. Pp 448-453.

26. Fröhlich H, Speer N, Poustka A, Beissbarth T. GOSim--an R-package for computation of information theoretic GO similarities between terms and gene products. BMC Bioinformatics. 2007 May 22;8:166.

27. Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. Next generation software for functional trend analysis.Bioinformatics. 2009 Nov 15;25(22):3043-4.

28. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011 Feb 1;27(3):431-2.

29. Klie S, Martens L, Vizcaíno JA, Côté R, Jones P, Apweiler R, Hinneburg A, Hermjakob H. Analyzing large-scale proteomics projects with latent semantic indexing. J Proteome Res. 2008 Jan;7(1):182-91.

30. RamaniAK, Bunescu RC, Mooney RJ, Marcotte EM. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. Genome Biol. 2005;6(5):R40.

31. StumpfMP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. Proc Natl Acad Sci U S A. 2008 May 13;105(19):6959-64.

32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

33. Xu T, Du L, Zhou Y. Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. BMC Bioinformatics. 2008 Nov 6;9:472.

34. Pestova TV, Kolupaeva VG, Lomakin IB, Pilipenko EV, Shatsky IN, Agol VI, Hellen CU. Molecular mechanisms of translation initiation in eukaryotes. Proc Natl Acad Sci U S A. 2001 Jun 19;98(13):7029-36.

35. Eulalio A, Behm-Ansmant I, Schweizer D, Izaurralde E. P-body formation is a consequence, not the cause, of RNA-mediated gene silencing. Mol Cell Biol. 2007 Jun;27(11):3970-81.

36. Yang HS, Jansen AP, Komar AA, Zheng X, Merrick WC, Costes S, Lockett SJ, Sonenberg N, Colburn NH. The transformation suppressor Pdcd4 is a novel eukaryotic translation initiation factor 4A binding protein that inhibits translation. Mol Cell Biol. 2003 Jan;23(1):26-37.

37. Piwko W, Olma MH, Held M, Bianco JN, Pedrioli PG, Hofmann K, Pasero P, Gerlich DW, Peter M. RNAi-based screening identifies the Mms22L-Nfkbil2 complex as a novel regulator of DNA replication in human cells. EMBO J. 2010 Dec 15;29(24):4210-22.

38. Groth A, Lukas J, Nigg EA, Silljé HH, Wernstedt C, Bartek J, Hansen K. Human Tousled like kinases are targeted by an ATM- and Chk1-dependent DNA damage checkpoint.EMBO J. 2003 Apr 1;22(7):1676-87.

39. Redon S, Reichenbach P, Lingner J. Protein RNA and protein protein interactions mediate association of human EST1A/SMG6 with telomerase. Nucleic Acids Res. 2007;35(20):7011-22.

40. Banumathy G, Somaiah N, Zhang R, Tang Y, Hoffmann J, Andrake M, Ceulemans H, Schultz D, Marmorstein R, Adams PD. Human UBN1 is an ortholog of yeast Hpc2p and has an essential role in the HIRA/ASF1a chromatin-remodeling pathway in senescent cells. Mol Cell Biol. 2009 Feb;29(3):758-70.

41. Zhang R, Poustovoitov MV, Ye X, Santos HA, Chen W, Daganzo SM, Erzberger JP, Serebriiskii IG, Canutescu AA, Dunbrack RL, Pehrson JR, Berger JM, Kaufman PD, Adams PD. Formation of MacroH2A-containing senescence-associated heterochromatin foci and senescence driven by ASF1a and HIRA.Dev Cell. 2005 Jan;8(1):19-30.

42. Xu, H., Freitas, M. A., MassMatrix: A database search program for rapid characterization of proteins and peptides from tandem mass spectrometry data. Proteomics **2009**, 9, (6), 1548–1555.

43. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry *Anal Chem,* **2003***, 75*, 4646-58.

44. Nelson, E. K., Piehler, B., Eckels, J., Rauch, A., Bellew, M., Hussey, P., Ramsay, S., Nathe, C., Lum, K., Krouse, K., Stearns, D., Connolly, B., Skillman, T. & Igra, M. LabKey Server: an open source platform for scientific data integration, analysis and collaboration. *BMC Bioinformatics,* **2011***, 12*, 71

45. Wang, M.; Weiss, M.; Simonovic, M.; Haertinger, G.; Schrimpf, S. P.; Hengartner, M. O.; von Mering, C. *Molecular & cellular proteomics***2012**.10.1074/mcp.O111.014704

46. Tsai, Y.-C.; Greco, T. M.; Boonmee, A.; Miteva, Y.; Cristea, I. M. *Molecular & cellular proteomics***2012**, 11, M111.015156.

47. Nesvizhskii, A. I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* 2010, 73(11), 2092-123.

48. Lavallee-Adam, M.; Cloutier, P.; Coulombe, B.; Blanchette, M., Modeling contaminants in AP-MS/MS experiments. *Journal of proteome research* 2011, 10(2), 886-95.

49. Dazard, J. E.; Saha, S.; Ewing, R. M., ROCS: A reproducibility index and confidence score for interaction proteomics. *BMC bioinformatics* 2012, 13(1), 128.

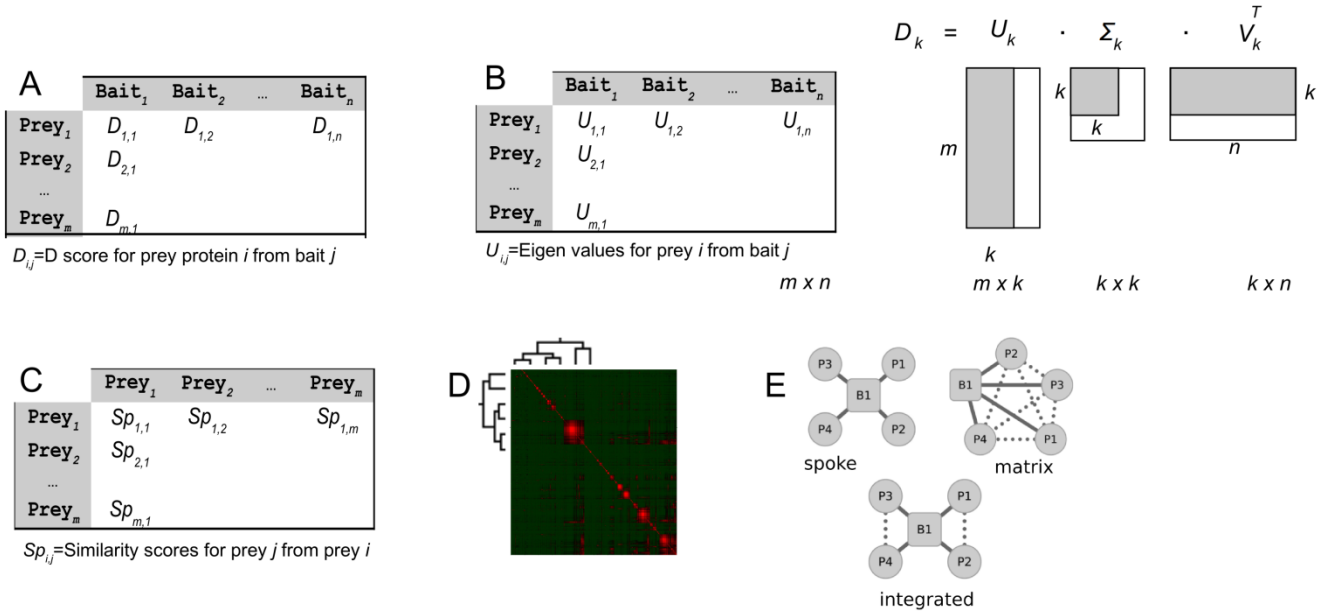

Figure 1. Data analysis work-flow. A. D-score[7] matrix with score for each pair of bait-prey proteins. **B.** D-score matrix approximation using singular value decomposition. **C.** Pairwise cosine similarity computed for each vector of prey scores. **D.** Topological overlap matrix created using hierarchical clustering to group prey proteins with similar prey-prey profiles. **E.** Spoke, matrix and integrated models for a hypothetical bait (B1) and 4 prey proteins (P1-P4). Integrated model incorporates selected edges from spoke and matrix models (solid lines lines represent bait-prey interactions, dotted lines represent prey-prey interactions). Figure 1B adapted with permission from (Klie et al, Journal of Proteome Research 7, 182-19). Copyright (2008) American Chemical Society.

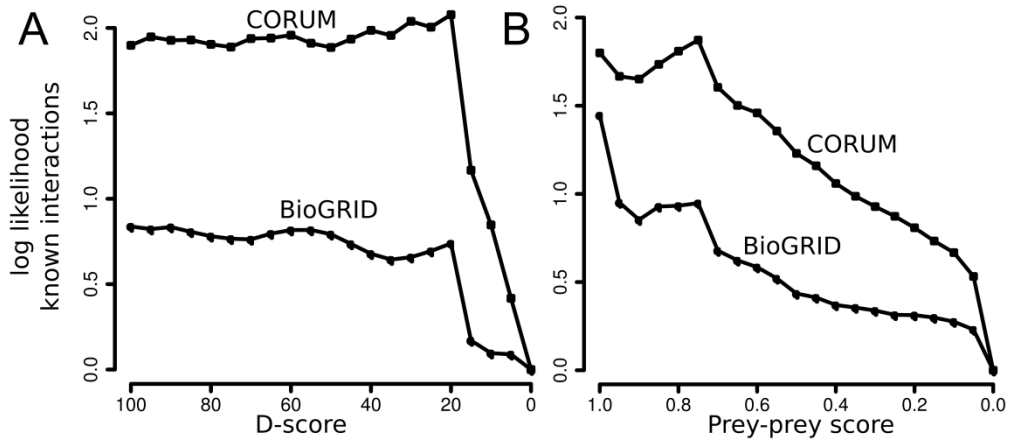

Figure 2. Comparison with known interactions. A. Bait-prey interactions compared to known interactions. Bait-prey protein pairs within shared 2-hop BioGRID network or present in same CORUM complex counted as 'known'. Log likelihood of relative enrichment of known vs. unknown interactions computed for each D-score threshold. **B.** As **A**, with log likelihood computed for each prey-prey protein pair and corresponding prey-prey interaction score threshold.

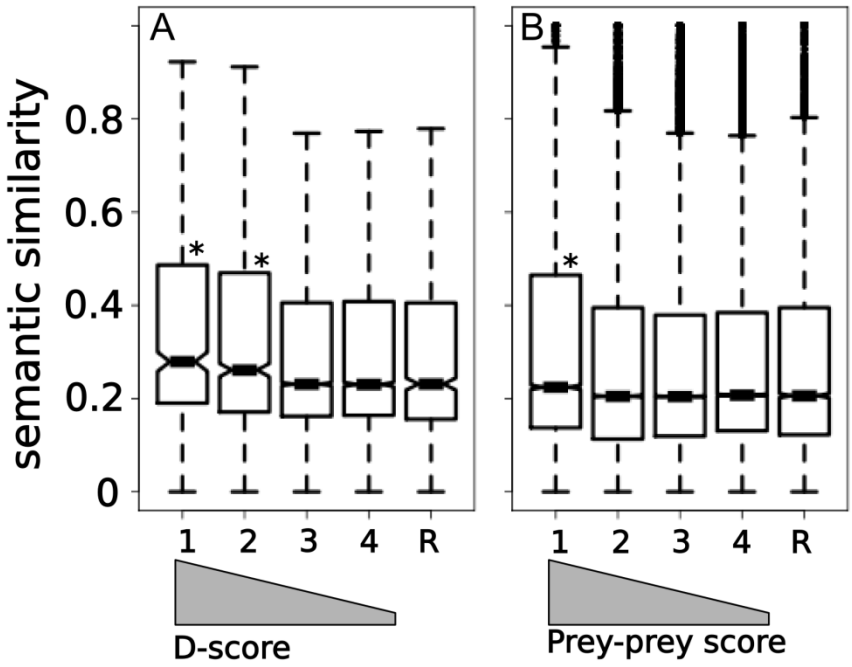

Figure 3. GO semantic similarity distributions and protein interaction scores. A. Box plots represent distributions of semantic similarity scores for bait-prey protein pairs binned according to D-scores (4 bins) for molecular function gene ontologies. Bins 1-4 represent D-score bins of D-score > 100, 100>D-score>20, 20>D-score>5 and D-score<5 respectively. Bins 1 and 2 have statistically significantly (*) higher semantic similarity than random set (R) (Wilcoxon Test P-values are 1.5E-08 and 3.3E-06). **B.** Box plots represent distributions of semantic similarity scores for prey-prey protein pairs binned according to prey-prey scores (4 bins) for molecular function (MF) gene ontologies. Bins 1-4 represent Prey-prey scores of >0.75, 0.75>PPS>0.5, 0.5>PPS>0.25 and PPS<0.25 respectively. Bin 1 has significantly (*) higher semantic similarity than random set (R) of prey-prey similarity scores (Wilcoxon Test P value is 2.1E-12). Median of each distribution is represented by horizontal bar in each box plot.

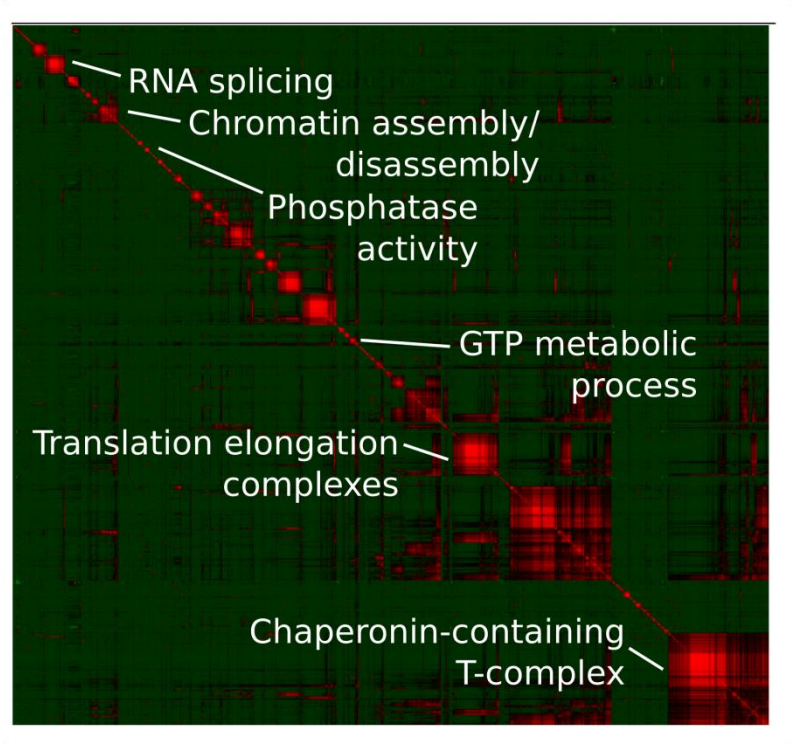

Figure 4. Topological overlap matrix of prey-prey interactions. The prey-prey similarity matrix (2242 X 2242) was used to create a topological overlap matrix using hierarchical clustering to group preys with similar prey-prey vectors. Red areas indicate prey-prey vector similarity and diagonal shows 107 protein modules. Selected modules are labeled with significant GO terms.

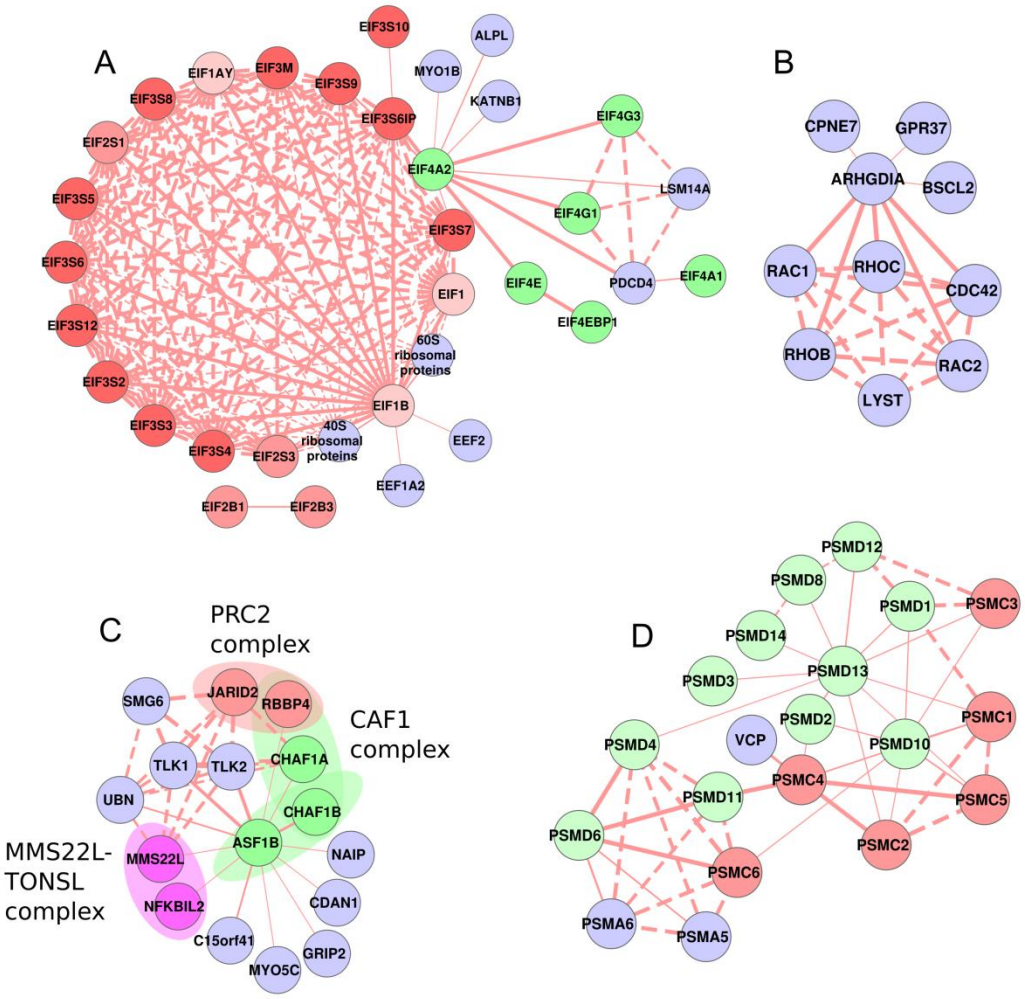

Figure 5. Selected network models integrating high-scoring bait-prey and prey-prey edges. A. Eukaryotic Initiation Factor (EIF) bait proteins (EIF1B, EIF3S10, EIF2B1, EIF4EBP1, EIF4A1 and EIF4A2) and their associated prey proteins (D-score>20) were combined with associated high-scoring prey-prey interactions. EIF1/2/3, EIF4 and EIF2 components are represented in red, green and gray nodes respectively. **B**. G-protein signaling complex. ARHGDIA bait and associated GTP binding proteins, RAC1 and RAC2; GTPase, RHOB and RHOC are identified preys with extensive inter-

connectivity (high prey-prey scores). **C**. Chromatin metabolism/ CAF1 complex, PRC2 complex and MMS22L-TONSL complex are shown in green, red and pink color respectively. **D.** Proteasome complex**,** four bait proteins corresponding to proteasome, PSMD6, PSMD10, PSMD13 and PSMC4 and 16 prey proteins forming highly enriched eukaryotic proteasome complex. The gray nodes correspond to proteasome core complex subunit (PSMA5/6); green nodes to 26S proteasome non-ATPase regulatory subunit (PSMD*); red nodes represents 26S protease regulatory subunits (PSMC*). Solid edges indicate bait-prey interactions and dashed edges indicate prey-prey interactions, edge thickness indicates bait-prey or prey-prey scores appropriately (thicker edge/higher score).

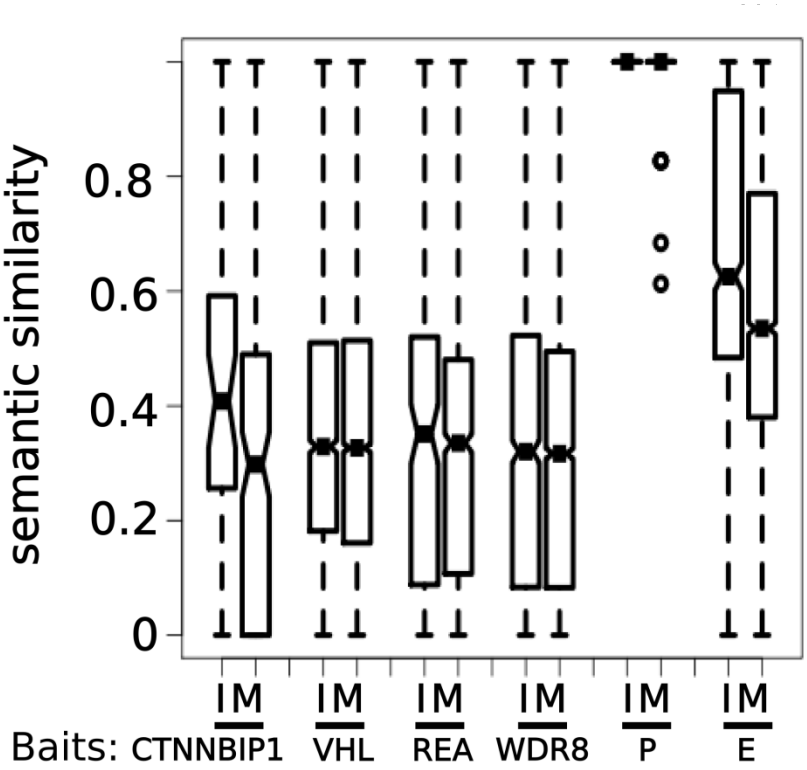

Figure 6. Biological coherence of integrated and matrix network models of protein complexes. Semantic similarity (GO biological process) distributions of protein-protein interactions in Integrated (I) and Matrix (M) models. Networks were constructed for data from 4 single baits (CTNNBIP1, VHL, REA and WDR8) and two models incorporating multiple baits, Proteasome complex (P) and Eukaryotic Initiation Factor (E) complexes. In 5 cases (CTNNBIP1, VHL, REA, WDR8, E), the Integrated model shows higher median semantic similarity than the Matrix model, and in 4 of these cases (CTNNBIP1,VHL,WDR8, E) the difference between Integrated and Matrix models is significant (Wilcoxon Test P-values<0.1).