

UC Irvine

UC Irvine Previously Published Works

Title

Analysis of affinity purification-related proteomic data for studying protein–protein interaction networks in cells

Permalink

<https://escholarship.org/uc/item/1pr271bn>

Journal

Briefings in Bioinformatics, 24(2)

ISSN

1467-5463

Authors

Kattan, Rebecca Elizabeth

Ayesh, Deena

Wang, Wenqi

Publication Date


2023-03-19

DOI

10.1093/bib/bbad010

Peer reviewed

Analysis of affinity purification-related proteomic data for studying protein–protein interaction networks in cells

Rebecca Elizabeth Kattan, Deena Ayesh and Wenqi Wang 

Corresponding author: Wenqi Wang, Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA 92697, USA.
E-mail: wenqiw6@uci.edu

Abstract

During intracellular signal transduction, protein–protein interactions (PPIs) facilitate protein complex assembly to regulate protein localization and function, which are critical for numerous cellular events. Over the years, multiple techniques have been developed to characterize PPIs to elucidate roles and regulatory mechanisms of proteins. Among them, the mass spectrometry (MS)-based interactome analysis has been increasing in popularity due to its unbiased and informative manner towards understanding PPI networks. However, with MS instrumentation advancing and yielding more data than ever, the analysis of a large amount of PPI-associated proteomic data to reveal *bona fide* interacting proteins become challenging. Here, we review the methods and bioinformatic resources that are commonly used in analyzing large interactome-related proteomic data and propose a simple guideline for identifying novel interacting proteins for biological research.

Keywords: protein–protein interaction, proteomics, affinity purification, bioinformatics

Introduction

Protein–protein interactions (PPIs) mediate cellular signal transduction cascades where specific interactions between proteins define unique and robust signalling outputs for different cellular events [1]. Dysregulation of PPIs has been frequently associated with various human diseases [2]. Therefore, better characterizing PPIs will create new opportunities for understanding normal physiology and treating human diseases.

Mass spectrometry (MS) is an instrument for analyzing small chemical and biological molecules [3], which has revolutionized scientific research in the past decades. Regarding PPI study in biological research, MS can be used in partner with multiple biochemical approaches to uncover *bona fide* interactors for proteins of interest. Commonly used methods for isolating protein complexes comprise immunoprecipitation, affinity purification (AP) [4] and protein-proximity labelling approaches like BioID [5] and APEX [6]. However, with MS technology and methods of capturing interactors continuously improving, the size and complexity of PPI-related proteomic data keep increasing. Manually converting such a large dataset to a meaningful biological list of PPIs is time- and labour-consuming, and often results in errors and biases.

To tackle this issue, several bioinformatic tools have been developed to facilitate the PPI-related proteomic data analysis, which include but are not limited to MS raw data filtration, gene ontology (GO) study, functional topology analysis and PPI network visualization. These methods make processing, analyzing and presenting large-scale MS data possible for regular biological research labs. To produce a reliable interactome dataset that can be used by other researchers, the refined PPI network also needs to

be experimentally validated and functionally characterized. However, integrating these tools and methods to produce meaningful and reliable interactome data for in-depth functional studies has been challenging.

Here, we review the current methods used for isolating and identifying interacting proteins by MS analysis and provide a brief summary on how to process MS raw data, interpret their biological significance, build PPI network and characterize biological functions for newly identified interacting proteins. We hope this guideline and the related bioinformatics resources can benefit new researchers who are interested in taking MS as an approach to investigate PPIs for their biological research.

Main

Isolation of interacting proteins for a protein of interest

Many biochemical methods have been developed to isolate the associated protein complex for a protein of interest (i.e. bait protein) [7], such as immunoprecipitation (IP) through antibodies against either bait protein or an epitope tag fused with bait protein, AP using single or tandem epitope tag [4] and protein-proximity labelling approaches using the modified biotin-protein ligase tag (e.g. BioID [5], APEX [6]). Despite their different mechanisms, the overall goal of these purification techniques is to capture and reserve true binding proteins while minimizing the non-specific ones for bait protein.

Attempting to capture all the interacting proteins whether they are transient or stable has been a difficult task for the current purification methods. Although performing IP with a bait protein

Rebecca Elizabeth Kattan is a graduate student at University of California, Irvine. Her research interests include proteomics and cancer signaling pathways.

Deena Ayesh is an undergraduate student at University of California, Irvine. Her research interests include proteomics and cancer signaling pathways.

Wenqi Wang is an Associate Professor at University of California, Irvine. His research interests include elucidation of the protein-protein interaction networks underlying organ size control and the role of their dysregulation in cancer development.

Received: November 7, 2022. **Revised:** December 22, 2022. **Accepted:** January 2, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

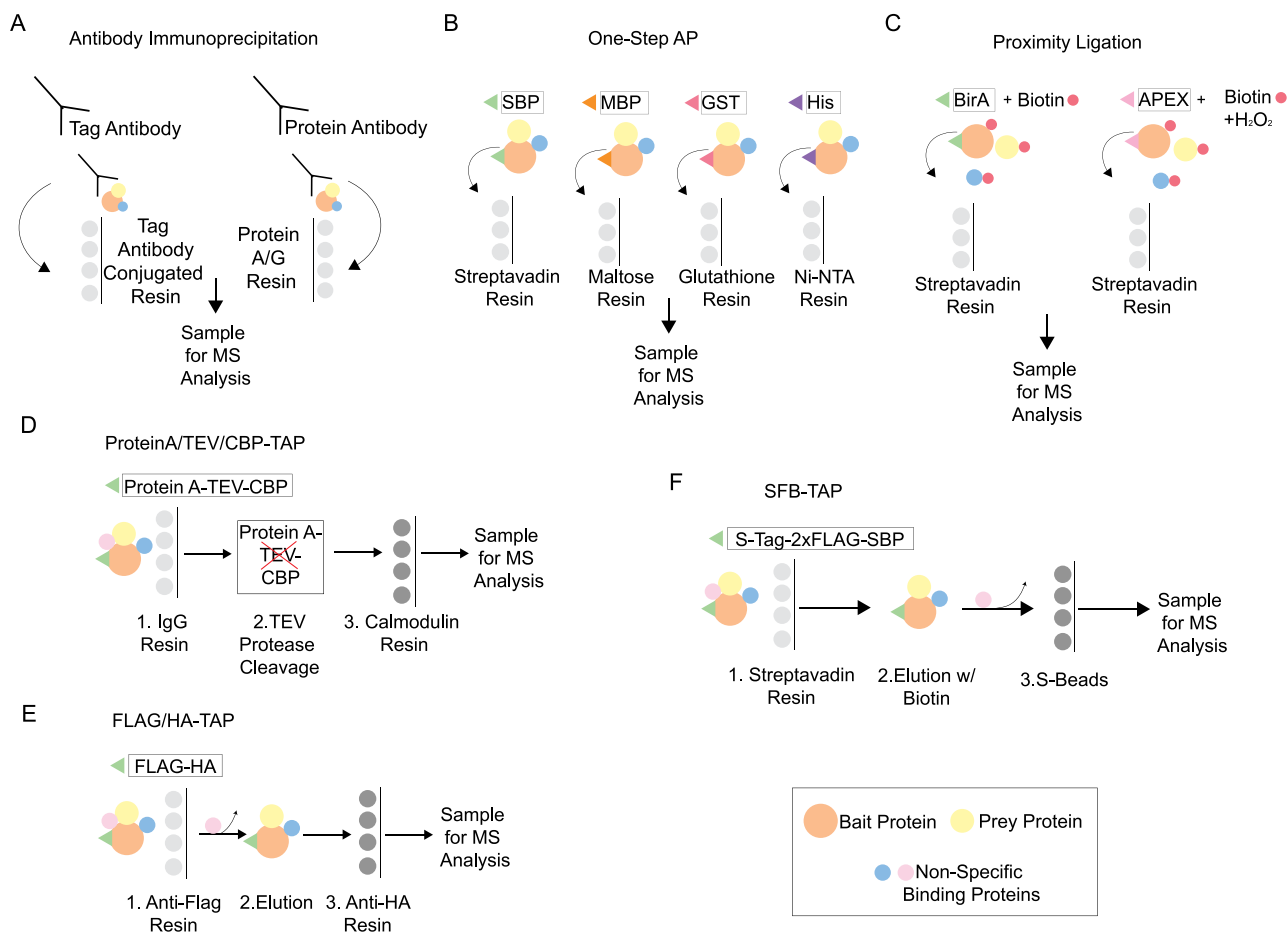


Figure 1. Illustration of commonly used methods for isolating associated protein complex for a protein of interest. The commonly used methods for purifying bait protein-associated protein complex from cells for MS analysis include antibody-based immunoprecipitation (A), one-step AP (B), proximity ligation-based AP (C), ProteinA/TEV/CBP-based TAP (D), FLAG/HA-based TAP (E) and SFB-based TAP (F).

antibody (Figure 1A) can purify its associated protein complex at endogenous level, antibody availability and its IP efficiency often limit the use of this method. One way to overcome these issues is to utilize well-established antibodies for epitope tags, such as Flag, Myc, HA, GFP and indirectly isolate the binding proteins for the bait protein fused with an epitope tag (Figure 1A). However, exogenously expressing tag-fused protein often results in non-specific and artificial binding proteins due to overexpression. In addition, both methods share the common antibody leakage issue, which can affect sample preparation for MS analysis.

AP can overcome the antibody leakage issue and has been widely used for PPI studies. To achieve so, epitope tags with high binding affinity to corresponding agarose beads have been developed (Figure 1B). For example, a commonly used tag for AP is streptavidin binding protein (SBP) (Figure 1B), which has high binding affinity to streptavidin beads [8]. However, streptavidin beads also bind endogenous biotinylated proteins in cells, resulting in non-specific interactions [8]. Maltose-binding protein (MBP) [9] and glutathione S-transferase (GST) [10] tags are often used for protein complex purification, because they can strongly bind to the agarose beads conjugated with amylose and glutathione, respectively (Figure 1B). However, these two epitope tags are both large in protein size and can cause potential folding issues for a bait protein, thus hindering the isolation of its true binding proteins. In contrast, hexa-histidine (His) tag [11] is a small tag that binds to immobilized nickel beads (Figure 1B), which offers

a better solution for AP as compared with MBP and GST tags. However, His tag-mediated protein purification requires the step of optimizing imidazole concentration for different bait proteins, adding additional work to ensure the protein purification quality.

Recently, proximity labelling methods have been widely used for interactome study (Figure 1C) because it allows identification of transient and weak binding proteins for a bait protein. Like the previously mentioned SBP tag, this method relies on streptavidin beads to purify the newly biotinylated proteins, though the endogenous modification of biotinylation leads to the issue of isolating non-specific binding proteins.

Over time, methods for purifying protein complex associated with a protein of interest have been evolving to increase their performance and efficiency [12]. Some labs incorporated different epitope tags as one for protein complex purification, allowing further elimination of non-specific binding interactions through multiple purification and washing steps [13]. For example, tandem affinity purification (TAP) was developed to fuse two epitope tags with one bait protein to reduce the non-specific binding proteins [4, 13]. Initially, TAP was designed with two IgG-binding units of protein A of *Staphylococcus aureus* (Protein A) and the calmodulin-binding peptide (CBP) [12] (Figure 1D), whereas its upgraded version includes tags like FLAG and HA in tandem [14] (Figure 1E) to achieve a better purification performance.

In the past years, we have been extensively utilizing an S-Flag-SBP (SFB) triple tagged system for TAP (Figure 1F) and revealed

new regulators/effectors for multiple key proteins involved in growth control and cancer development [15–28]. As for it, SBP is a small peptide that binds effectively to streptavidin beads, which can be easily eluted with a biotin-containing solution [8]; S protein is another small peptide that binds efficiently with S protein beads and can be also used for AP [29]. The Flag tag in this system is used to detect bait protein expression by Western blot and verify its localization by immunofluorescence. As a routine practice in our lab, cells stably expressing the SFB-tagged bait protein are generated by either lentiviral infection or single colony isolation. After verifying the bait protein expression and localization using anti-Flag antibody, these stable cells will be expanded to a large scale and lysed for TAP. As shown in Figure 1F, the first step of this purification method involves the use of streptavidin beads to isolate the associated protein complex by binding to SBP tag. After biotin elution, S protein beads are used for the second round of purification. The whole process can be finished in several hours and only one buffer solution (i.e. NETN buffer) is used from the beginning to end, which greatly reduces the chance of protein degradation and protein complex dissociation.

In addition to purifying the associated protein complex from cell lysates, several library screening approaches are available for identifying interacting proteins for a protein of interest, such as yeast two-hybrid (Y2H) [30], bimolecular fluorescence complementation (BIFC) [31], which are developed based on the protein-fragment complementation strategy [32]. For Y2H screening, a transcription factor is split into the DNA-binding domain (DBD) and activating domain (AD). If AD and DBD are brought into proximity, they can activate downstream reporter gene by binding onto its upstream activating sequence (UAS). In general, a protein of interest is fused with DBD, while its candidate interacting proteins are fused with AD and prepared as a cDNA library for screening their interactions with the bait protein. PPI indirectly connects DBD and AD to activate the transcription of reporter gene through its UAS. For BIFC, a fluorescent protein (e.g. CFP, GFP, YFP) is split into two fragments, which are, respectively, fused with a bait protein and its candidate interacting proteins in a format of cDNA library. The binding between bait protein and its interacting proteins will bring the two fragments of a fluorescent protein within proximity, allowing the fluorescent protein to reform and emit its fluorescent signal. Detection and quantification of such fluorescent signal can be achieved by fluorescent microscope and flow cytometry. Notably, these two assays are both performed in a live cell system, making the characterization of *in vivo* PPIs possible. However, as these complementation assays require generation of two separate fusion proteins, technical issues need to be taken into consideration, which include but are not limited to the effects on protein localization/function as caused by fragment fusion, overexpression-induced artificial effects and issues with the reforming efficiency for the split fragments from two fusion proteins. Previous studies also raise concerns regarding high false positive and false negative rates for Y2H and BIFC assays. Technically, generating the large-scale cDNA library and setting up conditions for screening is costly and labour / time-consuming, making these assays difficult to be widely used by researchers.

MS data generation, submission, processing and quality evaluation

Upon the completion of protein complex purification, a sample can be prepared in a format either on beads or in a polyacrylamide gel and processed by an MS facility. After trypsin digestion, produced peptides are eluted through high-performance liquid chromatography (HPLC), subjected to electrospray ionization, and

loaded into a mass spectrometer, where peptides are detected, isolated and fragmented to produce a tandem mass spectrum of specific fragment ions for each peptide. Peptide sequences (i.e. protein identity) are determined by matching protein databases (e.g. UniProt) with the fragmentation pattern acquired by the software program SEQUEST. Spectral matches are filtered to contain a false discovery rate (FDR) of less than 1% at the peptide level using the target-decoy method [33]. The protein inference is considered followed the general rules [34] with manual annotation based on experiences applied when necessary.

Users will then be provided with an extensive list of identified proteins from the sample, which are referred as MS raw data. Now it has become a standard practice by scientific journals that these raw data should be shared through public repositories before publication [35]. One commonly used repository is the Proteomics IDentifications database (PRIDE) [36, 37], which is designed to receive raw protein and peptide files for a MS experiment. To deposit MS data to the PRIDE, users first need to gather the MS data files from the MS facility including raw, result, search and peak files. These files are then uploaded through the ProteomeX-change (PX) submission tool [38], where a two-step assessment process via PRIDE is provided for checking data quality [39]. After submission, a PX accession number and permanent digital object identifier (DOI) will be issued for publication use.

Next, we usually use a pipeline to deconvolute the MS raw data into a short list of high confident interacting proteins (HCIPs) for a bait protein (Figure 2A). To achieve so, a web-accessible resource named the contaminant repository for AP (CRAPome) [40] is often used to filter out commonly identified prey proteins (i.e. non-specific binding proteins) by comparing to control experiments provided by either CRAPome or users. Control experiments are a group of unrelated MS raw datasets that are usually produced under similar experimental settings [40]. Based on the quantitative comparisons of prey abundance (using spectral counts) against the prey abundances across control experiments, a significance analysis of interactome (SAINT) score [41] will be assigned to each prey, allowing users to generate a list of HCIPs based on a suitable cutoff value of SAINT score.

Specifically, SAINT identifies false interactions by estimating the spectral count distribution from negative controls. For experiments that are produced with multiple replicates, a probability score is assigned to estimate the FDR [41], allowing users to determine the reliability of interactions. Recently, SAINT has been updated to SAINTExpress [42], which provides a topology-assisted probability score (TopoAvgP), incorporating the prior knowledge of the target interactome into the scoring step. In addition, SAINTExpress provides a simpler fold-change (FC) score based on the ratio of averaged normalized spectral counts between experiments and controls.

In addition to CRAPome, scoring algorithms like comparative proteomic analysis software suite (CompPASS) [14], mass spectrometry interaction statistics (MiST) [43] and Minkowski distance-based unified scoring environment (MUSE) [18] are also available for generating HCIP list. As compared with CompPASS and MiST, SAINT can be applied to datasets of all sizes and perform filtering quantification simply using spectral counts rather than other MS parameters [41]. In addition, SAINT removes interactions with spectral counts less than two, making the filtering process more robust [41]. Different from SAINT, MiST provides a more complete dataset analysis by incorporating multiple MS parameters-based measures, such as protein abundance (i.e. peak intensities), invariability of abundance over replicated experiments (i.e. reproducibility), uniqueness of an observed interaction

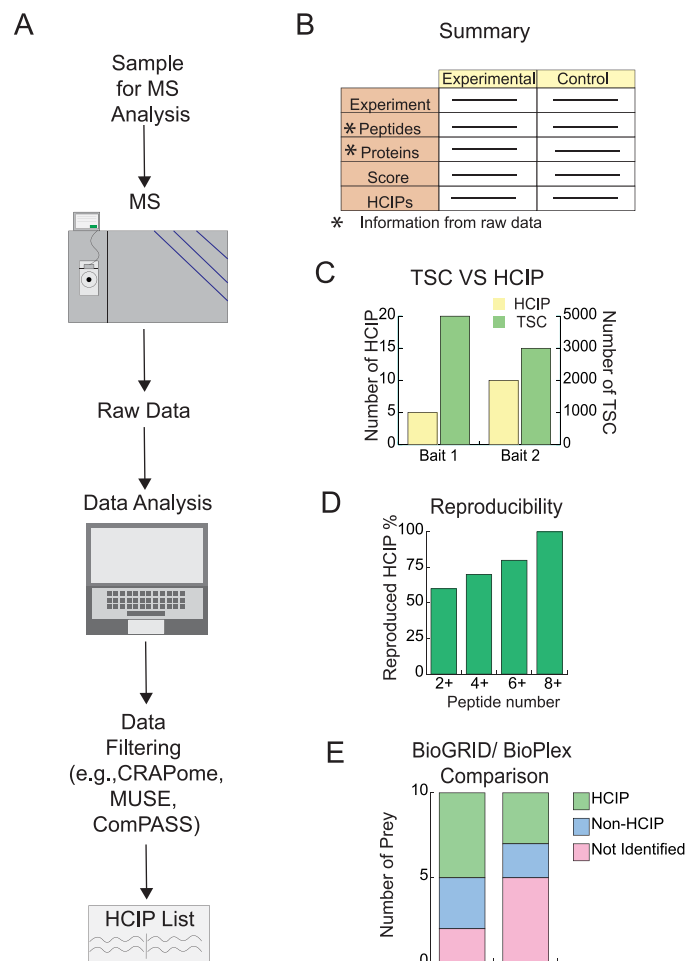


Figure 2. Schematic overview of the interactome-related proteomics data processing. **(A)** Schematic workflow for the filtration of MS data to generate a high confidence interacting proteins (HCIP) list used for further bioinformatics analysis. **(B)** Organization of MS summary data can include the number of experiments, peptide information, protein information, cutoff score for HCIP generation and HCIP count. **(C)** Schematic illustration of the comparison between the TSCs and HCIP number for an interactome study. **(D)** Reproducibility rate can be illustrated based on different prey peptide cutoffs. **(E)** HCIPs can be compared with public PPI resources like BioGRID and BioPlex to reveal known and new interactions in the produced interactome dataset.

across all purifications (i.e. specificity) [43]. It is plausible to use different methods to analyze the same MS raw data, although extra bioinformatic work will be required.

For publication purpose, we usually provide an MS data filtration report as a table by detailing the numbers of total experiments, control experiments, and identified peptides and proteins, respectively (Figure 2B). In addition, the cutoff value (e.g. a SAINT score) chosen for HCIP generation and the number of total HCIPs generated in this study are often included in this report table (Figure 2B).

To evaluate the quality of the produced HCIP dataset, additional analyses are usually performed. For example, the total spectral count (TSC) can be presented along with the number of HCIPs (Figure 2C), allowing readers to evaluate the HCIP rate of each sample in the dataset. Biological replicates are required to be performed for current proteomic studies; therefore, reproducibility is another key factor to assess the variation among experiments for the bait proteins [44]. This can be determined by calculating the HCIP correlation R value under different numbered peptides in the dataset (usually from low to high), where the reproducibility rate at each peptide cutoff number can be shown (Figure 2D). Another way to qualitatively evaluate the produced HCIP dataset is to compare them with some available

PPI databases, such as biological general repository for interaction data sets (BioGRID) [45], biophysical interactions of ORFeome-based complexes (BioPlex) [46], search tool for the retrieval of interacting genes/proteins (STRING) [47]. These databases comprise numerous reported protein interactions that have been discovered through various experimental assays including AP-MS, proximity label-MS, Y2H, immunoprecipitation, biochemistry, immunofluorescence. Comparing HCIPs with these reported interactions not only helps to evaluate the quality of the HCIP dataset from a different perspective, but also allows to reveal new PPIs through the current interactome study for functional investigation in future.

Annotation of HCIP dataset

Once a list of HCIPs have been generated, a standard practice is to deconvolute their underlying biological connections. This step is crucial for revealing potential functional processes, signalling pathways, cellular components and human diseases that bait proteins may be involved through their HCIPs. To achieve so, several web-based annotation tools such as the database for annotation, visualization and integrated discovery (DAVID) [48], protein analysis through evolutionary relationships (PANTHER) [49] and Metascape [50] are available for GO analysis of HCIPs.

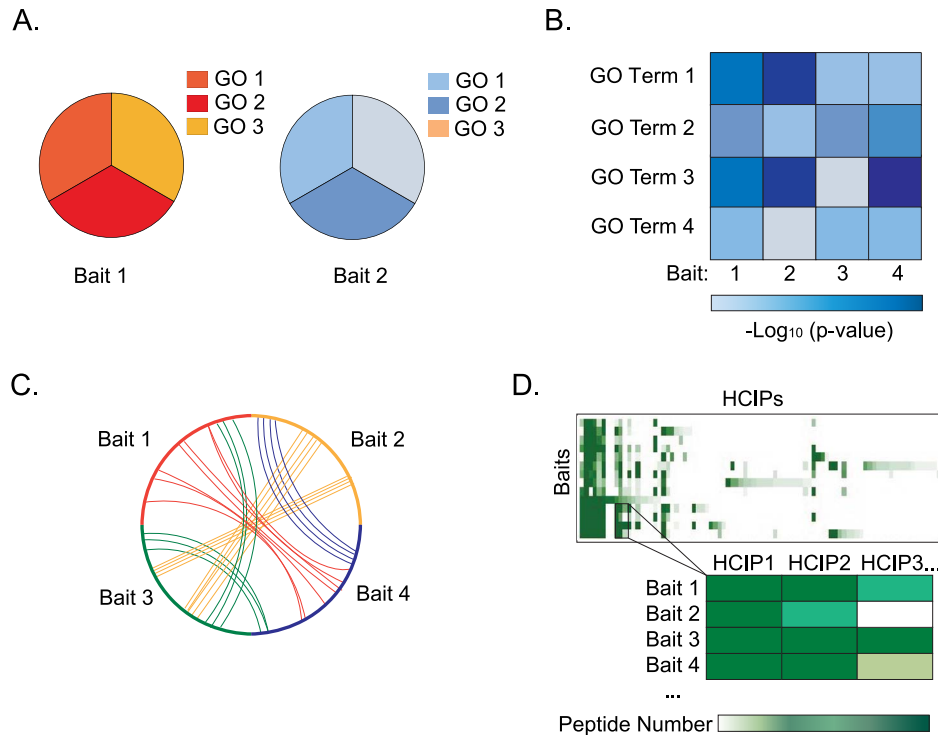


Figure 3. GO and topological analyses of interactome data. **(A)** GO analysis can be applied and presented for bait proteins individually. **(B)** GO analysis can be applied across a series of bait proteins and presented as a heatmap. **(C)** Topological analysis of multi-bait interactome data can be presented as Circos plot to reveal the bait proteins that share overlapping HCIPs. **(D)** Topological analysis of multi-bait interactome data can be presented as heatmap to reveal the overlapping HCIPs among bait proteins.

The common ontologies described in these tools include biological process, molecular function, cellular components, pathways, diseases and tissue distributions. As for a multiple-bait interactome study, bait proteins can be either analyzed individually for its HCIPs-associated GO terms (Figure 3A) or compared globally and visualized as a heatmap (Figure 3B). The latter usually helps reveal the bait proteins who share overlapped cellular functions through their HCIPs.

To reveal the shared HCIPs across different bait proteins, clustering methods through Circos plot (Figure 3C) or heatmap (Figure 3D) are often used. On a global scale, Circos plot can provide easy visualization of relationships between different bait proteins [51], whereas heatmap can show details of the overlapped HCIPs among different bait proteins to reveal potential PPI sub-networks within a multi-bait interactome study.

Notably, innovative technologies like artificial intelligence (AI) machine learning greatly facilitate the predictions of protein structures and PPIs, providing additional tools for HCIP dataset annotation. For example, AlphaFold [52, 53] and RoseTTAFold [54] are both open-sourced AI tools, which can provide structural information for the identified HCIPs and predict their protein complex formations with the bait protein. In addition, TissueNet [55] and integrated interactions database (IID) [56] tools can offer tissue expression information for HCIPs, while weighted gene co-expression network analysis (WGCNA) [57] can be incorporated to annotate functional correlations between bait protein and its HCIPs.

Visualization of PPI network

Once HCIPs have been generated and annotated, a PPI network can be built up to provide an informatic overview of bait

protein-associated interactome. Such PPI network can be generated using Cytoscape [58, 59] through various freely available plugins [60] or R program, a programming language for statistical computing and graphic [61].

Visualizing PPI network can easily present PPIs identified from different experiments and is useful when looking for unique connections and patterns among bait and prey proteins [62]. There are different ways visualizing a PPI network along with necessary experimental and/or biological information. Specifically, PPI network can be organized in a format of prey nodes surrounding their baits, where each node simply represents prey alone (Figure 4A). Moreover, these prey nodes can be complemented with further information by adding various 'visual features', such as different shapes, sizes, colours, patterns, outline thickness, to convey their experimental and/or biological details. For example, the size of the node can convey its identified TSC, where larger nodes represent preys with more TSC (Figure 4B). Nodes also can be made in different colours to represent the prey-associated GO terms, such as biological processes, localization, cellular components (Figure 4C). Another commonly presented information is to incorporate data gathered from BioGRID or STRING database to indicate the known interacting proteins within the PPI network. In addition, if reciprocal MS studies are performed using the identified HCIPs as bait proteins, the PPI network can be further enhanced through their connecting lines using either uni- or bi- directional arrows to indicate the relationship between bait proteins and its HCIPs (Figure 4D). Simultaneously visualizing all these representations (e.g. shape, size, colour, line direction) will make a PPI network summarizing all the experimental and biological information comprehensively achieved.

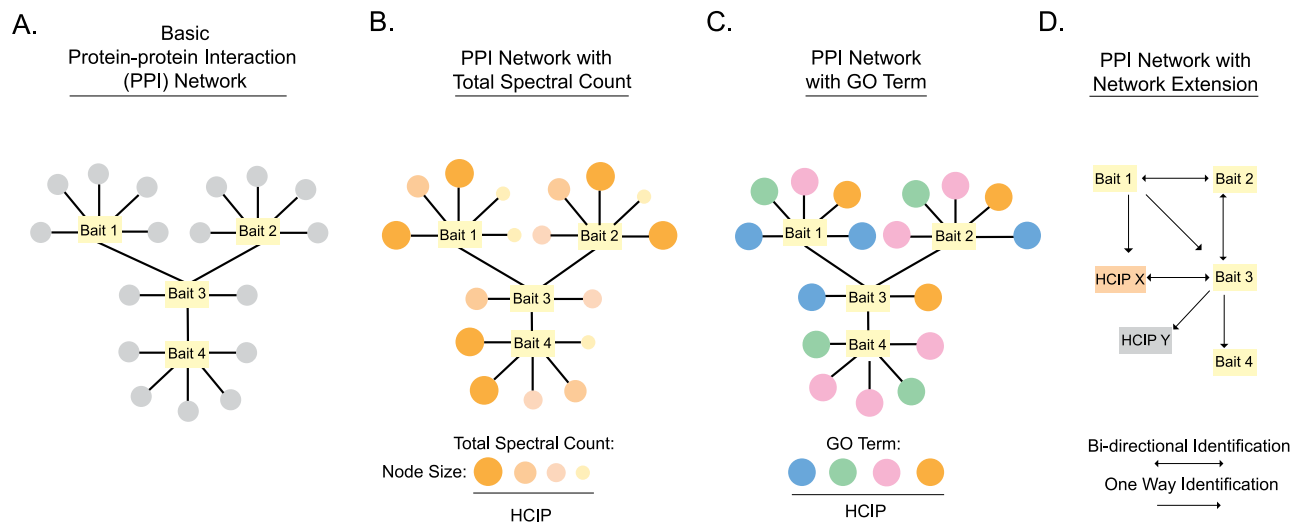


Figure 4. Visualization of PPI networks. (A) A PPI network can be organized in a format of prey nodes surrounding their baits, where nodes simply represent preys alone. (B) A PPI network can be organized in a format of prey nodes surrounding their baits, where nodes represent preys and their TSC. The size of the node conveys its identified TSC, where larger nodes represent preys with more TSC. (C) A PPI network can be organized in a format of prey nodes surrounding their baits, where nodes represent preys and their GO terms that are labelled in different colours. (D) The information from reciprocal MS studies using HCIPs as bait proteins can be included in the PPI network and indicated as connecting lines with either uni- or bi- directional arrows. Many figure labelings are missing.

Designing functional assays for characterizing newly identified HCIPs

After generating a list of HCIPs and annotating their biological functions, a group of HCIPs of interest first need to be experimentally validated. Typically, immunoprecipitation and pull-down are the choice assays for assessing the interaction between bait protein and its HCIPs. As mentioned earlier, antibody-based immunoprecipitation offers a way to examine PPI at endogenous level (Figure 5A); however, difficulties may arise in identifying a suitable antibody for use. To combat this, protein can be fused with a tag (e.g. Flag, HA, Myc, GFP, SFB), whose antibody or antibody-conjugated beads are available to help validate the PPI at a level of overexpression (Figure 5B). In addition, immunofluorescent staining is usually used to assess the co-localization between a bait protein and its HCIPs, confirming their complex formation from a different perspective (Figure 5C).

With the interaction between bait protein and its HCIPs validated, functional significance underlying their complex formation can be further explored. For example, the binding regions between bait protein and its HCIPs can be mapped in detail through generating a series of truncation and/or deletion mutants for both bait protein and its HCIPs (Figure 5D). In addition, we usually knockout (KO) or knockdown (KD) HCIPs in functionally relevant cells to examine the potential effects on the bait protein-dependent signalling events or biological functions (Figure 5E). If confirmed, these KO or KD cells will be reconstituted with wild-type HCIP or its mutant protein that fails to bind bait protein (Figure 5D) and used to determine whether their complex formation is required for the related cellular functions. For example, to examine the roles of one HCIP of interest in regulating the bait protein-dependent cell proliferation and migration (Figure 5F), altered proliferation and migration will be examined in the HCIP KO cells. If there is a change in cell proliferation/migration, rescue experiments will be performed by reconstituting the HCIP KO cells with wild-type HCIP and its bait protein non-binding mutants (Figure 5F). Using this strategy, we can provide both functional and mechanistic insights into the newly identified HCIPs through the interactome study.

Conclusions and perspectives

In this study, we reviewed the commonly used methods and bioinformatic resources for characterizing the interacting proteins for a protein of interest and illustrated a pipeline for analyzing the related MS data. This proposed pipeline is mainly composed of three steps. First, it assigns each identified prey with a confidence score, allowing users to generate a list of HCIPs for a bait protein. Second, it provides a series of bioinformatic resources (Table 1) for users to annotate HCIPs, build up PPI network and visualize interactome data informatically. Third, it suggests the strategies for follow-up data validation and functional investigation for newly identified HCIPs. In the past years, we have been frequently using this pipeline to define and characterize the PPI networks for different signalling pathways and protein families [17, 18, 23–26, 28], fully testifying its feasibility for addressing biological questions in different fields. Here, we would like to pinpoint several key factors that may affect the outcome of the interactome analysis for researchers who may be interested in trying this method for their own studies.

First, using different cell lines may lead to the difference in the identified HCIPs due to protein abundance variation between cells; therefore, the cell line choice should be considered prior to starting. In general, cell lines should be chosen based on scientific questions and their related biological contexts, whereas other issues, such as cell proliferation rate, cell culture costs, the way for cell collection, are also taken into consideration. We usually use HEK293T cells for protein complex purification due to their ease of growth and collection in a large quantity, but later move to functional cell lines to study biological functions for HCIPs.

Second, technical caveats for isolating associated protein complex for a bait protein should be taken into consideration. For example, choosing appropriate tag is crucial, as it could cause protein structure change and introduce false positive/negative hits. In addition, users should be aware of the positioning of the tag (e.g. N terminus, C terminus), as it may alter bait protein cellular localization and function. Regarding this point, several methods are available to characterize PPIs without using an epitope tag. For example, thermal proximity coaggregation approach can be

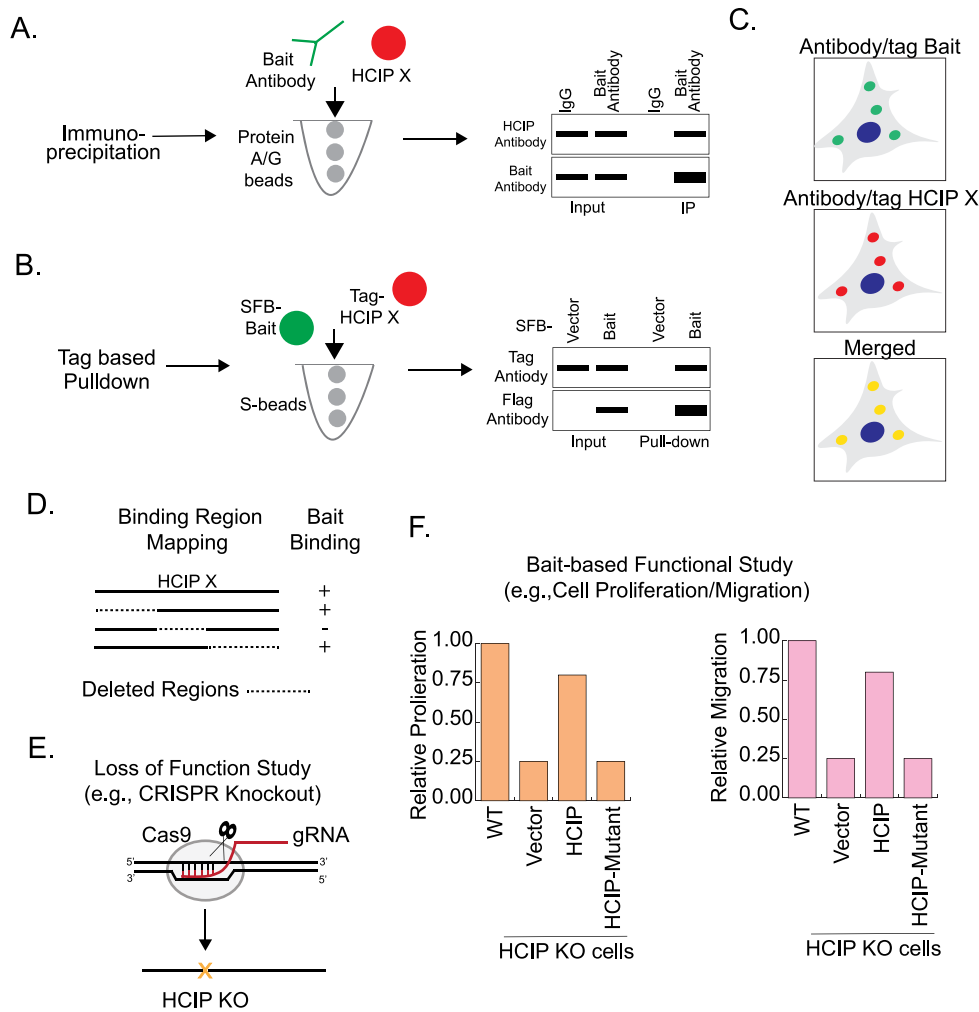


Figure 5. Validation and functional characterization of HCIPs produced through an interactome study. **(A–B)** Illustration of experimental assays used for validating the complex formation between bait proteins and HCIPs. To examine the interaction between a bait protein and its HCIP, cell lysates can be subjected to immunoprecipitation (A) and pull-down (B) assays. **(C)** Immunofluorescence assay can be used to examine the co-localization of bait proteins with their HCIPs. **(D)** The bait protein-HCIP complex formation can be further characterized through mapping the regions required for their interaction. **(E)** Illustration of the HCIP knockout (KO) cell generation using CRISPR/Cas9 technique. **(F)** Rescue experiments can be performed in the HCIP KO cells to determine whether the bait protein binding is required for HCIP to regulate the related cellular functions such as cell proliferation and migration.

used to examine protein complex dynamics in cells [63]. Size exclusion chromatography can separate different protein complexes based on their size [64]. Combined with MS analysis, these approaches provide options for elucidating interacting proteins for untagged bait protein, although additional factors could be introduced to affect protein complex formation (e.g. temperature, chromatography sample preparation). These approaches can be concurrently performed when analyzing limited bait proteins, whereas this strategy may not be feasible for a large-scale multi-bait interactome study.

Expressing a bait protein in cells will also bring in the overexpression issue. To address it, knock-in approach can be adopted to integrate a tag into the bait protein-coding region via CRISPR technique, so we can directly purify endogenous bait protein-associated protein complex. To accelerate the progress, an inducible lentiviral system can be used to establish the stable cells for a bait protein, where doxycycline concentration is optimized to make the level of exogenously expressed bait protein close to that of endogenous one [16, 65].

In addition, the steps of cell lysing and followed protein complex purification can lead to protein degradation, loss of weak

and transient interactors, and non-specific binding [66]. To reduce these problems, isolating associated protein complex from cells should be finished in a timely manner and avoid frequent changes of buffer systems between different steps.

Third, identifying *bona fide* interacting partners can be hindered by a vast number of contaminants (i.e. non-specific binding proteins) during the purification of bait protein-associated protein complex. This issue can be solved by including a group of control experiments for MS data filtration. Before that, users are suggested to carefully examine their control experiments to ensure that they are appropriate and unrelated to their bait proteins. Another way to reduce contaminants is to apply different purification approaches (e.g. TAP and BioID) (Figure 1) to the same bait protein and then compare their identified HCIPs. Not only would this confirm true binding proteins for a bait protein, but also allow the user to reduce the contaminants due to technical issues, thus making the interactome analysis more robust.

Lastly, interactome data analysis has become more comprehensive with online databases and software constantly evolving, allowing the generation of more informative PPI datasets based

Table 1. Summary of available bioinformatic resources for analyzing interactome-related proteomics data

Name	Annotation	Website	Reference
CRAPome	Data Filtering	https://reprint-apms.org/	[40]
CompPASS	Data Filtering		[14]
MUSE	Data Filtering		[23]
MiST	Data Filtering	https://modbase.compbio.ucsf.edu/mist/	[43]
BioPlex	Compare Datasets	https://bioplex.hms.harvard.edu/	[46]
BioGRID	Compare Datasets	https://thebiogrid.org/	[45]
AlphaFold	Structural Analysis	https://alphafold.ebi.ac.uk/	[52, 53]
RoseTTAFold	Structural Analysis	https://www.ipd.uw.edu/2021/07/rosettafold-accurate-protein-structure-prediction-accessible-to-all/	[54]
TissueNet	Tissue Association for PPI	https://netbio.bgu.ac.il/tissuenet3/	[55]
IID	Condition Association for PPI	http://iid.ophid.utoronto.ca/	[56]
WGCNA	Explore PPI via Gene Expression Profiles		[57]
PANTHER	Gene Ontology	http://pantherdb.org/	[49]
Metascape	Gene Ontology	http://metascape.org/	[50]
DAVID Bioinformatics	Gene Ontology	https://david.ncifcrf.gov/	[48]
R Studio	Visualization/Graphs Clustering/Heatmaps	https://www.rstudio.com/	Open-Source License
Cytoscape	Visualization of PPI network	https://cytoscape.org/	[58, 59]

The commonly used bioinformatic resources for interactome studies are listed, which include the tools for proteomics data filtration, GO analysis, PPI databases for HCIP compare, PPI network visualization.

on needs. For example, PPI dataset can be further integrated with cancer-related databases (e.g. TCGA), which can help annotate the produced HCIPs from a cancer-related perspective and provide opportunities for identifying new therapeutic strategies for cancer treatment.

Collectively, we review the commonly used methods/resources for characterizing cellular PPI networks and propose a simple pipeline for researchers to process the related large-scale MS data. As mentioned earlier, deconvoluting the MS data into a list of HCIPs and validating the HCIP dataset are just the beginning of the study. The goal of the entire work is to reveal valuable HCIPs for in-depth functional studies to advance our understanding of the mechanisms underlying the related biological questions. We hope this work would not only help alleviate the fears newcomers may face when trying to piece together bioinformatic methods/resources to analyze large-scale proteomic data, but also aid users with a user-friendly pipeline that incorporates details behind the methods/resources needed to identify *bona fide* interacting proteins for their research.

Key Points

- A review of the commonly used biochemical methods for identifying interacting proteins for a protein of interest.
- A summary of the bioinformatic tools and resources for analyzing interactome-based mass spectrometry data.
- A proposed guideline for taking proteomics as an approach to study protein–protein interactions in biological research.

Author contribution

W.W. conceived and supervised the study; R.E.K., D.A. and W.W. wrote the manuscript.

Acknowledgment

We thank all the members in the Wang lab for constructive discussion.

Funding

NIH grants (R01GM126048, R01GM143233) and American Cancer Society Research Scholar grants (RSG-18-009-01-CCG, TLC-21-165-01-TLC to W.W.); NIH National Center for Research Resources and the National Center for Advancing Translational Sciences (UL1 TR001414) through the UC Irvine Institute for Clinical and Translational Science (ICTS) pilot award.

References

1. Pawson T, Nash P. Protein-protein interactions define specificity in signal transduction. *Genes Dev* 2000;**14**:1027–47.
2. Ryan DP, Matthews JM. Protein-protein interactions in human disease. *Curr Opin Struct Biol* 2005;**15**:441–6.
3. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;**422**:198–207.
4. Li Y. The tandem affinity purification technology: an overview. *Biotechnol Lett* 2011;**33**:1487–99.
5. Roux KJ, Kim DI, Raida M, et al. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J Cell Biol* 2012;**196**:801–10.
6. Lam SS, Martell JD, Kamer KJ, et al. Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nat Methods* 2015;**12**:51–4.
7. Rao VS, Srinivas K, Sujini GN, et al. Protein-protein interaction detection: methods and analysis. *Int J Proteomics* 2014;**2014**:147648.
8. Keefe AD, Wilson DS, Seelig B, et al. One-step purification of recombinant proteins using a nanomolar-affinity streptavidin-binding peptide, the SBP-tag. *Protein Expr Purif* 2001;**23**:440–6.

9. Kellermann OK, Ferenci T. Maltose-binding protein from *Escherichia coli*. *Methods Enzymol* 1982;**90 Pt E**:459–63.
10. Smith DB, Johnson KS. Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene* 1988;**67**:31–40.
11. Janknecht R, de Martynoff G, Lou J, et al. Rapid and efficient purification of native histidine-tagged protein expressed by recombinant vaccinia virus. *Proc Natl Acad Sci USA* 1991;**88**:8972–6.
12. Rigaut G, Shevchenko A, Rutz B, et al. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* 1999;**17**:1030–2.
13. Dunham WH, Mullin M, Gingras AC. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* 2012;**12**:1576–90.
14. Sowa ME, Bennett EJ, Gygi SP, et al. Defining the human deubiquitinating enzyme interaction landscape. *Cell* 2009;**138**:389–403.
15. Wang W, Huang J, Chen J. Angiotensin-like proteins associate with and negatively regulate YAP1. *J Biol Chem* 2011;**286**:4364–70.
16. Wang W, Huang J, Wang X, et al. PTPN14 is required for the density-dependent control of YAP1. *Genes Dev* 2012;**26**:1959–71.
17. Wang W, Li X, Huang J, et al. Defining the protein-protein interaction network of the human hippo pathway. *Mol Cell Proteomics* 2014;**13**:119–31.
18. Li X, Wang W, Wang J, et al. Proteomic analyses reveal distinct chromatin-associated and soluble transcription factor complexes. *Mol Syst Biol* 2015;**11**:775.
19. Wang W, Li X, Lee M, et al. FOXKs promote Wnt/beta-catenin signaling by translocating DVL into the nucleus. *Dev Cell* 2015;**32**:707–18.
20. Wang W, Xiao ZD, Li X, et al. AMPK modulates hippo pathway activity to regulate energy homeostasis. *Nat Cell Biol* 2015;**17**:490–9.
21. Wang W, Li N, Li X, et al. Tankyrase inhibitors target YAP by stabilizing Angiotensin family proteins. *Cell Rep* 2015;**13**:524–32.
22. Li X, Wang W, Xi Y, et al. FOXR2 interacts with MYC to promote its transcriptional activities and tumorigenesis. *Cell Rep* 2016;**16**:487–97.
23. Li X, Tran KM, Aziz KE, et al. Defining the protein-protein interaction network of the human protein tyrosine phosphatase family. *Mol Cell Proteomics* 2016;**15**:3030–44.
24. Li X, Han H, Zhou MT, et al. Proteomic analysis of the human tankyrase protein interaction network reveals its role in pexophagy. *Cell Rep* 2017;**20**:737–49.
25. Vargas RE, Duong VT, Han H, et al. Elucidation of WW domain ligand binding specificities in the hippo pathway reveals STXP4 as YAP inhibitor. *EMBO J* 2020;**39**:e102406.
26. Seo G, Han H, Vargas RE, et al. MAP4K Interactome reveals STRN4 as a key STRIPAK complex component in hippo pathway regulation. *Cell Rep* 2020;**32**:107860.
27. Bian W, Tang M, Jiang H, et al. Low-density-lipoprotein-receptor-related protein 1 mediates notch pathway activation. *Dev Cell* 2021;**56**:2902–2919 e2908.
28. Kattan RE, Han H, Seo G, et al. Interactome analysis of human phospholipase D and phosphatidic acid-associated protein network. *Mol Cell Proteomics* 2022;**21**:100195.
29. Kim JS, Raines RT. Ribonuclease S-peptide as a carrier in fusion proteins. *Protein Sci* 1993;**2**:348–56.
30. Young KH. Yeast two-hybrid: so many interactions, (in) so little time. *Biol Reprod* 1998;**58**:302–11.
31. Kerppola TK. Design and implementation of bimolecular fluorescence complementation (BiFC) assays for the visualization of protein interactions in living cells. *Nat Protoc* 2006;**1**:1278–86.
32. Li P, Wang L, Di LJ. Applications of protein fragment complementation assays for analyzing biomolecular interactions and biochemical networks in living cells. *J Proteome Res* 2019;**18**:2987–98.
33. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007;**4**:207–14.
34. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 2005;**4**:1419–40.
35. Prince JT, Carlson MW, Wang R, et al. The need for a public proteomics repository. *Nat Biotechnol* 2004;**22**:471–2.
36. Jones P, Cote RG, Cho SY, et al. PRIDE: new developments and new datasets. *Nucleic Acids Res* 2008;**36**:D878–83.
37. Ternent T, Csordas A, Qi D, et al. How to submit MS proteomics data to ProteomeXchange via the PRIDE database. *Proteomics* 2014;**14**:2233–41.
38. Jarnuczak AF, Vizcaino JA. Using the PRIDE database and ProteomeXchange for submitting and accessing public proteomics datasets. *Curr Protoc Bioinformatics* 2017;**59**:13.31.11–13.31.12.
39. Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 2019;**47**:D442–50.
40. Mellacheruvu D, Wright Z, Couzens AL, et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* 2013;**10**:730–6.
41. Choi H, Larsen B, Lin ZY, et al. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat Methods* 2011;**8**:70–3.
42. Teo G, Liu G, Zhang J, et al. SAINTexpress: improvements and additional features in significance analysis of INTeractome software. *J Proteomics* 2014;**100**:37–43.
43. Jager S, Cimermancic P, Gulbahce N, et al. Global landscape of HIV-human protein complexes. *Nature* 2011;**481**:365–70.
44. Tabb DL, Vega-Montoto L, Rudnick PA, et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res* 2010;**9**:761–76.
45. Chatr-Aryamontri A, Oughtred R, Boucher L, et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 2017;**45**:D369–79.
46. Huttlin EL, Ting L, Bruckner RJ, et al. The BioPlex network: a systematic exploration of the human interactome. *Cell* 2015;**162**:425–40.
47. Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 2017;**45**:D362–8.
48. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;**4**:44–57.
49. Mi H, Lazareva-Ulitsky B, Loo R, et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* 2005;**33**:D284–8.
50. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 2019;**10**:1523.
51. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009;**19**:1639–45.

52. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
53. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* 2022;**13**:1265.
54. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.
55. Barshir R, Basha O, Eluk A, et al. The TissueNet database of human tissue protein-protein interactions. *Nucleic Acids Res* 2013;**41**:D841–4.
56. Kotlyar M, Pastrello C, Sheahan N, et al. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res* 2016;**44**:D536–41.
57. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;**4**:Article17.
58. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
59. Cline MS, Smoot M, Cerami E, et al. Integration of biological networks and gene expression data using cytoscape. *Nat Protoc* 2007;**2**:2366–82.
60. Saito R, Smoot ME, Ono K, et al. A travel guide to cytoscape plugins. *Nat Methods* 2012;**9**:1069–76.
61. Jones PJ, Mair P, McNally RJ. Visualizing psychological networks: a tutorial in R. *Front Psychol* 2018;**9**:1742.
62. Merico D, Gfeller D, Bader GD. How to visually interpret biological data using networks. *Nat Biotechnol* 2009;**27**:921–4.
63. Tan CSH, Go KD, Bisteau X, et al. Thermal proximity coaggregation for system-wide profiling of protein complex dynamics in cells. *Science* 2018;**359**:1170–7.
64. Bloustine J, Berejnov V, Fraden S. Measurements of protein-protein interactions by size exclusion chromatography. *Biophys J* 2003;**85**:2619–23.
65. Han H, Nakaoka HJ, Hofmann L, et al. The hippo pathway kinases LATS1 and LATS2 attenuate cellular responses to heavy metals through phosphorylating MTF1. *Nat Cell Biol* 2022;**24**:74–87.
66. Miteva YV, Budayeva HG, Cristea IM. Proteomics-based methods for discovery, quantification, and validation of protein-protein interactions. *Anal Chem* 2013;**85**:749–68.