Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions

John H Morris[1], Giselle M Knudsen[1], Erik Verschueren[2], Jeffrey R Johnson[2], Peter Cimermancic[2,3], Alexander L Greninger[4] & Alexander R Pico[5]

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California, USA. [2]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, California, USA. [3]Graduate Group in Bioinformatics, University of California, San Francisco, San Francisco, California, USA. [4]School of Medicine, University of California, San Francisco, San Francisco, California, USA. [5]Gladstone Institutes, University of California, San Francisco, San Francisco, California, USA. Correspondence should be addressed to A.R.P. (apico@gladstone.ucsf.edu).

Published online 2 October 2014; doi:10.1038/nprot.2014.164

By determining protein-protein interactions in normal, diseased and infected cells, we can improve our understanding of cellular systems and their reaction to various perturbations. In this protocol, we discuss how to use data obtained in affinity purification–mass spectrometry (AP-MS) experiments to generate meaningful interaction networks and effective figures. We begin with an overview of common epitope tagging, expression and AP practices, followed by liquid chromatography–MS (LC-MS) data collection. We then provide a detailed procedure covering a pipeline approach to (i) pre-processing the data by filtering against contaminant lists such as the Contaminant Repository for Affinity Purification (CRAPome) and normalization using the spectral index (SI_N) or normalized spectral abundance factor (NSAF); (ii) scoring via methods such as MiST, SAInt and CompPASS; and (iii) testing the resulting scores. Data formats familiar to MS practitioners are then transformed to those most useful for network-based analyses. The protocol also explores methods available in Cytoscape to visualize and analyze these types of interaction data. The scoring pipeline can take anywhere from 1 d to 1 week, depending on one's familiarity with the tools and data peculiarities. Similarly, the network analysis and visualization protocol in Cytoscape takes 2–4 h to complete with the provided sample data, but we recommend taking days or even weeks to explore one's data and find the right questions.

INTRODUCTION

AP-MS arose as a result of improved methods to enrich samples and perform separation chromatography, as well as advances in the resolution and sensitivity of mass spectrometers. AP-MS studies produce large amounts of information-rich data that detail protein-protein interactions in a variety of organisms and biological systems. These interactions help characterize the functions of proteins, provide detailed catalogs of proteins involved in protein complexes and biological processes, and can reveal networks of biological processes at local and proteome-wide scales. By bridging these molecular measurements to the systems level, we can better understand the genetic, epigenetic and protein-based associations of these proteins with disease. Today, the standards for analyzing protein-protein interactions span a wide spectrum that includes well-established protocols for sample preparation, diverse interaction scoring and clustering algorithms, methods for graph theory and data mining, and biological networks. This protocol describes how to analyze AP-MS data to produce meaningful networks. It presents some of the current thinking and practices in MS and network biology and is intended to guide practitioners, mentors and instructors in these fields.

All function is local

Analysis of protein-protein interactions can yield insight into the functional relationships between proteins. For example, the interactions between co-regulated metabolic enzymes can reveal the assembly of higher-order noncovalent protein complexes, called metabolons, which channel products to substrates along sequential steps of a metabolic pathway[1,2]. Protein-protein interactions can also reveal relationships between substrates and post-translational modifying enzymes, such as kinases, phosphatases, acetyltransferases and proteases, or with coactivator/co-repressor assemblies that modulate the specificity and activation of core complex functions[3–11]. Recent analyses of protein-protein interactions between host and pathogen have yielded new information on how pathogens use cellular processes for their own replication, division, budding, invasion and immune evasion. These analyses revealed basic cellular processes, such as membrane organization and protein trafficking[12–20]. Analysis of protein-protein interactions is orthogonal, but also complementary, to genetic interaction experiments; both provide lists of candidate interactions and implicate functional relationships[21–24].

AP-MS has become a standard method for discovering protein-protein interactions. Traditional methods couple native protein immunoprecipitation with immunological or mass spectrometric detection. AP-MS methods, however, use epitope tags on target 'bait' proteins of interest as affinity capture probes for the identification of the coassociating 'prey' proteins, without requiring purchase or development of specific antibodies for each new bait protein. Furthermore, antibodies raised against proteins of interest, or bait, may disrupt protein-protein interactions if the epitope coincides with a protein interface needed for an interaction. Affinity tagging has been applied most extensively in yeast, in which over 70% of the expressed proteins have been mapped into protein-protein interactions, which can be resolved into ~500 major complexes by clustering and network analysis[25–27]. Numerous protein complexes and sets of proteins from humans, *Drosophila* and *Arabidopsis* have also been analyzed[28–33].

The goal of this protocol is to provide a template for interpreting protein-protein interaction data derived from AP-MS experiments. By analyzing these data, rich networks can be created to detect even transient interaction partners that may be present at low expression levels. As the quality of the network depends on that of the interaction data, our protocol begins with an overview of selecting the proper epitope tag, cell line, biochemical conditions and instrumentation for experimentation. After these considerations, we provide step-by-step guidelines for scoring AP-MS data. We then place these data into context by comparing and integrating them with available large data sets that contain known protein-protein interactions and Gene Ontology (GO) annotations. Here we work within Cytoscape[34,35], a common tool used to visualize and analyze networks. We discuss various approaches to integrate and analyze data, tailored for AP-MS data, as well as common research questions, and then we conclude with steps for generating high-quality images of interaction networks for publication figures.

Experimental considerations

Bait selection. The first step to designing an AP-MS study is to select the proteins of interest, or baits, that will be used to characterize a network of protein-protein interactions. As the proteins in this set of baits are scored against each other, they should be selected to maximize the likelihood of identifying interactions that are unique to the bait compared with other proteins in the set. For example, if a bait protein contains an RNA-binding domain, then at least one other negative-control protein containing an RNA-binding domain should be included. This approach helps identify proteins that uniquely interact with the protein of interest compared with the class of RNA-binding proteins in general.

Controls. Within each bait set, include a positive control and a negative control. Positive-control bait may be a protein that previously underwent AP-MS analysis and that has a high-confidence set of interacting proteins identified by reciprocal binding assays. For example, the HIV-1 Vif protein is well established to enrich for elongin-B and elongin-C, cullin-5, Rbx2 and CBFb[10]. GFP is recommended for a negative-control bait that is not expected to have specific interactions in most organisms[10]; any interactions with this protein may be related to the epitope tag and resin-capture system.

Cell lines. Ideally, bait proteins would be expressed in the background of the interactome targeted for interrogation. However, an important consideration is that the input material needed for AP and MS analysis is reasonably large (<25 million cells yield microgram-scale quantities of expressed bait protein)[13,17,36]. Thus, the experimental design must balance optimizing bait expression with maintaining relevance to the biology or tropism of the proteins of interest. For the purposes of rapid and high-throughput screening, transient transfection of an epitope-tagged protein is a rapid approach that is also amenable to genetic manipulation through mutagenesis or alternate tagging strategies[10].

Unfortunately, the variety of cell lines amenable to transient transfection is limited. Thus, depending on the scale of the study, one may need to generate a more relevant and stable cell line that expresses the bait proteins. Stable cell lines may be engineered to incorporate inducible expression, such as using a line that stably expresses the Tet-repressor protein. This method may optimize the expression of the bait protein to approximate endogenous expression levels[37]. Engineering stable expression systems may require the development of multiple clones to select for optimal expression levels, which generally requires much more time[36]. Another possibility is to express and purify a bait protein in one heterologous expression system, such as HEK293, and then incubate this bait *in trans* with the desired lysate. This method is less desirable, as proteins of interest are not expressed *in vivo* but are extracted and bound to isolated protein *in vitro*, thereby requiring relevant interactions to be stably formed in a lysate[38].

Genome engineering approaches, such as clustered regularly interspersed short palindromic repeats (CRISPR) or transcription activator–like effector nuclease (TALEN) technologies, can also be considered when generating cell lines that express affinity-tagged proteins[39]. These methods help directly fuse affinity tags to proteins of interest within their locations in the genome, thus maintaining genomic contexts that regulate endogenous gene expression levels. Currently, these approaches are somewhat specialized and can be expensive, depending on the scale of the study; however, they are rapidly improving and will probably become standard procedure in many laboratories within a few years. Furthermore, strategies for cellular reprogramming convert nearly any cell type into an induced pluripotent stem (iPS) cell that can be differentiated into a number of specialized cell types. These iPS cells can be generated from cell material taken directly from patients. iPS cells can also be subjected to genome engineering (e.g., introducing a single point mutation), which maintains an isogenic background for comparing mutant and wild-type proteins of interest.

Affinity tags. A number of synthetic or naturally occurring affinity tags have been implemented in AP-MS studies. Common epitope tags include FLAG, Strep, Myc, hemagglutinin, protein A, His6-tag, calmodulin-binding protein, GFP and maltose-binding protein, which are reviewed in ref. 40. In some cases, multiple tags are fused together, such as 2×Strep3×FLAG, which is widely used by the authors of this protocol and which contains five tandem affinity tags of two types. Most combinations of these tags would be sufficient for AP-MS analysis, but the conditions required to bind and elute different tags may be substantially different and should be identified in the literature. For example, efficiently eluting a 3×FLAG-tagged protein from anti-FLAG resin requires detergent in the elution buffer. Notably, we recommend tagging both the N and C termini, in case one terminus disrupts interactions or protein function. Our tag of choice is generally the 2×Strep tag for lower reagent costs and ability to elute with a small molecule that is LC–tandem MS (LC-MS/MS) compatible (desthiobiotin), rather than requiring competition with an excess of peptide (such as for FLAG elution). Selection of a tandem affinity purification (TAP) versus a single-affinity step strategy is usually made when highly refined or stable complexes are sought (reviewed in ref. 41). On the other hand, a rapid capture-release strategy with a single affinity step is expected to produce more interaction candidates, including interesting transient interactions but also with additional false positives[41]. We note that every tag has its own specific background protein profile, such as biotin-cofactor enzymes (e.g., carboxylases) associated with the Strep tag system or STK38 and the associated CRAPome for FLAG-tagged baits[42].

Affinity purification. Early AP-MS approaches used tandem affinity tags to perform sequential APs that yielded highly pure protein complexes that could be characterized by MS[25]. At that time, analyzing a complex protein mixture with MS platforms was limited, which caused nonspecifically interacting proteins in the purification to obscure the identification of true members of a protein complex. However, more stringent purification procedures generally require substantially more starting material and include strict wash steps. Unfortunately, these stringent conditions may exclude identifying interactions that are transient or that occur with low-abundance proteins.

More modern MS platforms are increasingly sensitive in characterizing mixtures of complex proteins. In parallel, using single-step purification procedures, magnetic beads rather than agarose resins (faster wash times), cryogenic lysis of the cell pellet or in-line chromatographic strategies, potentially with microfluidic devices, can all improve sample recovery and manipulation speed for capturing transient interactions[8,43–45]. Unfortunately, these methods also increase the degree to which nonspecifically interacting proteins are represented in the final sample analyzed by MS[29,36,46]. Thus, the challenge in protein network analysis has shifted from optimizing methods to obtain highly pure protein complexes to developing bioinformatics methods to differentiate true protein interactions from background. This challenge is the focus of the following section.

We recommend that the purification be eluted in-solution, rather than, for example, eluted by boiling the affinity resin with SDS sample buffer and running the sample into a gel. First, strategies that maintain the purified sample in-solution commonly use competitive elution that specifically elutes the bait protein from the resin (e.g., eluting from anti-FLAG resin with FLAG peptide and eluting from Strep-Tactin resin with biotin). Eluting with SDS sample buffer is highly nonspecific and includes proteins that bind nonspecifically to the naked affinity resin. Second, in-gel digestion procedures are time-consuming and inefficient. For example, hydrophobic peptides may be difficult to extract from gel pieces, and gel lanes are often separated into multiple slices that require a separate MS analysis[47].

We also recommend that gel electrophoresis and silver staining be performed to assay the quality of expression and AP, as well as complexity of the eluate. Although the sensitivity of mass spectrometers has now advanced beyond silver staining, a band at the predicted molecular weight of the tagged bait proteins is almost always readily detected by silver staining. In addition, the availability of a silver-stained gel allows one to confirm by MS the true molecular weight of identified proteins, such that alternative splice forms or otherwise modified proteins can be accurately identified. Western blots are also recommended to verify that epitope-tagged proteins are expressed and seem intact.

MS analysis. Sample preparation methods are fairly standardized and generally include denaturation of proteins with chaotropes such as urea or guanidinium hydrochloride, reduction of disulfide bonds with a reducing agent, alkylation of free cysteine residues and proteolytic digestion with trypsin or an alternate protease such as LysC[48,49]. After digestion, samples are typically desalted with C18 cartridges or tips to remove salts and other digestion impurities.

Commonly, the MS platform for AP-MS analyses is implemented by LC-MS/MS. There are countless platforms capable of identifying proteins, most of which include an up-front, in-line HPLC equipped with C18 chromatography columns. These columns separate peptides with an organic gradient that elutes directly into the mass spectrometer coupled with a nano-electrospray source for ionization. Data are generally acquired by standardized methods for data-dependent acquisition that collect the maximum number of informative mass spectra within a short time. High-resolution and high mass accuracy systems (e.g., systems equipped with Orbitrap or time-of-flight mass analyzers) have been used in many studies, but they are not absolutely required. Low-resolution systems are also excellent for most AP-MS applications, and they are often more sensitive than high-resolution systems. To identify HIV-human protein-protein interactions, our laboratory regularly uses a Thermo Scientific Velos Pro dual linear ion trap system that replicates work performed using an Orbitrap XL system[12]. Cases in which a high-resolution system would be required include assessing quantitative differences in protein interactions, such as the response of protein interactions to drugs, or comparisons between protein mutations that may cause more subtle changes to protein interactions. Stable isotope-labeling strategies may also be appropriate for obtaining quantitative comparisons of interactions and are reviewed elsewhere[50].

Matching raw MS data to peptide sequences can be performed using a number of proteomics software packages that include suitable scoring for ranking the probability of correct peptide identification and protein inference. Common packages include commercially licensed Mascot[51] and SEQUEST[52], and open-source MaxQuant[53], X!Tandem[54] or Protein Prospector[55] packages. All of these database search programs are suitable for MS data processing; our laboratories use the Protein Prospector suite. It is recommended that raw MS data be searched against a database concatenated with a 'decoy' database containing all the original database sequences randomized or reversed in amino acid sequence order[56]. Decoy hits can be used to estimate the false-positive rate of peptide and protein identifications, and thus they can be used to select a peptide identification score cutoff that results in a desired false-positive rate. Generally, a protein false-positive rate of 1% is accepted in the field. See the review in Nesvizhskii[57] for details about matching raw data to peptide sequences.

After these steps, which are commonly done by a core facility or collaborators with expertise in MS–based proteomics, users should obtain a table that lists for each sample the accession numbers and descriptions for proteins identified and the number of peptides identified for each protein, as well as supporting database search statistics commonly reported as an expectation value for evaluating the quality of the identification. Another output is a measure of peptide, and therefore protein abundance, such as the 'spectral count', which is the number of total MS/MS spectra assigned to a protein in a sample. Spectral counting approximates the relative protein abundance in samples, which can be used for comparative statistical analysis of protein enrichment across multiple AP-MS experiments[58,59]. For data obtained on high-resolution instruments, the peptide measurements can also be output as either intensity or peak areas for a given peptide species using label-free quantification results. This method more

quantitatively measures peptide and thus protein abundance in different samples.

Further considerations. Although AP-MS may reveal interacting proteins, it does not fully detail the assembly of the protein complex. This limitation is because AP-MS does not easily distinguish direct from indirect interactions (i.e., with intermediate interactors between bait and prey proteins), and it does not resolve highly connected proteins that may participate in multiple distinct complexes and cellular functions. Furthermore, AP-MS is biochemically constrained and does not necessarily capture proteins in a natural physiological state or consider compartmentalization and other features of the cell. These limitations can yield misleading results. Inferences around the association rate, stability and stoichiometry of complex components may eventually be accessible with the incorporation of more quantitative proteomic methods[60]. However, additional experiments are still required to elucidate the molecular interactions of complex formation. Confirmatory binary interaction analysis, obtained through two-hybrid methods or high-density mutagenesis mapping of interactions, can help resolve the direct protein-protein interactions. Now, standard mammalian two-hybrid analysis is performed in relevant cell lines[61] and has served our own projects in confirming interactions between candidate virus and human protein-protein interactions[46,62]. To reach higher levels of molecular precision, high-resolution biophysical analyses, such as calorimetry, sedimentation analysis, fluorescence monitoring, surface plasmon resonance and other structural methods can be used.

The sample data sets

For the purposes of this protocol, we consider two different sample data sets: the recently published data on the interaction between the host and HIV proteins by Jäger *et al.*[12] and a larger network of yeast protein-protein interactions by Collins *et al.*[63].

The Jäger data set is characteristic of a focused, quantitative AP-MS experiment that has relatively few bait proteins and pulls down a larger set of prey proteins. Although these particular data resulted from an experiment on host-pathogen interactions, similar data sets might result from experiments looking for targeted protein-protein interactions, potentially under different biological conditions. We take the Jäger data set from 'raw' AP-MS results (**Supplementary Data 1**) to a normalized, filtered and scored data table (**Supplementary Data 2**). This table is imported into Cytoscape for network analysis and visualization. The goal in analyzing the network found from a low-density data set of this form is generally to understand the biological context of the interactions and to form hypotheses about biological function. This typically involves augmenting the network with other known interactions, as we do in the network analysis protocol.

The Collins data set (**Supplementary Data 3**) is a combination of two large AP-MS yeast interactome experiments[26,27,64], which we have chosen on the basis of its public availability and high quality. This data set (and the two it combines) attempts to systematically determine a large portion of the interactome of a species, which involves using all proteins as bait proteins to get all possible combinations of proteins. In analyzing higher-density data sets such as these, the goal is often to find or confirm complexes[65] and to analyze the relationship between nodes and edges in the network (network topology) to find 'hubs' in the network[66]. We thus refer to this data set in the clustering steps of the network analysis protocol and produce a Cytoscape session file with the clustering results (**Supplementary Data 4**).

A step-by-step tutorial following the network analysis procedure (Steps 11–33) in Cytoscape is available as **Supplementary Methods**.

Computational scoring pipeline

It is useful to categorize raw AP-MS results into four classes of protein-protein interactions: (Class I) interactions that occur in the cell (i.e., biologically relevant complexes); (Class II) physically existing interactions that do not occur in the cell and that are only observed as an artifact of sample preparation (e.g., cell lysis allowing proteins from different compartments to interact); (Class III) interactions involving contaminant proteins; and (Class IV) physically non-existing interactions detected by error. The goal of scoring is to reliably highlight interactions of Class I on the basis of interaction properties (features) measured or derived from single and cross-experiments to diminish Class II interactions, and then to use multiple approaches to filter out Class III and IV errors. Although some experimental techniques can discriminate between biologically relevant and irrelevant interactions (e.g., stable isotope labeling by amino acids in cell culture (SILAC) labeling, larger numbers of technical and biological replicates), these approaches are not readily available on a large scale. Instead, a variety of computational scoring pipelines have been developed to identify biologically relevant interactions among a large number of irrelevant interactions in raw AP-MS data. These scoring pipelines usually comprise three stages: pre-processing, scoring and testing.

Pre-processing. In the pre-processing stage, we recommend a number of quality control steps to clean up the list of identified prey. Common filtering procedures involve removing known contaminants, as well as protein groups that have no uniquely identified peptides assigned to them. Contaminant or background proteins can originate from the epitope-tagging system, the solid support or other biological factors such as protein misfolding. For example, avidin captures both Strep-tagged constructs and endogenously biotinylated proteins, and FLAG-tagged capture has highly specific, reproducible interactions with a cadre of proteins, called 'frequent fliers'[67]. The recently published CRAPome[42] compiles a freely available list of contaminants identified by the AP-MS community. Another, more elusive, source of contamination is from 'carry-over' proteins that were left behind during previous runs on a given machine. Carry-over contamination is higher when a particular protein is overexpressed to increase the efficiency of the AP, and it is longer lasting when the protein is hydrophobic. Although this problem can be addressed by performing elaborate washing conditions, we found that additional *in silico* filtering steps minimize the impact of carry-over. A computational approach to this problem is to sort all MS results in the order they were run and to scan for half-life-like patterns of decreasing raw values (e.g., intensities, spectral counts or number of uniquely identified peptides) in consecutive runs.

A crucial step in pre-processing is normalizing the raw MS values so that they can be used as a quantitative feature to average or compare the abundance of a given protein across different APs. Raw protein measurements derived from AP-MS data include the

TABLE 1 | AP-MS scoring algorithms.

Algorithm	Pros	Cons
MiST	Intuitive feature set	Bias toward 'one-hit-wonders' or background proteins when over-weighing specificity or reproducibility, respectively
	Includes specificity as feature	Works best with a larger number of baits
	Robust performance	
	Configurable feature weights	
SAInt	Discriminates between background or true interactions for abundant proteins	Poor prediction of low-abundance and specific-interacting proteins
	Configurable parameters	May be computationally expensive with large data sets
	Robust performance on any data set size	
CompPASS	Intuitive feature set	Convoluted formula without feature weights
	Includes specificity as feature	Unbalanced scores for unevenly distributed number of replicates in a data set
Z-score	Simple and intuitive	Limited feature set
		Does not work on small data sets

Comparison of pros and cons for a set of scoring algorithms commonly used with AP-MS data.

number of unique peptides, spectral counts and MS1 intensities. Unfortunately, data from different AP-MS runs (or replicates) contain inherent biases and variations, resulting in a signal corrupted by systematic or even random changes[68]. Therefore, properly quantifying and normalizing AP-MS runs allow the user to directly compare runs. Several quantification approaches have been developed, such as the SI_N, exponentially modified protein abundance index, NSAF and distributed NSAF[69]. The SI_N and NSAF scores are the most linear and reproducible across different technical and biological replicates[69], and they were chosen for use in this protocol. The result of this step is a 2D bait-prey matrix of runs (columns) by all detected proteins (rows) with values corresponding to an abundance metric that quantifies each bait-prey interaction.

Before proceeding to the scoring stage, we recommend comparing all replicate AP-MS runs to identify potential errors (e.g., bait annotation mistakes and technical errors) and to collect more data on pull-downs with baits that show a high degree of variation. Multiple biological and/or technical replicates of a sample (we recommend at least three, and in some cases up to seven replicates, depending on sample availability and instrument time) are required to ensure more accurate and comprehensive determination of biologically relevant interactions[12,68]. To compare all AP-MS runs, use any distance metric (e.g., Pearson product-moment correlation coefficient and Euclidean distance metric) on the 2D bait-prey matrix to analyze the replicate correlation with unsupervised clustering (e.g., hierarchical clustering) and heat map visualization. The result of the pre-processing stage is therefore a quantified, normalized and filtered 2D bait-prey matrix of interaction data from which a number of biologically irrelevant Class III and IV interactions have been removed.

Scoring. In the scoring stage of the pipeline, AP-MS bait-prey pairs are assigned scores to help distinguish biologically relevant interactions from those that are irrelevant. Several scoring methods have been developed, including the following: (i) *Z*-score for calculation of specificity[46]; (ii) significance scores for pairwise co-occurrence of interactions based on randomly shuffled preys[70]; (iii) a combination of machine-learning algorithms using probabilistic mass spectra scores and measurement of reproducibility[27,63]; (iv) a composite score consisting of bait-prey abundance, specificity and reproducibility (CompPASS[32] and MiST[10]); and (v) a mixture model with Bayesian statistical inference (SAInt[71–73]). The performance of these algorithms depends on the nature of the data. For example, scoring methods are affected differently by the topology of the protein-protein interaction network, the number of baits, replicates and control experiments, the level of expected contamination and the size of the data set. Refer to **Table 1** for a comparison of the performance of scoring algorithms taking these factors into account. In practice, multiple scoring algorithms should be applied to a given AP-MS data set and then assessed on a case-by-case basis in the testing stage.

Testing. The testing stage resolves which (if any) of the scoring methods, or their composites, most accurately predicted the biologically relevant bait-prey pairs. Testing should include one or both of the following strategies: (i) evaluation of a benchmark of known 'ground truth' bait-prey pairs, or (ii) selection of a subset of the unknown bait-prey pairs with the best scoring from different rankings, followed by additional characterization through orthogonal experiments (for example, an interaction can be validated by coimmunoprecipitation studies using antibodies against the prey protein). In the Jäger paper[12], the authors applied the

MiST, SAInt and CompPASS scores to our AP-MS data set, and then tested it on a benchmark of 39 known HIV-human protein-protein interactions. They found that the MiST score is the most accurate among all the tested scores. For example, at the threshold of 0.75 (resulting in 387 bait-prey pairs from experiments done in HEK cell-line), the recall number of known bait-prey pairs for the SAInt, CompPASS and MiST scores was 19, 29 and 32, respectively. At the same threshold, the recall number of bait-prey pairs involving ribosomal proteins (most probably Class II–IV interactions) for the MiST score was only 3, compared with 32 and 75 for SAInt and CompPASS, respectively. In the detailed protocol that follows, we outline the steps that transform a raw AP-MS data file with Prospector search results (**Supplementary Data 1**) into a set of scored bait-prey interactions (**Supplementary Data 2**) by the MiST algorithm.

Data transformation and network analysis

After the scoring pipeline, our AP-MS data were transformed from a raw set of values to a quantified, normalized and filtered 2D bait-prey matrix and then to a table of accurately scored bait-prey interactions (**Supplementary Data 2**). This table is a common result format for AP-MS interaction studies, which is the input for the network analysis and visualization of our data. This table is also an acceptable format for data deposition into public repositories, such as IMEx (http://www.imexconsortium.org/submit-your-data). Public deposition is strongly recommended—and sometimes required—before publishing related findings. In general, making interaction data publically accessible improves many of the methods and resources used in the scoring and network analysis procedures presented in this protocol.

By nature, the results of an AP-MS experiment are networks. Cytoscape is a widely used tool for analyzing and visualizing biological networks (http://cytoscape-publications.tumblr.com/archive). This tool has been used for analyzing expression data[35] and genetic interactions[74], as well as protein-protein interaction data derived from yeast two-hybrid experiments[75] and AP-MS[27]. An advantage of Cytoscape is its open architecture, which allows its core functionality to be extended by the development of apps. An analysis of an AP-MS network might use several of these apps, all of which may be accessed through the Cytoscape App Store (http://apps.cytoscape.org).

The network analysis section of the protocol below follows what a Cytoscape user might do to analyze an AP-MS data set; however, the specific steps might be performed in a number of different packages or platforms. Bioinformaticians might use Python scripts, R, MATLAB or any number of different tools along the way. Cytoscape is a convenient tool for end users, but it is not the only way to analyze and visualize a data set.

Network analysis. The fundamental goal of analyzing the scored AP-MS data is to form or confirm hypotheses about cellular function. This might involve enriching the data set with known interactions or determining the functions of specific interactions or groups of interactions. For our protocol, we assume that a bait protein uniquely and independently binds to each prey protein (i.e., the 'spoke' model)[65]. For our low-density data, the protocol outlined below augments the network to find evidence in known protein-protein interaction databases (e.g., BioGRID[76], Reactome[77], STRING[78], CORUM[79] and IntAct[80]) that suggest a more complex interaction. These additions to the network are useful for forming hypotheses, but we strongly recommend that the user conduct confirmatory experiments to determine the exact nature of the complex. This may be particularly important in perturbed systems (e.g., diseased or infected cells) in which the publically available protein-protein interaction databases might only apply to normal cells. For our high-density data (such as our Collins example; **Supplementary Data 3**), we have often used the prey proteins as bait proteins so that we can assume the spoke model and use clustering approaches to determine when the interactions suggest complexes[50,81].

In the protocol below, we use several methods to understand the data set, although this is not a complete set of available network analysis techniques. To understand the scope of the available approaches, users can refer to recent reviews[28,50,82]. For our protocol, we use network augmentation, functional annotation, enrichment analysis, topology analysis and clustering. A brief introduction of each of these approaches follows.

Network augmentation

Network augmentation uses additional known protein-protein interactions to enhance experimental results. For low-density data sets, this method can determine whether multiple prey proteins might have been pulled down as a complex rather than as a single interaction. As previously mentioned, there are a large number of public databases of known protein-protein interactions. We recommend only using interactions that have been experimentally verified, either by direct immunoprecipitation assays or some other biochemical technique. High-throughput techniques such as yeast two-hybrid screening might have too much noise to give accurate information, but CORUM[79] is valuable for human complexes. Most of the public repositories provide information about the type of assay used to determine the interaction and allow you to select only certain types. Resources such as GeneMANIA[83] and PSICQUIC[84] also support queries across multiple repositories.

Once you have chosen a set of public repositories, you will then find the interactions between your prey proteins and add them to your network. In the PROCEDURE, we use Cytoscape's features for this process, but there are several other ways to add these interactions to your network. Network augmentation may also be useful to explore potential interactions with your prey proteins and other proteins not in your experimental set. For example, one Cytoscape app, the Agilent Literature Search App[85], searches literature abstracts to find relevant terms that suggest interactions. Another option is to add a gene of interest to your network to see whether it interacts with your prey proteins. As with other augmentation approaches, these tools are useful for forming hypotheses and should be verified using other experimental approaches.

Functional annotation

The goal in functional annotation is to note information about the known biological function of prey proteins or bait-prey interactions. Annotation is often done using GO[86], and, in Cytoscape, it involves the addition of GO terms to the node (protein) or edge (interaction) of interest. GO provides a very rich set of terms, which can cause the number of terms associated with a particular protein to become overwhelming. Thus, a reduced

(or 'slimmed') set of terms, called GO-slim (http://www.geneontology.org/GO.slims.shtml), is often more useful for annotating your network at the right level, and it is less redundant.

Enrichment analysis

Enrichment analysis determines whether a subset of your network is enriched in some associated function[87,88]. One approach to this analysis is a statistical technique that determines whether some terms are over-represented (enriched) in a subset of the data set. For example, consider an AP-MS experiment looking at the interactions between HIV and human cells. To determine whether HIV pathogen proteins target specific aspects of the cellular machinery, the network can be functionally annotated with GO, and then each of the bait protein's interaction partners can undergo enrichment analysis. If these proteins are more likely to share a set of GO terms than one might expect, on the basis of a random assignment of all terms in the network, those proteins are likely to be involved in that biological process or cellular function (depending on the branch of GO). There are several tools for performing ontology-based enrichment analysis, including web-based (e.g., DAVID[89,90]) and various Cytoscape apps (e.g., BiNGO[91], ClueGO[92], NOA[93] and ReactomeFI[94]). When using these tools, be aware of the source and version of their underlying ontology data.

An additional statistical approach to enrichment analysis is to use a 'guilt-by-association' technique[95]. In this technique, if a gene associates with a group of genes that tend to have a particular function, that gene is likely to also have that function. This technique might be useful when a protein in an AP-MS experiment has no annotated function but several interaction partners. We do not use this technique in our protocol, but several Cytoscape apps provide tools that use guilt by association to functionally annotate genes or proteins in an interaction network (GeneMANIA[83]).

Topology analysis

Another common technique for analyzing AP-MS networks is to evaluate the topology of the network. This method is seldom used for low-density networks (bait proteins will obviously be hubs), but it might be informative in higher-density networks in which the proteins represent a significant proportion of a cell's proteome, or when a particular cellular function or mechanism is being studied. Network topology is often defined by measurements that include the following:

- Node degree—the number of nodes that interact with this node,
- Degree centrality—a measure of how central a node is based on normalizing its node degree (can also be a measure of the entire network),
- Betweenness centrality—the tendency of a node to be on the shortest path between other nodes in the network,
- Closeness centrality—the tendency of a node to be close (as measured by shortest path) to other nodes in the network, and
- Eigenvector centrality—a measure of the influence of a node in a network that considers connections to higher scoring nodes more than connections to lower scoring nodes.

Many of the above measures may be summed over all nodes to get a sense of the topology of the network as a whole. In terms of looking at protein-protein interaction networks generated from AP-MS experiments, proteins that have a high node degree (hubs) or have a high betweenness centrality are candidate essential proteins. To explore network topology measures, see Pavlopoulos *et al.*[96], Grindrod and Kibble[97], Koschützki and Schreiber[98] and Vidal *et al.*[99]. Although Cytoscape 3 includes the Network Analyzer[100] functionality in the core, other apps such as CentiScaPe[101] are also available.

Clustering

Clustering (also known as unsupervised classification) is a very common technique used in high-density networks for finding close associations that might suggest complexes or partitioning the network for simpler visualization and analysis. Several algorithms have been published for finding complexes within networks, including MCL[102], MCODE[103], RNSC[104] and SPC[105] (see the comparison by Brohee and van Helden[106] or the overview by Moschopoulos *et al.*[107]). Krogan *et al.*[27] used MCL to find putative complexes in their data, but Collins *et al.*[63] used a hierarchical cluster approach to generate an adjacency matrix view of the network. The clusterMaker2 app[108] for Cytoscape supports many of these algorithms and visualizations, although there are several other apps that implement one or more clustering algorithms.

Network visualization

There are two major goals for network visualization: exploration and communication. Exploration is useful to form hypotheses and to develop an understanding of the data set, whereas communication often involves abstracting or filtering the data set to show the relevant information. Abstracting or filtering data too early in the analysis of a data set can obscure important relationships. In contrast, not filtering irrelevant data or abstracting data results can obscure the evidence for specific hypotheses. We explore both of these approaches in our visualization steps below. This protocol focuses on visualizing the network as a node-link diagram, in which nodes represent the proteins and links (edges) represent the interactions. Although there are alternative representations for interaction networks, such as adjacency matrices, they are much less commonly used.

MATERIALS

EQUIPMENT

Hardware requirements

- Personal computer with Internet access; we also recommend a screen with 1,920 × 1,080 (HD) resolution and at least 8 GB of RAM

Software requirements

- Microsoft Excel
- R (downloaded from http://cran.r-project.org/)
- MiST software, available from https://github.com/everschueren/mist
- SAInt software, available from http://saint-apms.sourceforge.net/
- Java Standard Edition, version 6 or 7 (download from http://java.oracle.com)
- Cytoscape 3.1 or later. Cytoscape may be downloaded from http://www.cytoscape.org
- Cytoscape apps. These may be downloaded from http://apps.cytoscape.org or through Cytoscape's App Manager.

• These include: clusterMaker2 (http://apps.cytoscape.org/apps/clusterMaker2); BridgeDB (http://apps.cytoscape.org/apps/BridgeDB); BiNGO (http://apps.cytoscape.org/apps/bingo); and
• enhancedGraphics (http://apps.cytoscape.org/apps/enhancedGraphics)

EQUIPMENT SETUP

Data sets Jäger *et al.*[12] is a systematic study using AP-MS of the interactions between HIV proteins and human proteins in two different human cell lines. **Supplementary Data 1** corresponds to Supplementary Data 1 in Jäger *et al.*[12]. **Supplementary Data 2** in this paper corresponds to Supplementary Data 3 in Jäger *et al.*[12].

The study by Collins *et al.*[63] is a combination of two high-quality AP-MS studies[27,64] of the yeast interactome that uses the overlap of the two data sets to compute a probabilistic score for each edge to overcome some of the noise inherent in the data. The Cytoscape session file **Supplementary Data 3** is derived from the public repository provided by the authors at http://interactome-cmp.ucsf.edu/ (see *Physical Interactions* → *PE Scores* → *Downloads*).

PROCEDURE

Part 1—scoring the data: filtering the raw input ● TIMING ~30 min

▲ CRITICAL A step-by-step tutorial following the network analysis procedure (Steps 11–33) in Cytoscape is available as **Supplementary Methods**.

1| Filter the Prey column in the 'raw' AP-MS output file (**Supplementary Data 1**) to remove nonunique and 'decoy' protein groups identified by the search algorithm.

? TROUBLESHOOTING

2| (Optional) Filter the Prey column to remove common contaminants that are described in peer-annotated resources such as CRAPome[42], the MaxQuant contaminant file[53] or a custom list of commonly identified contaminants in your laboratory.

Preparing the input matrix for scoring ● TIMING ~1–2 h

3| Organize the data set into a 2D bait-prey matrix for MiST or SAInt. MiST and SAInt input matrix formats are compatible, but the CompPASS format is not publicly available. See the **Supplementary Data 1** tabs 'HEK_MiST_input' and 'Jurkat_MiST_input' for examples. Rows 1–3 are the matrix header; populate these rows as described here:

Row 1	Experiment identifier for each unique AP-MS run (immunoprecipitation column)
Row 2	Bait identifier that groups a set of replicated purifications (Bait column)
Row 3	Baits to exclude when computing specificity feature values. If no baits are excluded, set this value to the bait name to exclude itself

4| Populate the columns of a single spreadsheet as described in the table below. For the quantification values in columns 5 to the end, any of the following MS measurement types are allowed as long as the choice is consistent across samples. We ordered them from most quantitative but more prone to noise to least quantitative but more robust: Summed MS1 intensity measurements per protein (these are the values we used in the 'Intensity' column of **Supplementary Data 1**); summed spectral counts per protein; and number of uniquely identified peptides.

Column 1	Protein identifiers as a unique list of all identified proteins in all immunoprecipitations (Prey column)
Column 2	Protein Peptide Atlas identifiers (note: for SAInt input format compatibility, this column is required; or set to 1; for MiST values can be set to zero)
Column 3	Protein sequence length (length column) or molecular weight, which will be used to scale abundance. Both are allowed
Column 4	Prey type (PreyType column): for SAInt, fill the column with C for known contaminants, R for known non-contaminants (especially hubs), and N for all other proteins. For MiST, the column can contain any value
Column 5–end	Quantification values for each protein (row) in each individual AP-MS run (column). Use a neutral value, such as 0, for prey not identified in a particular run

5| (Optional) Identify carry-over proteins from previous MS runs by inspecting quantification values that have a half-life–like pattern in a sequence of consecutive runs (columns). If a pattern diminished systematically and the potentially carried-over protein is not detected in a biological replicate, then the likelihood that carry-over caused this pattern is high. In this case, we recommend either to set the carry-over value in the matrix to your neutral value or to keep a list of carry-over bait-protein pairs to inspect after scoring.

Inspecting reproducibility (optional) ● **TIMING ~4 h**

6| Inspect the reproducibility between biological replicates and identify mislabeled experiments, if any, by computing a run-times-run correlation matrix of matched preys. Compute a pairwise Pearson correlation matrix for all experiments using the built-in cor() function in R.

7| Cluster and visualize this correlation matrix with the heatmap() function in R, which performs a hierarchical clustering of the correlation matrix. Verify that biological replicates cluster in groups, not including negative controls. Remove low-quality replicates.

Computing the score ● **TIMING 30 min–1 d**

8| Compute the bait-prey score by running the most suitable scoring algorithms (**Table 1**) for your data set. The run time for any scoring algorithm is polynomial with respect to the data set size.

9| (Optional) Evaluate the accuracy of the scoring algorithm(s) on a data set of known high-confidence interactions. If such a data set is unavailable, then generate one by experimentally testing a few (<10) high-scoring bait-prey pairs[12].

10| Sort bait-prey scores in descending order and determine a threshold. If you evaluated the accuracy of the scoring algorithm (Step 9), then choose a threshold on the basis of a false-positive rate of 5%. If not, use the threshold recommended by the developers (i.e., MiST score 0.75 (ref. 10), SAInt score 0.95 (ref. 73) and CompPASS score 95th percentile[32]).

Part 2—network analysis: data import ● **TIMING ~30 min–2 h**

▲ **CRITICAL** See **Supplementary Methods** for a more detailed, step-by-step process.

11| Import AP-MS data from the table of scored bait-prey interactions generated in Steps 1–10 (**Supplementary Data 2**). The table of scored interactions includes additional information about the bait and prey proteins and their interactions. Each row in the table represents a scored bait-prey interaction. Edge bait-prey interactions are represented in the network as an edge between the bait and prey proteins, which are represented as nodes. Interaction information (e.g., scores) is associated with the edges, and information about the bait and prey proteins is associated with the nodes. For some software (e.g., Cytoscape), multiple passes are needed for the import process. First, import the interactions (edges) with all interaction data. Do not overfilter the interaction data too early; most software packages provide filtering tools that allow you to remove edges later (**Fig. 1**). Note that for the Jäger data set (**Supplementary Data 2**) the Score Average column is imported as an edge attribute.

12| Import information about the bait proteins (e.g., protein name, identifier and any scoring you want to associate with the protein, rather than the edge). When you are working with the low-density sample data set, only import the Bait column.

13| Import information about the prey proteins (e.g., protein name, identifier and any scoring you want to associate with the protein, rather than the edge). In the example shown in **Figure 2**, the following information was imported: Gene Symbol, GeneID, Protein Name, PreyAccesion, HEKScore and JurkatScore. It is worth noting that HEKScore and JurkatScore are more properly associated with edges. In this case, they are associated with the prey proteins for visualization purposes (Step 31).

14| (Optional) *Map identifiers*. An unfortunate challenge in bioinformatics is that many identifiers (e.g., UniProt, Entrez Gene and sequence identifiers) are associated with the same biological entity. To use existing protein-protein interaction data, a consistent identifier must exist between your data set and the public interaction data. For AP-MS data, UniProt protein identifiers are a good choice. To map the identifiers, choose an imported column to map from (e.g., Entrez Gene; GeneID column in **Supplementary Data 4**), and map that column to UniProt (see **Supplementary Methods** for Cytoscape instructions).

? TROUBLESHOOTING

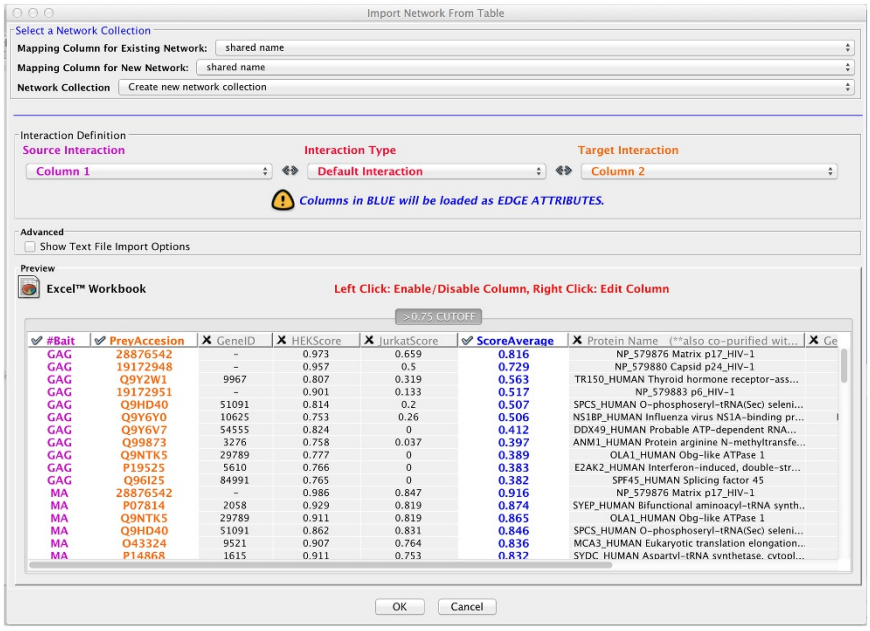


Figure 1 | Screenshot of a network import dialog showing the import of Supplementary Table 4 from Jäger *et al.*[12]. Purple and orange columns represent the source and target of an interaction, respectively, and blue columns represent data to be associated with the interaction. Columns may be indicated as interaction data by clicking on the column header.

Enriching the network with public interaction data (optional)

● **TIMING ~4 h**

15| Enrich your data with existing protein-protein interaction data (Steps 17 and 18 can be done in the opposite order, depending on the tools being used). Enriching a data set with public protein-protein interaction data can provide missing connections between proteins that were actually pulled down as a complex. First, choose a data repository from which to extract data. The repository should share identifiers with those either imported or mapped in previous steps.

? TROUBLESHOOTING

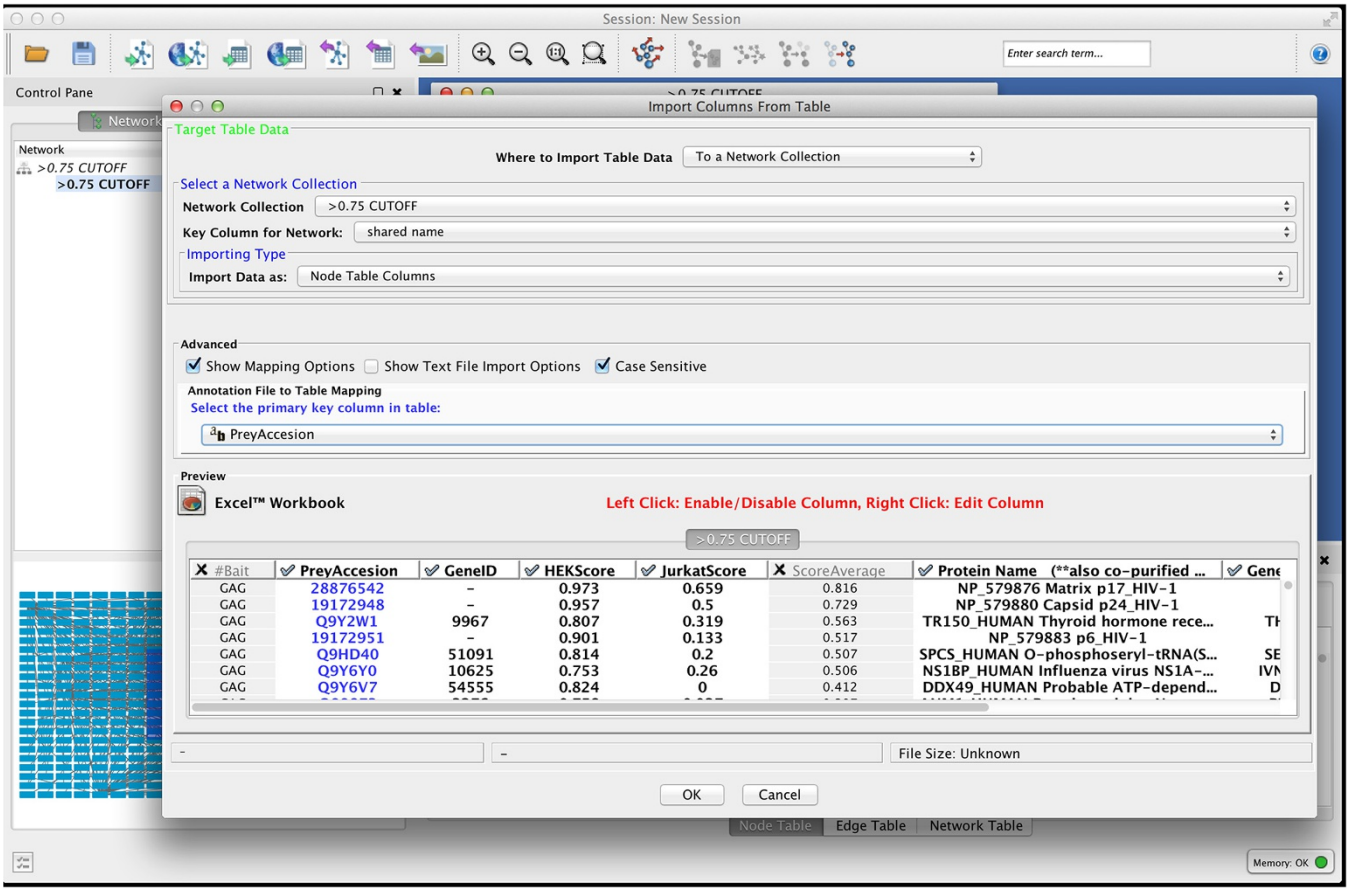


Figure 2 | Screenshot of a table import dialog showing the import of Supplementary Table 4 from Jäger *et al.*[12]. The blue column indicates the key column that is used to map the data to the nodes. Grayed columns are not imported.

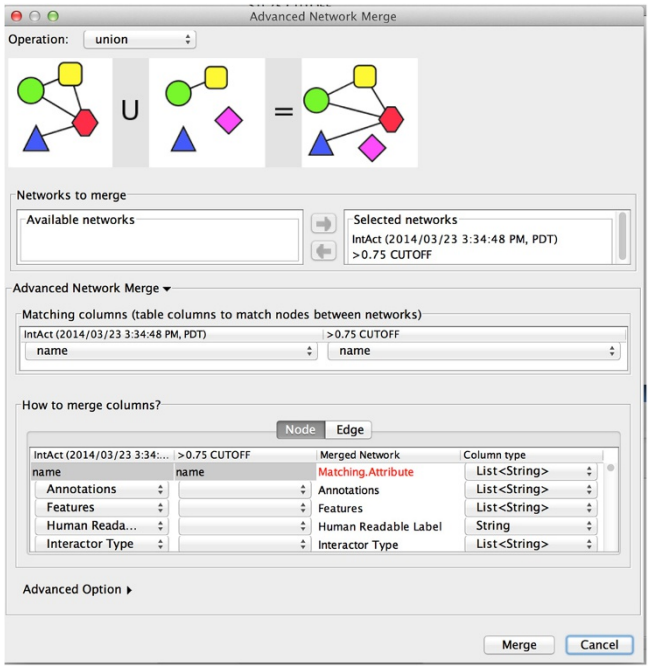


Figure 3 | Screenshot showing the merging of the IntAct imported network with our original AP-MS data set. The Advanced Network Merge tab was opened, but no changes were made.

16| By using all of the prey proteins, search the repository for interaction data. Note that public repositories search for interaction data for each protein provided, but not between all proteins. This step adds a number of proteins (and potentially small molecules) that are not part of the original data set.

17| Merge the AP-MS experimental network with the public protein-protein interaction network (**Fig. 3**).

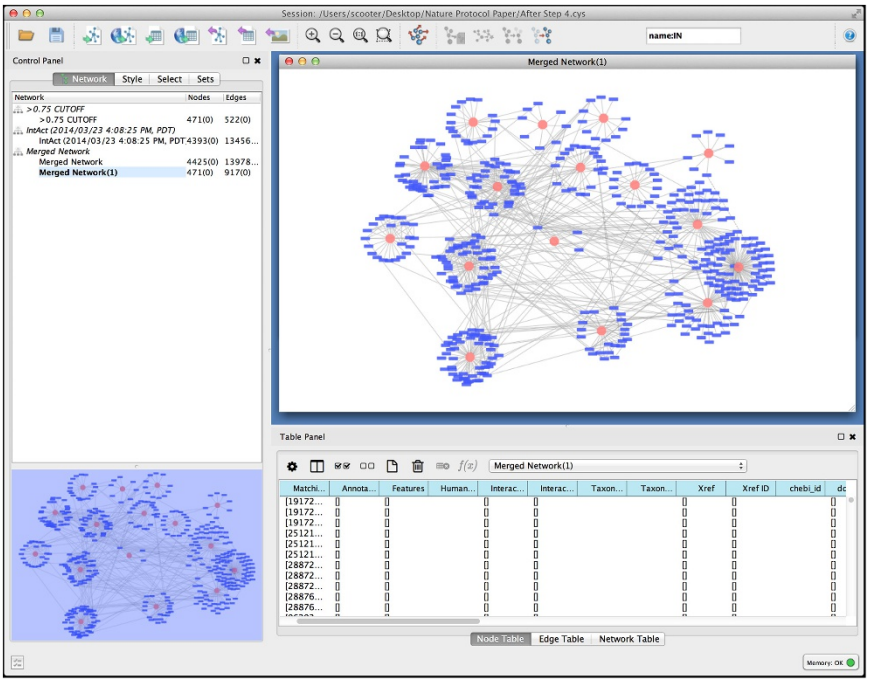
18| Filter the resulting network to only include interactions between proteins (both bait and prey) that were part of the AP-MS data set. The public network may also have pulled in a number of interactions between the AP-MS prey proteins and small molecules. Depending on the specific experiment, these interactions could be removed.

19| Visualize the resulting network and filter out any other interactions that are not relevant or are of uncertain quality. For example, some of the added interactions may be computationally derived and will not have good experimental evidence to justify them. These interactions should be removed.

20| To simplify the network, remove loops and collapse duplicate edges (**Fig. 4**).

Functional annotation ● TIMING ~2–3 h

21| Functionally annotate the proteins with GO[86] terms. We recommend including both term identifiers and descriptors. The first are easy to map to other resources and the second provide the human readable text. There are several ways to add GO terms and descriptions to the network. Most involve mapping from a protein identifier to the terms annotated to that identifier. This process may need to be done in two phases: first, map from the protein identifier to the list of GO identifiers; then, map the list of GO identifiers to a list of GO descriptions.



Network topology analysis (optional) ● TIMING ~1 h

22| For high-density data sets in which prey proteins have also been used as baits, we recommend calculating various network parameters and measuring the network topology. These measures can provide clues to the potential essentiality of proteins, critical pathways and overall network topology (e.g., scale-free). First, calculate the node degree for all nodes. High-degree nodes (hubs) are often either essential genes or promiscuous binders. Look at the

Figure 4 | The merged and filtered network. This figure was produced by selecting each bait protein, extending the selection to all first neighbors and then doing a force-directed layout that was restricted to the selected nodes. The colors and shapes were adjusted by creating a visual style in Cytoscape.

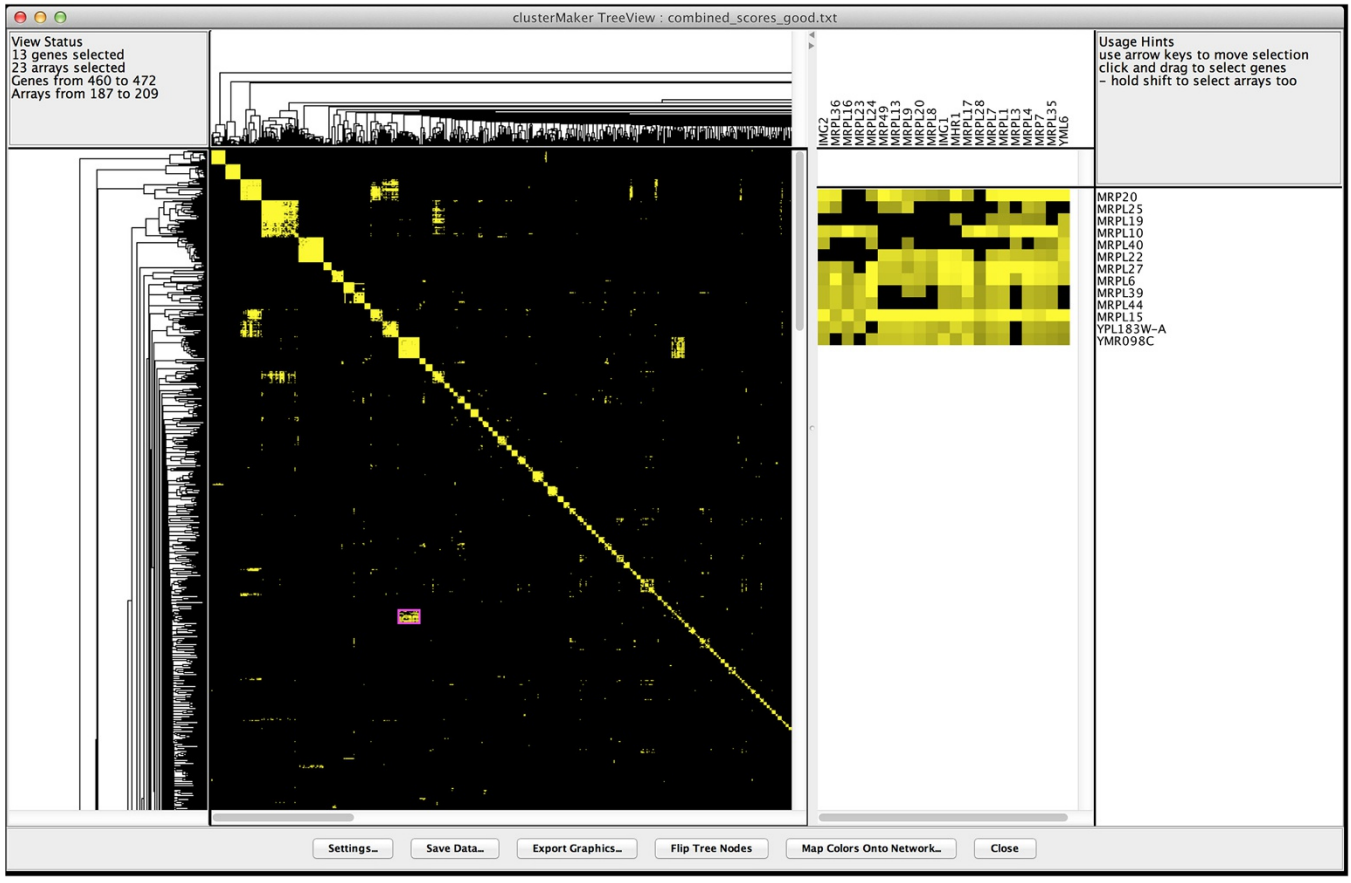


Figure 5 | Screenshot of the JTreeView viewer from clusterMaker2 after a hierarchical cluster of the PE scores from the Collins *et al.*[63] data set. The highlighted section shows the off-axis grouping between proteins of the mitochondrial ribosome. PE, purification enrichment.

high-degree proteins and check against a repository of known promiscuous binders (e.g., CRAPome[42]). Consider removing promiscuous binders from the network.

23| Calculate betweenness centrality (and any other centrality measures of interest) for the network. Betweenness centrality is another indication of the potential essentiality[109] of proteins.

Clustering analysis (optional) ● **TIMING ~2–3 h**

24| Clustering using any of the published techniques can provide useful clues to protein complexes and their interactions. Although several algorithms partition networks, two common approaches are MCL[102] and MCODE[103]. Before using a network partition algorithm, calculate a hierarchical cluster using an edge score as the metric to get an overall sense of the network clusters.

25| Visualize the resulting cluster as a heat map, and look for tight clusters on the diagonal and off-axis interactions. Tight clusters on the diagonal suggest complexes, whereas off-axis interaction might suggest co-complexes or interactions between complexes. **Figure 5** shows the heat map from the Cytoscape app clusterMaker2 (ref. 108) using the data set from Collins *et al.*[63] after performing the hierarchical cluster.

26| Partition the network using MCL[102], MCODE[103] or some other cluster algorithm that portions networks. The Cytoscape app clusterMaker2 (ref. 108) includes both of these algorithms and several others.

27| Visualize the resulting network. As with hierarchical clustering, clusters strongly suggest complexes. Adding inter-cluster edges back may infer some inter-complex interactions. **Figure 6** shows the partitioned network resulting from an MCL cluster of the Collins *et al.*[63] data set performed by the Cytoscape app clusterMaker2 (ref. 108).

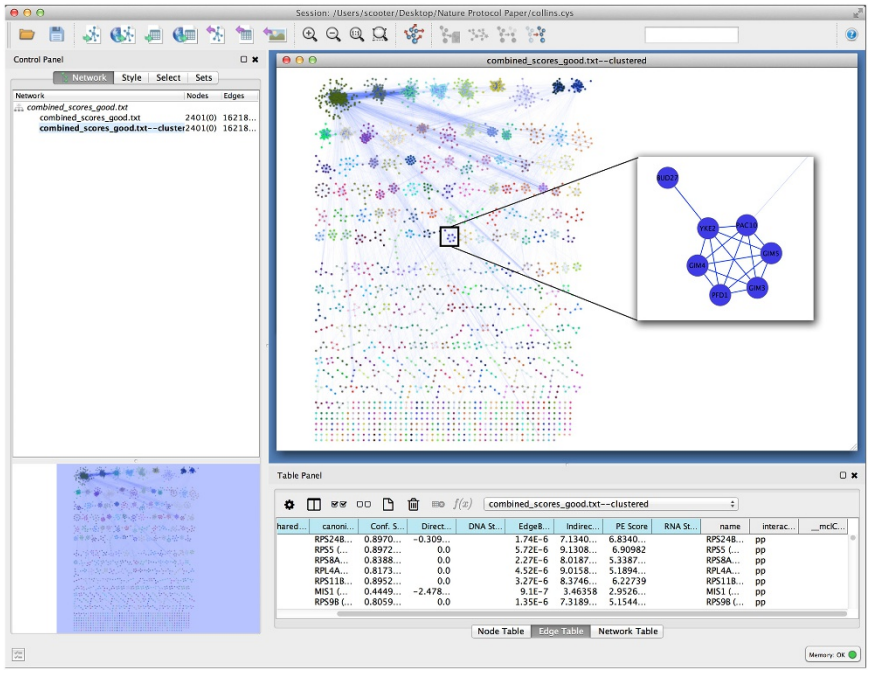


Figure 6 | Cytoscape screenshot showing the clustered network resulting from an MCL cluster on the high-density data set from Collins *et al.*[63]. The inset highlights an individual cluster.

Over-representation analysis

● TIMING ~3–4 h

28| Over-representation analysis can provide clues about the functions of protein-protein interactions, complexes or specific bait proteins. For each of the clusters calculated in Step 7 for high-density data, or bait protein and all of its interactors for low-density data, use a tool such as DAVID[90], BiNGO[91] or ClueGO[92] to determine whether the cluster or group is functionally enriched for any particular set of terms or pathways.

? TROUBLESHOOTING

Network visualization ● TIMING 2 h–1 d

29| The final visualization of the results depends on several factors: significant results, generated hypotheses and the audience. Network visualization tools provide a number of methods for adding visual styles to nodes and edges that represent associated data. These steps lead to a particular visual result, similar one shown in Jäger *et al.*[12]. See **Figure 7** for an image generated with Cytoscape and the enhancedGraphics app. First, change the color and size of the bait proteins to distinguish them from the prey.

30| Map the thickness of the AP-MS results to the Average Score so that results with a higher score show as thicker lines. If you enriched the network with public interactions (Steps 15–20), make added interactions thinner and even slightly transparent.

31| To show the differential binding for Jurkat and HEK cells, split the nodes in half and use different color gradients for the left half scaled by the JurkatScore and the right half scaled by the HEKScore.

32| Some manual positioning of nodes may be required, but a good starting point is to use a force-directed layout.

33| Save the resulting image as a PDF, SVG or some other vector format. A Cytoscape session with the results from this workflow is available in **Supplementary Data 4**.

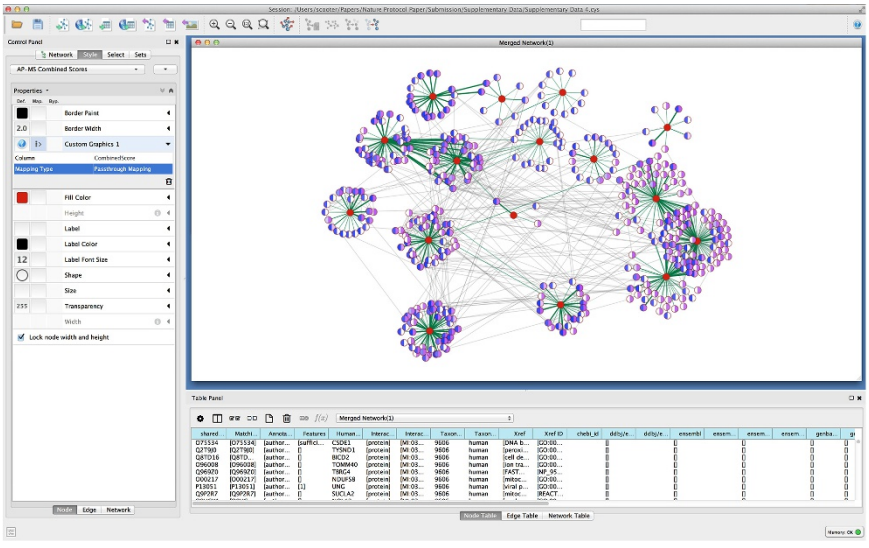


Figure 7 | Final visualized network from the Jäger *et al.*[12] data set showing the Jurkat (purple gradient on the right-hand side of the nodes) and HEK293 scores (blue gradient on the left). Interactions from the AP-MS experiment are shown as green lines in which the thickness of the line reflects the average score and interactions merged from IntAct are shown as gray lines.

? TROUBLESHOOTING

Troubleshooting tips are provided in **Table 2** for common issues with bioinformatics tools used in this protocol. More specific tips for performing network analysis steps with Cytoscape and associated apps are included in the detailed tutorial (**Supplementary Methods**).

TABLE 2 | Troubleshooting table.

Step	Problem	Possible solution
1	How to choose threshold for protein identification search	Choose a threshold that produces <1% false-discovery rate (i.e., < or = 1 decoy protein ID among a list of 100 protein hits). Orthogonal validation experiments are recommended for any proteins that score at the bottom of the list at statistical levels near the decoy matches
	No proteins are retained after filtering of contaminants	Reconsider the experimental design (i.e., selection of cell line, epitope tagging construct and purification procedure). Confirm that the bait is expressed (e.g., by immunodetection). Confirm the bait is captured and released by the purification procedure with a positive control such as GFP with the same epitope tag. Consider an alternative tagging strategy in case of interference. Consider cloning a subdomain of the protein of interest. If the protein is toxic to the cells, *in vitro* transcription and translation or Tet-inducible systems may be more suitable
14,15	No mapping IDs (or interactions) are returned	Confirm the type of source IDs. For a given type, e.g., Entrez Gene, confirm the nature of the ID, e.g., numeric identifiers or gene names
	Mapping (or interaction query) takes too long	Try a small sample query to test whether the service is working at all. If connecting via web services, check your internet connection and firewall settings. If available, consider downloading their database for local access
28	The GO terms returned are different from those that were expected or those obtained using another tool	Each ontology analysis tool should provide information on which versions of their source ontologies, e.g., GO, are used. Consider the date of the current version, the frequency of updates and any pre-processing of the ontology data

● **TIMING**

Part 1—scoring the data

Steps 1 and 2, filtering the raw input: ~30 min
Steps 3–5, preparing input matrix for scoring: ~1–2 h
Steps 6 and 7, inspecting reproducibility (optional): ~4 h
Steps 8–10, computing the score: 30 min–1 d

Part 2—network analysis

Steps 11–14, data import: ~30 min–2 h
Steps 15–20, enriching the network with public interaction data (optional): ~4 h
Step 21, functional annotation: ~2–3 h
Steps 22 and 23, network topology analysis (optional): ~1 h
Steps 24–27, clustering analysis (optional): ~2–3 h
Step 28, over-representation analysis: ~3–4 h
Steps 29–33, network visualization: 2 h–1 d

ANTICIPATED RESULTS

This protocol traces the steps from raw AP-MS data to scoring and network analysis to visualization and image export. Herein, the protocol applied to the low-density data set from Jäger *et al.*[12] provides results that support hypotheses about the function of HIV proteins VPU, VIF and VPR. Over-representation analysis and the combined scores from the scoring protocol suggest that VPU has a statistically significant interaction with proteins involved in ATP synthesis, whereas VIF and VPR have a statistically significant interaction with proteins involved in catabolic processes and RNA splicing, respectively (Step 28). In the high-density data set, the hierarchical cluster heat map suggests strong off-axis interactions between two sets of mitochondrial ribosomal proteins, and it shows the full extent of the large subunit of the mitochondrial ribosome complex in yeast (**Fig. 5**). The MCL cluster of the same data shows these proteins grouped together in a single cluster (**Fig. 6**). This protocol demonstrates some of the advantages of using multiple algorithms to explore AP-MS data in a network context (**Fig. 7**).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS The work of J.H.M. and A.R.P. is supported by grant no. P41 GM103504 (the National Resource for Network Biology (NRNB)). J.H.M. is also supported by grant no. P41 GM103311 (Resource for Biocomputing, Visualization, and Informatics (RBVI)). G.M.K. is supported by the National Institute of General Medical Sciences (NIGMS) grant no. 8P41 GM103481. E.V. and J.R.J. are supported by US National Institutes of Health grant nos. P50 GM082250, P01 AI090935, P01AI091575 and P01 AI106754. P.C. is supported by a Howard Hughes Medical Institute Predoctoral Fellowship. A.L.G. is supported by the Walter K. Evans Prememorial Fellowship.

AUTHOR CONTRIBUTIONS G.M.K., A.L.G. and J.R.J. contributed the Introduction and Experimental considerations; E.V. and P.C. contributed to Part 1 of the protocol (scoring pipeline) and its associated Supplementary Data; J.H.M. and A.R.P. contributed to Part 2 of the protocol (network analysis) including Cytoscape files and associated Supplementary Methods.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

1. Sumegi, B., Sherry, A.D., Malloy, C.R., Evans, C. & Srere, P.A. Is there tight channelling in the tricarboxylic acid cycle metabolon? *Biochem. Soc. Trans.* **19**, 1002–1005 (1991).
2. De la Fuente, I.M. *et al.* Global self-regulation of the cellular metabolic structure. *PLoS ONE* **5**, e9484 (2010).
3. Li, J. & Buchner, J. Structure, function and regulation of the hsp90 machinery. *Biomed. J.* **36**, 106–117 (2013).
4. Gao, W., Bohl, C.E. & Dalton, J.T. Chemistry and structural biology of androgen receptor. *Chem. Rev.* **105**, 3352–3370 (2005).
5. Obsil, T. & Obsilova, V. Structure/function relationships underlying regulation of FOXO transcription factors. *Oncogene* **27**, 2263–2275 (2008).
6. Rivera-Molina, F.E. & Novick, P.J. A Rab GAP cascade defines the boundary between two Rab GTPases on the secretory pathway. *Proc. Natl. Acad. Sci. USA* **106**, 14408–14413 (2009).
7. Ortiz, D., Medkova, M., Walch-Solimena, C. & Novick, P. Ypt32 recruits the Sec4p guanine nucleotide exchange factor, Sec2p, to secretory vesicles; evidence for a Rab cascade in yeast. *J. Cell Biol.* **157**, 1005–1015 (2002).
8. Chen, G.I. & Gingras, A.C. Affinity-purification mass spectrometry (AP-MS) of serine/threonine phosphatases. *Methods* **42**, 298–305 (2007).
9. Couzens, A.L. *et al.* Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. *Sci. Signal.* **6**, rs15 (2013).
10. Jäger, S. *et al.* Purification and characterization of HIV-human protein complexes. *Methods* **53**, 13–19 (2011).
11. Joshi, P. *et al.* The functional interactome landscape of the human histone deacetylase family. *Mol. Syst. Biol.* **9**, 672 (2013).
12. Jäger, S. *et al.* Global landscape of HIV-human protein complexes. *Nature* **481**, 365–370 (2012).
13. Greninger, A.L., Knudsen, G.M., Betegon, M., Burlingame, A.L. & DeRisi, J.L. ACBD3 interaction with TBC1 domain 22 protein is differentially affected by enteroviral and kobuviral 3A protein binding. *mBio* **4**, e00098–00013 (2013).
14. Dyer, M.D. *et al.* The human-bacterial pathogen protein interaction networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. *PLoS ONE* **5**, e12089 (2010).
15. Coiras, M. *et al.* Application of proteomics technology for analyzing the interactions between host cells and intracellular infectious agents. *Proteomics* **8**, 852–873 (2008).
16. Cristea, I.M. *et al.* Tracking and elucidating alphavirus-host protein interactions. *J. Biol. Chem.* **281**, 30269–30278 (2006).
17. Dyer, M.D., Murali, T.M. & Sobral, B.W. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog.* **4**, e32 (2008).
18. Filippova, M., Parkhurst, L. & Duerksen-Hughes, P.J. The human papillomavirus 16 E6 protein binds to Fas-associated death domain and protects cells from Fas-triggered apoptosis. *J. Biol. Chem.* **279**, 25729–25744 (2004).
19. Hartlova, A., Krocova, Z., Cerveny, L. & Stulik, J. A proteomic view of the host-pathogen interaction: the host perspective. *Proteomics* **11**, 3212–3220 (2011).
20. Henderson, B.R. & Percipalle, P. Interactions between HIV Rev and nuclear import and export factors: the Rev nuclear localisation signal mediates specific binding to human importin-. *J. Mol. Biol.* **274**, 693–707 (1997).
21. Breslow, D.K. *et al.* Orm family proteins mediate sphingolipid homeostasis. *Nature* **463**, 1048–1053 (2010).
22. Brandman, O. *et al.* A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* **151**, 1042–1054 (2012).
23. Jonikas, M.C. *et al.* Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science* **323**, 1693–1697 (2009).
24. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566 (2005).
25. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
26. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
27. Krogan, N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
28. Ewing, R.M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).
29. Goudreault, M. *et al.* 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* **8**, 157–171 (2009).
30. Guruharsha, K.G. *et al.* A protein complex network of *Drosophila melanogaster*. *Cell* **147**, 690–703 (2011).
31. Rubio, V. *et al.* An alternative tandem affinity purification strategy applied to *Arabidopsis* protein complex isolation. *Plant J.* **41**, 767–778 (2005).
32. Sowa, M.E., Bennett, E.J., Gygi, S.P. & Harper, J.W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389–403 (2009).
33. Zhou, Z., Licklider, L.J., Gygi, S.P. & Reed, R. Comprehensive proteomic analysis of the human spliceosome. *Nature* **419**, 182–185 (2002).
34. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
35. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
36. Moorman, N.J., Sharon-Friling, R., Shenk, T. & Cristea, I.M. A targeted spatial-temporal proteomics approach implicates multiple cellular trafficking pathways in human cytomegalovirus virion maturation. *Mol. Cell. Proteomics* **9**, 851–860 (2010).
37. Al-Hakim, A.K., Bashkurov, M., Gingras, A.C., Durocher, D. & Pelletier, L. Interaction proteomics identify NEURL4 and the HECT E3 ligase HERC2 as novel modulators of centrosome architecture. *Mol. Cell. Proteomics* **11**, M111 014233 (2012).
38. Dubois, F. *et al.* Differential 14-3-3 affinity capture reveals new downstream targets of phosphatidylinositol 3-kinase signaling. *Mol. Cell. Proteomics* **8**, 2487–2499 (2009).
39. Musunuru, K. Genome editing of human pluripotent stem cells to generate human cellular disease models. *Dis. Model Mech.* **6**, 896–904 (2013).
40. Chang, I.F. Mass spectrometry-based proteomic analysis of the epitope-tag affinity purified protein complexes in eukaryotes. *Proteomics* **6**, 6158–6166 (2006).
41. Westermarck, J., Ivaska, J. & Corthals, G.L. Identification of protein interactions involved in cellular signaling. *Mol. Cell. Proteomics* **12**, 1752–1763 (2013).
42. Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10**, 730–736 (2013).
43. Kean, M.J., Couzens, A.L. & Gingras, A.C. Mass spectrometry approaches to study mammalian kinase and phosphatase associated proteins. *Methods* **57**, 400–408 (2012).
44. Cristea, I.M., Williams, R., Chait, B.T. & Rout, M.P. Fluorescent proteins as proteomic probes. *Mol. Cell. Proteomics* **4**, 1933–1941 (2005).
45. Gerber, D., Maerkl, S.J. & Quake, S.R. An *in vitro* microfluidic approach to generating protein-interaction networks. *Nat. Methods* **6**, 71–74 (2009).
46. Greninger, A.L., Knudsen, G.M., Betegon, M., Burlingame, A.L. & Derisi, J.L. The 3A protein from multiple picornaviruses utilizes the Golgi adaptor protein ACBD3 to recruit PI4KIII. *J. Virol.* **86**, 3605–3616 (2012).
47. Granvogl, B., Ploscher, M. & Eichacker, L.A. Sample preparation by in-gel digestion for mass spectrometry–based proteomics. *Anal. Bioanal. Chem.* **389**, 991–1002 (2007).
48. Medzihradszky, K.F. In-solution digestion of proteins for mass spectrometry. *Methods Enzymol.* **405**, 50–65 (2005).
49. Medzihradszky, K.F., Leffler, H., Baldwin, M.A. & Burlingame, A.L. Protein identification by in-gel digestion, high-performance liquid chromatography, and mass spectrometry: peptide analysis by complementary ionization techniques. *J. Am. Soc. Mass Spectrom.* **12**, 215–221 (2001).

50. Kaake, R.M., Wang, X. & Huang, L. Profiling of protein interaction networks of protein complexes using affinity purification and quantitative mass spectrometry. *Mol. Cell. Proteomics* **9**, 1650–1665 (2010).
51. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
52. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
53. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
54. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).
55. Chalkley, R.J., Baker, P.R., Medzihradszky, K.F., Lynn, A.J. & Burlingame, A.L. In-depth analysis of tandem mass spectrometry data from disparate instrument types. *Mol. Cell. Proteomics* **7**, 2386–2398 (2008).
56. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
57. Nesvizhskii, A.I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73**, 2092–2123 (2010).
58. Choi, H., Fermin, D. & Nesvizhskii, A.I. Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* **7**, 2373–2385 (2008).
59. Liu, H., Sadygov, R.G. & Yates, J.R. III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
60. Gingras, A.C. & Raught, B. Beyond hairballs: the use of quantitative mass spectrometry data to understand protein-protein interactions. *FEBS Lett.* **586**, 2723–2731 (2012).
61. Iwabuchi, K., Li, B., Bartel, P. & Fields, S. Use of the two-hybrid system to identify the domain of p53 involved in oligomerization. *Oncogene* **8**, 1693–1696 (1993).
62. Sasaki, J., Ishikawa, K., Arita, M. & Taniguchi, K. ACBD3-mediated recruitment of PI4KB to picornavirus RNA replication sites. *EMBO J.* **31**, 754–766 (2012).
63. Collins, S.R. *et al.* Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6**, 439–450 (2007).
64. Gavin, A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
65. Bader, G.D. & Hogue, C.W. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.* **20**, 991–997 (2002).
66. Gursoy, A., Keskin, O. & Nussinov, R. Topological properties of protein interaction networks from a structural perspective. *Biochem. Soc. Trans.* **36**, 1398–1403 (2008).
67. Dunham, W.H., Mullin, M. & Gingras, A.C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* **12**, 1576–1590 (2012).
68. Griffin, N.M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **28**, 83–89 (2010).
69. McIlwain, S. *et al.* Estimating relative abundances of proteins from shotgun proteomics data. *BMC Bioinformatics* **13**, 308 (2012).
70. Yu, X., Ivanic, J., Wallqvist, A. & Reifman, J. A novel scoring approach for protein co-purification data reveals high interaction specificity. *PLoS Comput. Biol.* **5**, e1000515 (2009).
71. Breitkreutz, A. *et al.* A global protein kinase and phosphatase interaction network in yeast. *Science* **328**, 1043–1046 (2010).
72. Choi, H. *et al.* SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods* **8**, 70–73 (2011).
73. Choi, H. *et al.* Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT. *Curr. Protoc. Bioinform.* **39**, 8.15.1–8.15.23 (2012).
74. Michaut, M. *et al.* Protein complexes are central in the yeast genetic landscape. *PLoS Comput. Biol.* **7**, e1001092 (2011).
75. Bandyopadhyay, S. *et al.* A human MAP kinase interactome. *Nat. Methods* **7**, 801–805 (2010).
76. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* **41**, D816–823 (2013).
77. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472–477 (2014).
78. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–815 (2013).
79. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res.* **38**, D497–501 (2010).
80. Orchard, S. *et al.* The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–363 (2014).
81. Sardiu, M.E., Florens, L. & Washburn, M.P. Evaluation of clustering algorithms for protein complex and protein interaction network assembly. *J. Proteome Res.* **8**, 2944–2952 (2009).
82. Gavin, A.C., Maeda, K. & Kuhner, S. Recent advances in charting protein-protein interaction: mass spectrometry-based approaches. *Curr. Opin. Biotechnol.* **22**, 42–49 (2011).
83. Montojo, J. *et al.* GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics* **26**, 2927–2928 (2010).
84. Aranda, B. *et al.* PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* **8**, 528–529 (2011).
85. Vailaya, A. *et al.* An architecture for biological information extraction and representation. *Bioinformatics* **21**, 430–438 (2005).
86. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
87. Huang , D.W. *et al.* DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, W169–175 (2007).
88. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
89. Jiao, X. *et al.* DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* **28**, 1805–1806 (2012).
90. Huang , D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
91. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448–3449 (2005).
92. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
93. Zhang, C. *et al.* NOA: a Cytoscape plugin for network ontology analysis. *Bioinformatics* **29**, 2066–2067 (2013).
94. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.* **11**, R53 (2010).
95. Oliver, S. Guilt-by-association goes global. *Nature* **403**, 601–603 (2000).
96. Pavlopoulos, G.A. *et al.* Using graph theory to analyze biological networks. *BioData Mining* **4**, 10 (2011).
97. Grindrod, P. & Kibble, M. Review of uses of network and graph theory concepts within proteomics. *Exp. Rev. Proteomics* **1**, 229–238 (2004).
98. Koschutzki, D. & Schreiber, F. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul. Syst. Biol.* **2**, 193–201 (2008).
99. Vidal, M., Cusick, M.E. & Barabasi, A.L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
100. Doncheva, N.T., Assenov, Y., Domingues, F.S. & Albrecht, M. Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.* **7**, 670–685 (2012).
101. Scardoni, G., Petterlini, M. & Laudanna, C. Analyzing biological network parameters with CentiScaPe. *Bioinformatics* **25**, 2857–2859 (2009).
102. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
103. Bader, G.D. & Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
104. King, A.D., Przulj, N. & Jurisica, I. Protein complex prediction with RNSC. *Methods Mol. Biol.* **804**, 297–312 (2012).
105. Blatt, M., Wiseman, S. & Domany, E. Superparamagnetic clustering of data. *Phys. Rev. Lett.* **76**, 3251–3254 (1996).
106. Brohee, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**, 488 (2006).
107. Moschopoulos, C.N. *et al.* Which clustering algorithm is better for predicting protein complexes? *BMC Res. Notes* **4**, 549 (2011).
108. Morris, J.H. *et al.* clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* **12**, 436 (2011).
109. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* **3**, e59 (2007).