# CANVS: an easy-to-use application for the analysis and visualization of mass spectrometry–based protein–protein interaction/association data

Erick F. Velasquez[a], Yenni A. Garcia[a], Ivan Ramirez[a], Ankur A. Gholkar[a], and Jorge Z. Torres[a,b,c,*]
[a]Department of Chemistry and Biochemistry, [b]Molecular Biology Institute, and [c]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095

**ABSTRACT** **The elucidation of a protein's interaction/association network is important for defining its biological function. Mass spectrometry–based proteomic approaches have emerged as powerful tools for identifying protein–protein interactions (PPIs) and protein–protein associations (PPAs). However, interactome/association experiments are difficult to interpret, considering the complexity and abundance of data that are generated. Although tools have been developed to identify protein interactions/associations quantitatively, there is still a pressing need for easy-to-use tools that allow users to contextualize their results. To address this, we developed CANVS, a computational pipeline that cleans, analyzes, and visualizes mass spectrometry–based interactome/association data. CANVS is wrapped as an interactive Shiny dashboard with simple requirements, allowing users to interface easily with the pipeline, analyze complex experimental data, and create PPI/A networks. The application integrates systems biology databases such as BioGRID and CORUM to contextualize the results. Furthermore, CANVS features a Gene Ontology tool that allows users to identify relevant GO terms in their results and create visual networks with proteins associated with relevant GO terms. Overall, CANVS is an easy-to-use application that benefits all researchers, especially those who lack an established bioinformatic pipeline and are interested in studying interactome/association data.**

**Monitoring Editor**
Doug Kellogg
University of California, Santa Cruz

Received: May 20, 2021
Revised: Jul 26, 2021
Accepted: Aug 19, 2021

## INTRODUCTION

Direct protein–protein interactions (PPIs) and indirect protein–protein associations (PPAs) are critical to understanding the biological function of a protein of interest (POI). PPI/As can inform on how a POI is compartmentalized within a cell, how it forms higher-order complexes, how it is regulated, and how it coordinates with other proteins in a spatial and temporal manner to carry out specific cellular processes (Yugandhar *et al.*, 2019; Lu *et al.*, 2020). More broadly, PPI/A networks have been used to analyze the composition of cellular structures such as centrosomes, kinetochores, cilia, and other organelles and to define the function of cell cycle proteins (Torres *et al.*, 2011; Firat-Karalar and Stearns, 2015; Cheung *et al.*, 2016; Go *et al.*, 2019; Remnant *et al.*, 2019; Garcia *et al.*, 2021; Guo *et al.*, 2021). Several approaches have been used to identify PPIs, including yeast two-hybrid (Fields and Song, 1989; Chien *et al.*, 1991), fluorescence resonance energy transfer (Selvin, 1995), and protein fragment complementation assay (Michnick *et al.*, 2000). Although these classical methods are important for identifying and validating PPIs, mass spectrometry–based approaches have made high-throughput identification of PPIs possible (Yugandhar *et al.*, 2019). Affinity purification mass spectrometry (AP-MS) has become the conventional method of identifying PPIs, since it isolates protein

This article was published online ahead of print in MBoC in Press (http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E21-05-0257).

Conflicts of interest: The authors have no conflicts of interest.

*Address correspondence to: Jorge Z. Torres (torres@chem.ucla.edu).

Abbreviations used: AP-MS, affinity purification mass spectrometry; app, application; BioGRID, Biological General Repository for Interaction Datasets; CANVS, Clean Analyze Network Visualization Software; CORUM, Comprehensive Resource of Mammalian Protein Complexes; GO, Gene Ontology; MS, mass spectrometry; POI, protein of interest; PPA, protein–protein association; PPI, protein–protein interaction.

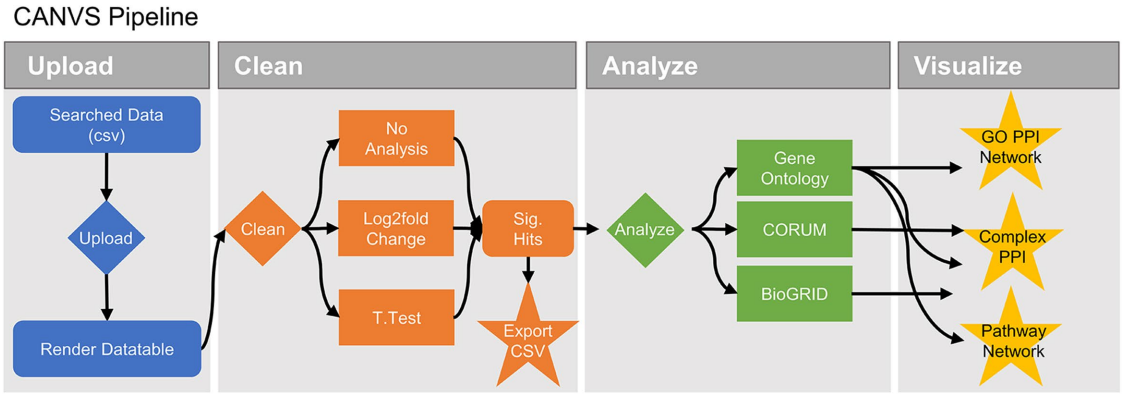


**FIGURE 1:** CANVS workflow. Mass spectrometry data files, comma-delimited text files, with protein UniProt accession numbers, protein descriptions, protein quantitative values (scores), and bait POIs are uploaded and rendered as interactive data tables. To clean the data, users can determine the significance of the identified proteins, given proper controls, using log 2–fold change and a Student's *t* test. Significant protein identifications are then analyzed by applying Gene Ontology (GO) terms, the Comprehensive Resource of Mammalian Protein Complexes (CORUM) database, and the Biological General Repository for Interaction Datasets (BioGRID) database. The visNetwork R package is then used to visualize the GO PPI/A, CORUM PPI/A, and BIOGRID PPI/A networks.

complexes from cell lysates under near-physiological conditions (Gingras *et al.*, 2007). More recently, the field has transitioned to defining PPAs, which may represent direct protein–protein interactions or local protein neighborhoods, through proximity-dependent biotinylation methods (Perkins *et al.*, 2010). A popular method is BioID, where a POI is tagged with a promiscuous biotin ligase, which biotinylates proteins in close proximity to the POI in the presence of biotin (Sears *et al.*, 2019). Overall, popular pipelines for PPI/A approaches involve identifying a POI, tagging the POI with an appropriate protein tag, expression of the tagged-POI, biochemical purifications, MS of purifications to identify proteins, and qualitative/quantitative bioinformatic analyses of identified proteins (Sears *et al.*, 2019; Yugandhar *et al.*, 2019).

Mass spectrometry-based approaches to identify PPI/As generate large amounts of PPI/A data, which are difficult to analyze and interpret. To overcome these issues, computational tools have been developed to analyze PPI/A data quantitatively (Choi *et al.*, 2011, 2012; Nesvizhskii, 2012; Teo *et al.*, 2014), to create visual representations of the analyses (Knight *et al.*, 2017), and to generate PPI/A networks (Shannon *et al.*, 2003; Szklarczyk *et al.*, 2019). One pipeline, APOSTL, integrates these steps and automates the process within a Galaxy framework (Kuenzi *et al.*, 2016). However, currently APOSTL does not allow users to filter search results by Gene Ontology–based (GO) terms or integrate protein-complex data. Furthermore, most PPI/A data analysis computational tools focus on the accuracy of protein identification, instead of on how the identified proteins might be associated at the molecular level. With this in mind, we sought to develop a computational pipeline that allowed users with no computational background to explore PPI/A data interactively within the context of relevant biological processes, molecular functions, cellular components, and protein–protein complex interactions.

Here, we present CANVS (Clean Analyze Network Visualization Software), an open access computational pipeline that cleans mass-spectrometry PPI/A data through statistical analyses and annotates identified proteins with proteoinformatic databases such as BioGRID (Biological General Repository for Interaction Datasets; Oughtred *et al.*, 2019) and CORUM (Comprehensive Resource of Mammalian Protein Complexes) (Giurgiu *et al.*, 2019) to create protein interaction/association networks. Furthermore, CANVS allows users to apply GO (Gene Ontology) (Ashburner *et al.*, 2000; Consortium GO, 2021) filters to create protein networks relevant to biological processes, cellular locations, or molecular functions of interest. To ensure accessibility to all researchers, CANVS is wrapped in a Golem framework (Fay *et al.*, 2021) and is deployed as a Shiny dashboard app that can be downloaded and installed locally on a Windows system (https://sourceforge.net/projects/canvs/files/). CANVS can be used as a standalone tool; however, the user can also upload results from other proteomic pipelines to generate protein networks quickly and identify proteins with relevant biological associations to a POI. Overall, CANVS provides an easy-to-use interactive framework where proteoinformatic resources are integrated to quantify, contextualize, and visualize data from PPI/A experiments that enable users to better understand the biological role of their POI.

## RESULTS AND DISCUSSION

### Features

CANVS is an open-access easy-to-use pipeline for studying protein–-protein interactome/association data. It was created so that researchers with no computational background can quickly analyze mass spectrometry data from affinity-based and proximity-based protein purifications, with an emphasis on identifying biologically interesting PPI/As that can be further validated and explored. Briefly, the CANVS pipeline (Figure 1) can be divided into four steps: 1) uploading data, 2) cleaning data to identify significant protein hits, 3) analyzing results by applying proteoinformatic databases, and 4) visualizing the resulting PPI/A networks. The CANVS pipeline takes advantage of user interface web development packages in R and is wrapped as a Shiny dashboard app that can be installed locally on a Windows system.

***Step 1. Data upload/preprocessing.*** CANVS accepts csv or tab-delimited text files with five columns containing information on: protein UniProt (Apweiler *et al.*, 2004) accession number, score or quantitative value, protein description, file name, and protein bait name (Supplemental Material—Uploading Data). Files are then merged, and the merged data table can be accessed in the *Upload* tab (Supplemental Figure S1). Additionally, the user has the option of uploading one data table each, for the control and experimental, that has all the information about replicates/conditions. The data tables

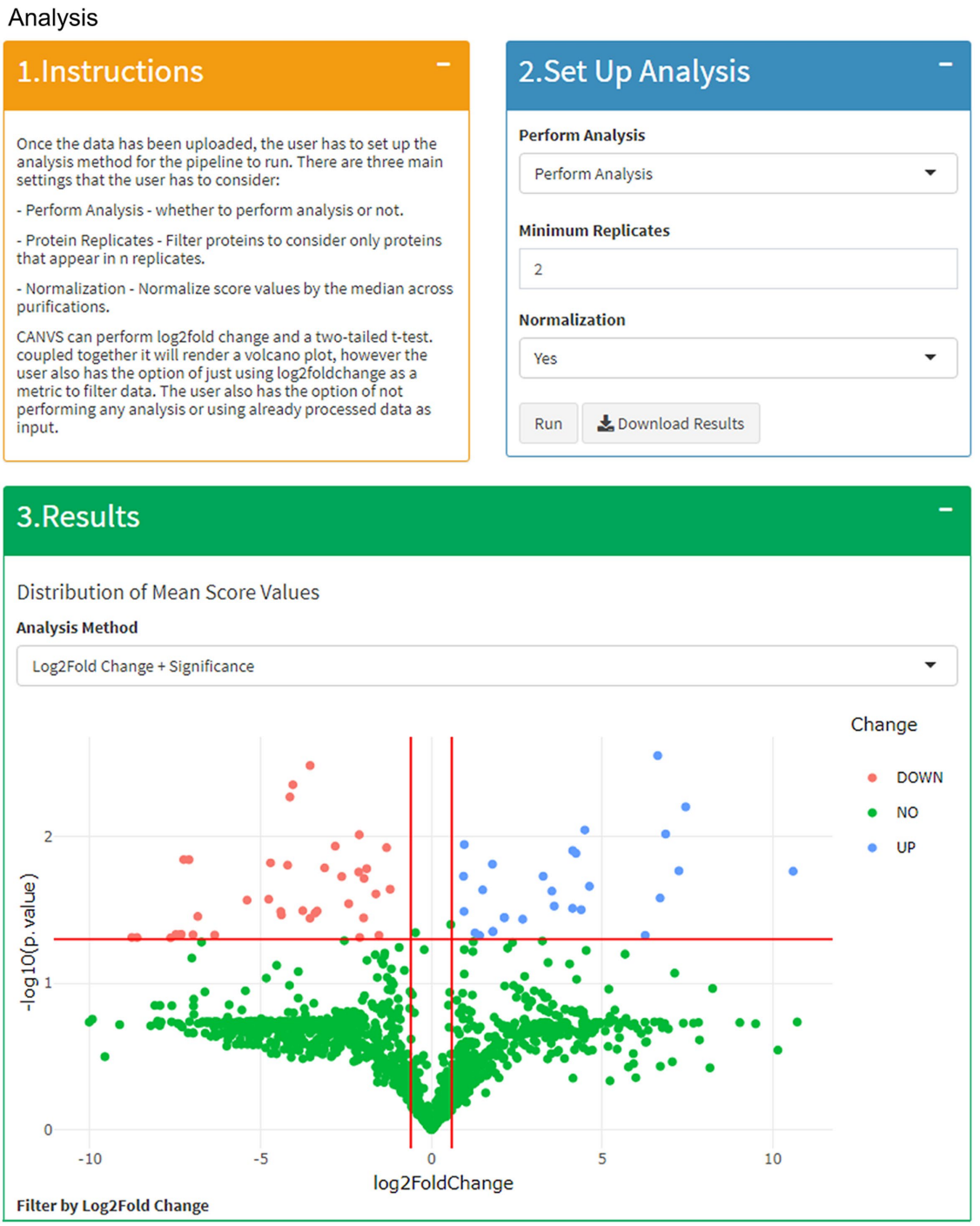


**FIGURE 2:** CANVS cleaning method. CANVS allows users to upload interaction/association MS data, filter by a minimum number of replicates a protein should be present in, normalize proteins by the median value of each purification, and apply significance statistics. CANVS calculates the log 2–fold change and *p* values that can be visualized in an interactive volcano plot. The user can then filter by a certain *p* value or fold change and the results are used in the pipeline for further analysis/visualization.

are interactive and the user can search for specific keywords in the search section. Preprocessing involves both determining how many replicates a protein needs to appear in to carry out the analysis and normalization across replicates and baits (Figure 2 and Supplemental Figures S2–S4). High-throughput MS-based approaches to identify PPI/As contain systematic biases due to steps in data processing and generation (Chawade *et al.*, 2014). To overcome these biases, the field has adopted normalization methods that aim to make samples more comparable across replicates/conditions (Chawade *et al.*, 2014; Välikangas *et al.*, 2018). CANVS gives the user the option to normalize by the median and scales samples so that each purification has the same median value (Välikangas *et al.*, 2018). We recommend using this method if no previous normalization was performed and if the user suspects variation across purifications due to human error (sample preparation) or analytical errors (device calibration, temperature fluctuations, etc.). Both settings are predefined by the

user in the *Clean* tab, where the user can filter by the number of replicates a protein must be identified in to be considered for further analysis and whether to normalize by the median (Figure 2). Please see the Supplemental Material for details and step-by-step user instructions.

***Step 2. Semi–quantitative/qualitative analysis of protein hits.*** After setting up the preprocessing options, users can determine whether to perform a statistical analysis to identify significant protein hits (Supplemental Material—Cleaning Data). CANVS uses spectral counts as a quantitative representation of protein abundance (Lundgren *et al.*, 2010), specifically the exponentially modified protein abundance index (emPAI), since this score considers the number of peptides per protein, an important metric in IP-MS (Ishihama *et al.*, 2005). Additionally, other label-free or labeled quantitative values can be used, including results from intensity-based quantification, stable isotope labeling using amino acids in cell culture (SILAC), and tandem mass tag (TMT) experiments (Thompson *et al.*, 2003; Ong and Mann, 2006; Zhang *et al.*, 2010). However, the data must be in an array format and representative of the abundance of a protein. If a protein is not present in either a control or an experimental purification, a missing value, CANVS assigns the missing protein half the minimum value of nonmissing proteins in the same purification (Wei *et al.*, 2018).

CANVS has two methods by which the user can filter data, using the log 2–fold change of proteins compared with a control or using a combination of log 2–fold change and significance statistics in the form of a two-tailed Student's *t* test (Student, 1908; Hubner *et al.*, 2010). Calculating the difference in the logarithmic mean protein intensities between experimental and control purifications allows users to identify nonspecific associations that center around zero (Hubner *et al.*, 2010; Singh *et al.*, 2016). If a sufficient number of replicates (we recommend three biological replicates and two technical replicates) are used in the analysis, the user has the option of calculating the significance in the log 2–fold change using a two-tailed Student's *t* test. Comparing the negative log of the *p* value with the log 2–fold change creates a volcano plot where background proteins cluster at zero (Singh *et al.*, 2016). The interactive volcano plot is displayed in the results box of the *Clean* tab (Figure 2 and Supplemental Figures S2–S4). The user then has the option of analyzing the data by log 2–fold change only or log 2–fold change and significance statistics. This is done by changing the drop-down menu under Analysis Method (Supplemental Material—Cleaning Data). The user can filter by a specific fold change of interest or a different *p* value; however, the preset options are set at a log change of 0.6 and a *p* value of 0.05. Additionally, if the user is concerned about the multiple testing problem and does not wish to use a strict *p*-value cutoff, an option is available to adjust the *p* value via the Benjamini and Hochberg method (Benjamini *et al.*, 2001). Briefly, adjusting the *p* value controls the false discovery rate and therefore corrects for the expected number of false positives among all positives that rejected the null hypothesis (Jafari and Ansari-Pour, 2019). Results appear in the form of a data table in the *Clean* tab, and are referenced throughout the rest of the pipeline.

Alternatively, users can elect not to carry out a statistical analysis and use CANVS solely to contextualize and visualize the results. To do so, users can select the *no analysis* option in the *Clean* tab and the program will consider data in the experimental as the results. Even when no statistical analysis is performed, the user must press the run button in the *Clean* tab to let CANVS know that the data uploaded in the experimental should be used as the results. This feature allows CANVS to be integrated easily with any other statistical pipeline of choice. The experimental data will then appear in the form of a data table in the *Clean* tab, and this data table will be referenced in the rest of the pipeline (Figure 2 and Supplemental Figures S2–S4).

***Step 3. Data analysis.*** CANVS features two tabs to analyze and generate visual network representations of the results (Supplemental Material—Analysis/Visualization). The first *Analyze/Visualize* tab (Supplemental Figures S5–S7) creates networks for all identified proteins. The *GO Analyze/Visualize* tab (Supplemental Figures S8 and S9) features a GO-based (Ashburner *et al.*, 2000; Consortium TGO, 2019) filtering tool that allows users to search the protein results for associations with GO terms of interest. The GO database classifies GO terms into three categories: biological processes (BP), molecular function (MF), and cellular component (CC). CANVS links proteins in the results to associated GO terms. Users can search by keyword and CANVS searches for GO terms that have that keyword. GO terms with the keyword and their respective subterms are selected. CANVS then renders two objects: a color-coded network representing how the selected GO terms are related and a data table with all the selected GO terms (Figure 3). Users can then select GO terms of interest in the data table, which instructs CANVS to filter the results for proteins that have the selected GO terms. Alternatively, if users want to filter the results with all of the GO terms in the table, no selection is necessary and the network visuals will reflect proteins associated with all of the GO terms. This feature is particularly helpful if users are interested in specific biological processes, molecular functions, and/or cellular components and want to search the results for proteins associated with such GO terms.

Once the user determines whether to include all proteins in the results or selects certain proteins based on GO terms, proteins in the result can be further annotated and contextualized. CANVS integrates two main databases: the Biological General Repository for Interaction Datasets (BioGRID v. 4.3; Oughtred *et al.*, 2019) and the Comprehensive Resource of Mammalian Protein Complexes (CORUM v. 3.0; Giurgiu *et al.*, 2019). BioGRID contextualizes results in terms of previously identified protein–protein interactions, whereas, CORUM contextualizes the results in terms of known protein–protein complex information. For both databases, the user first defines an organism of interest and annotations are performed within the context of that organism. See the Supplemental Material for a full list of organisms supported by CANVS, BioGRID, and CORUM. By analyzing the results from both databases, the user can identify known PPIs and complex-PPI/As present in the results.

***Step 4. Data visualization.*** After the results have been annotated (using CORUM or BioGRID) or selected (using associated GO terms), three types of networks are created: protein–protein network (Figure 4A), CORUM protein–protein complex network (Figure 4B and Supplemental Figure S10), and a BioGRID protein–protein network (Figure 4C and Supplemental Figure S11). Visual representations are powered by visNetwork, an R package that provides a framework for creating visual networks in an interactive environment (Almende, 2021). Considering a basic network, the user can interact with the graphs by zooming in on certain sections of a network, selecting certain proteins to highlight interactions, and downloading edited networks as png files. Networks can also be reset and created again, a helpful feature when considering different parameters that might influence which proteins are rendered in the networks. The user also has the option to download Cytoscape network files (Shannon *et al.*, 2003), which can be used in Cytoscape to recreate networks developed using CANVS. See the Supplemental

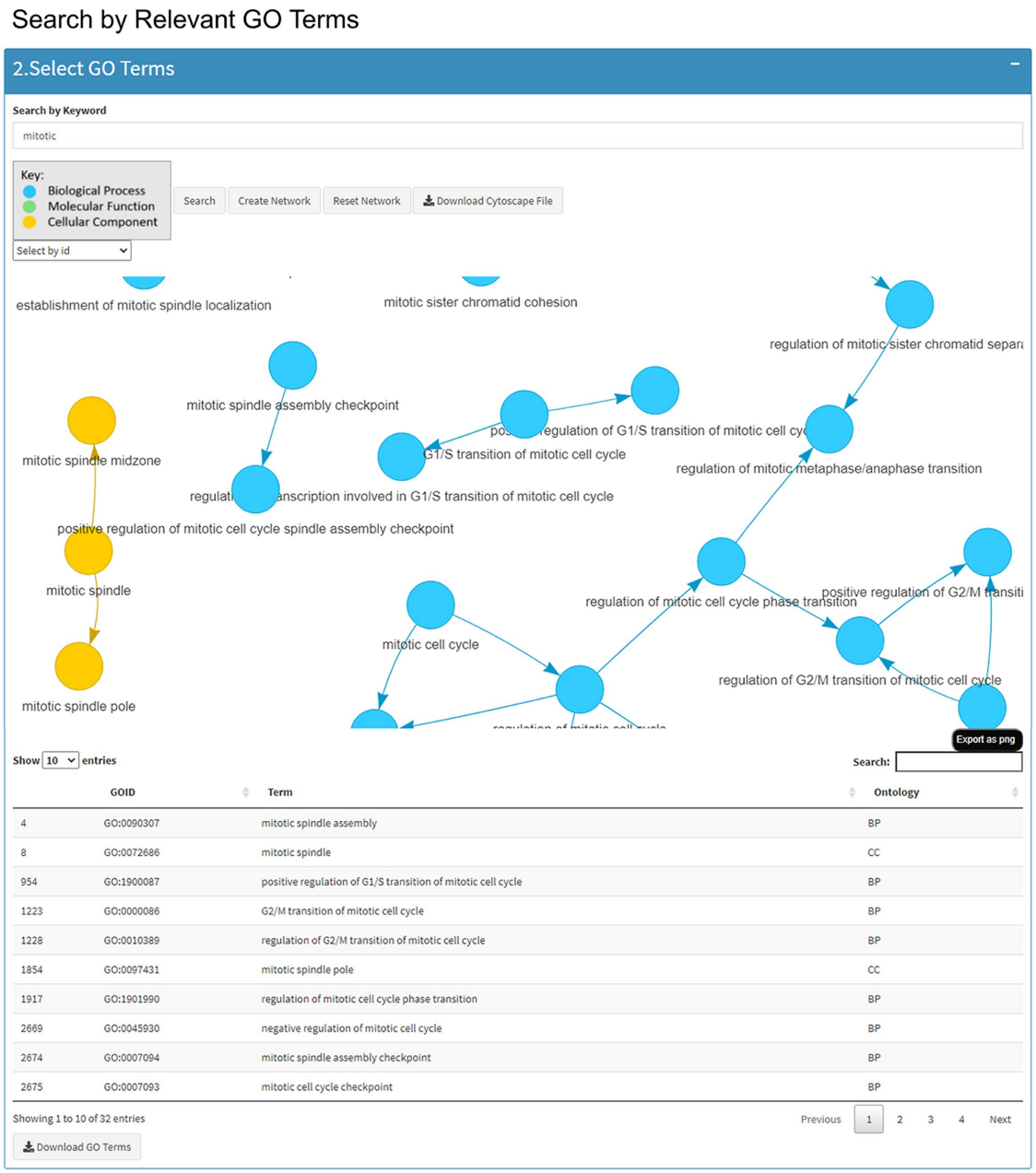


**FIGURE 3:** Selection of Gene Ontology GO terms and filtering results with selected GO terms. Users can perform a keyword search and GO terms containing the keyword/s of interest and any associated subterms are retrieved. Only GO terms associated with protein hits in the dataset will appear and can be selected and applied as a filter. Proteins with the associated GO terms of interest are included in the network tables.

Material for a detailed walkthrough on how to use CANVS network visuals.

## Conclusions

CANVS is an interactive tool that allows scientists to integrate systems biology databases and create PPI/A networks with biologically relevant results. The simple requirements of the application, along with its interactive networks, make CANVS a powerful tool to be used in conjunction with other tools or as a standalone pipeline. CANVS is particularly useful to researchers studying a POI or sets of proteins and wanting to contextualize the results of their interactome/association experiments. The integration of BioGRID allows the user to compare the results to a wide set of interactome experiments and create PPI networks based on previous PPI data. Similarly, the integration of CORUM allows users to identify protein complexes within their results, providing context as to how a POI or bait might be related to other proteins in the network. Additionally, by creating networks where specific molecular functions, biological processes, or cellular components are prioritized, the user can quickly parse through the results and concentrate on PPI/As that are relevant to their scientific question. Ultimately, additional databases and statistical methods can be integrated into the pipeline. It is important to note that protein interactions and associations should be validated biochemically. As examples, we recently used the

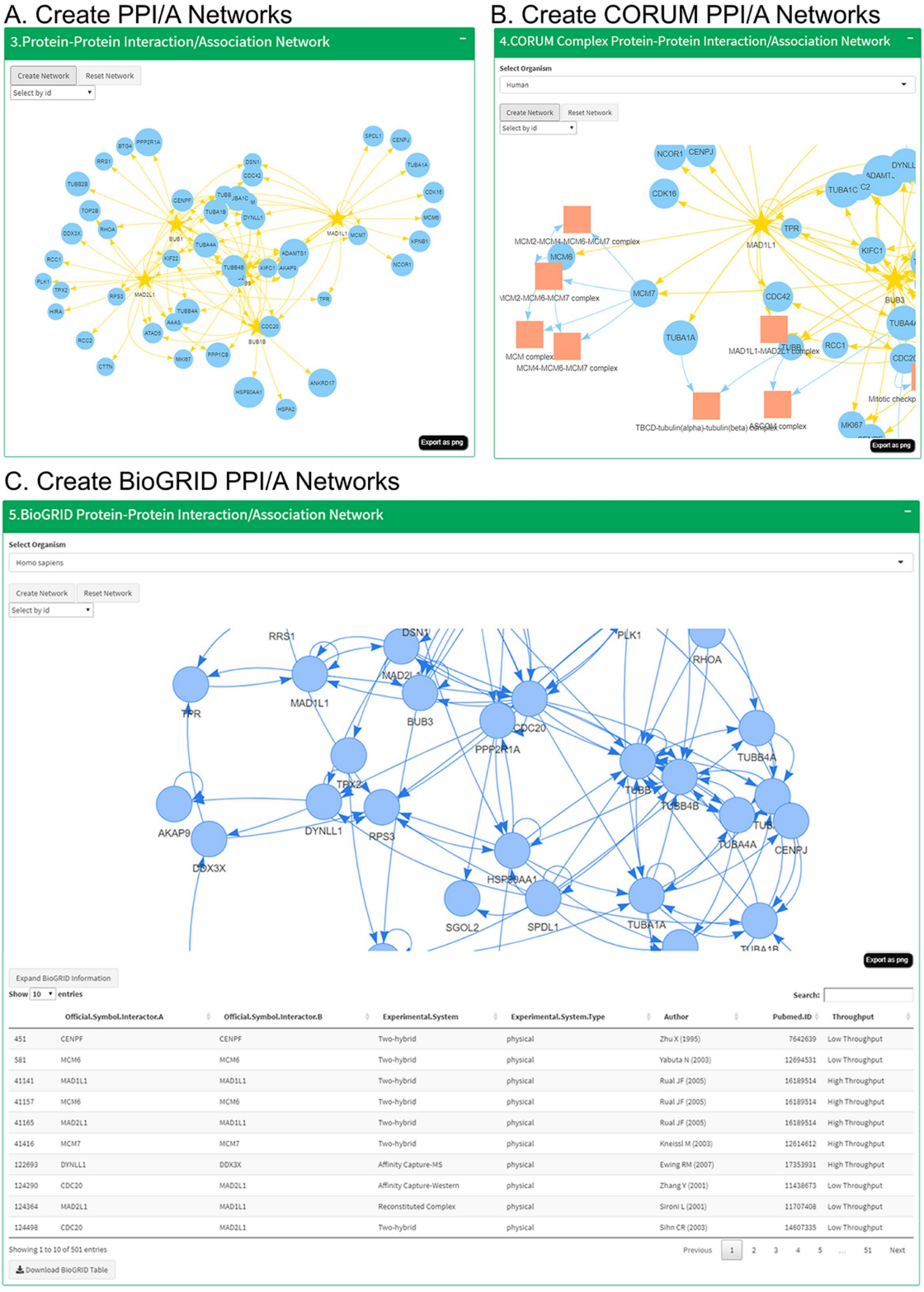


**FIGURE 4:** Creation of interactive PPI/A networks of (A) protein hits associated with the selected GO terms integrating (B) CORUM protein complex information and (C) BioGRID PPI information.

analytical framework included in CANVS to study the PPI/A networks of DUSP7, which helped to define its regulation of ERK2 during mitosis (Guo *et al.*, 2021), and to analyze the PPA networks of core spindle assembly checkpoint proteins (Garcia *et al.*, 2021). Overall, CANVS offers an interactive and easy-to-use solution to study PPI/A data that can be used by laboratories without an established proteomic analysis pipeline.

## MATERIALS AND METHODS

Request a protocol through *Bio-protocol*.

### Testing CANVS with mass spectrometry data from BioID experiments

To test CANVS, we utilized a previously published mass spectrometry data set from BioID-based experiments of core spindle assembly checkpoint proteins (Garcia *et al.*, 2021). Briefly, BioID2-tagged inducible HeLa stable cell lines were generated for core spindle assembly checkpoint (SAC) proteins (BUB1, BUB3, BUBR1, MAD1L1, MAD2L1). These cell lines were induced to express the BioID2-tagged core SAC proteins, incubated with biotin, and BioID purifications were performed in triplicate for each bait except for BUB3 which was performed in duplicate. A BioID2-tag alone cell line was used as control. The purifications were then analyzed by LC-MS/MS and peptide identification was conducted with Mascot (v2.4; Matrix Science, Boston, MA) against the UniProt human database (October 10, 2018). Search parameters included trypsin digestion allowing up to two missed cleavages, carbamidomethyl on cysteine as a fixed modification, oxidation of methionine as a variable modification, 10-ppm peptide mass tolerance, and 0.02-Da fragment mass tolerance. Peptides that surpassed a cut-off score of 20, assuming a 5% false discovery rate, were accepted. From the SAC protein BioID panel, 387 proteins where identified with at least two significant peptides and where further processed using CANVS. To perform a statistical analysis of proteins identified in both the control and experimental, emPAI scores (Ishihama *et al.*, 2005) were considered as quantifiable values. Files compatible with CANVS were then generated by summarizing search results by UniProt accession ID, protein description, quantifiable (emPAI score), file name, and associated bait. This dataset is available at the SourceForge directory (https://sourceforge.net/projects/canvs/files/), where users can download it for reference.

### Statistical analysis

Given quantitative scores, the significance of proteins shared between the experimental and the control can be determined using log 2–fold change values and significance statistics (Hubner *et al.*, 2010; Singh *et al.*, 2016). Proteins are first filtered by the number of replicates each protein is present in, a number chosen by the user. Proteins with the appropriate replicate count are further tested for significance. If a protein is not present in either a control or experimental purification, half of the minimum value of that purification is used (Wei *et al.*, 2018). Fold change is calculated by comparing the mean value of a protein across experimental purifications to the same mean value of a protein across control purifications. CANVS then calculates the log base 2 of the fold change, since it is beneficial to represent the distribution around zero, a value that indicates no change for a protein between the experimental and control. If a sufficient number of replicates (we suggest at least three biological replicates and two technical replicates) are used in the analysis, the user has the option of calculating the significance in the log 2–fold change using a two-tailed Student's *t* test. Additionally, the user can adjust the *p* value using the Benjamini and Hochberg method (Benjamini *et al.*, 2001).

### Integration of system biology databases

Three main system biology databases are integrated in CANVS: the Biological General Repository for Interaction Datasets (BioGRID v. 3.5; Oughtred *et al.*, 2019), the Comprehensive Resource of Mammalian Protein Complexes (CORUM v. 3.0; Giurgiu *et al.*, 2019), and Gene Ontology (Ashburner *et al.*, 2000; Consortium TGO, 2019). To incorporate each database, the entire database was downloaded and transformed into an R object that is then referenced in a function. In the case of Gene Ontology, both GO terms and sub-terms were merged by GO term ID. To identify data that correspond to certain proteins across all three databases, UniProt accession numbers are translated to common gene names using the R package org.HS.eg.db (Carlson, 2019). The package includes local versions of each database. Updates to databases in the package will be performed every six months.

### Development of Shiny dashboard

CANVS follows a Golem framework, with the motivation of creating an easy-to-use application with simplified modules (Fay *et al.*, 2021). Modules, functions, and helper functions were created using the framework outlined in the Golem package. To make the app interactive yet intuitive, a dashboard framework was implemented where the user is able to perform a part of the pipeline in each tab. Visual networks are created using visNetwork, an R package designed to create interactive network visuals (Almende, 2021). Briefly, protein results are formatted to include node information, edge information, and annotations for each node including ID and visual components that can be changed by the user (color, shape, etc.). The networks are then rendered in an interactive Shiny session in the format of a Shiny dashboard. The app was deployed using protocols from DesktopDeployR, a framework for deploying self-contained R-based applications in Windows (https://github.com/wleepang/DesktopDeployR). For step-by-step instructions on how to use CANVS, refer to the Supplemental Material or the main repository.

### Installation

CANVS is packaged as an R Shiny dashboard app, but since it includes data from large databases, the application is deployed locally. Currently the local version of the app can only be used in a Windows system. To install CANVS, download the CANVS.zip file in the main repository (https://sourceforge.net/projects/canvs/files/). Unzip the file in a location where you want to save the application. Open the folder and right click on the CANVS.bat file; then select *create a shortcut*. Name the shortcut CANVS and then drag the shortcut to the desktop. Then double click on the new shortcut and CANVS will open in a web browser window. Please make sure that there are no antivirus software blocking application connections to ports in the computer and that firewall does not block use of ports.


## ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health NIGMS, Grants R35GM139539 and R01GM117475, and the National Science Foundation, Grant MCB1912837 to J.Z.T. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Institutes of Health NIGMS or the National Science Foundation. Y.A.G. was supported by the UCLA Tumor Cell Biology Training Program (USHHS Ruth L. Kirschstein Institutional National Research Service Award #T32CA009056). E.F.V. was supported by a grant to the University of California, Los Angeles from the Howard Hughes Medical Institute through the James H. Gilliam Fellowships for Advanced Study Program and by a UCLA Molecular Biology Institute Whitcome Fellowship. I.R. was supported by a NSF Louis Stokes Alliances for Minority Participation Bridge to the Doctorate Fellowship and a Cota Robles Fellowship. We thank members of the Jackson laboratory at Stanford University for their insightful comments on improving CANVS.


## REFERENCES


Almende BV, Thieurmel B, Robert T (2021). CRAN: visNetwork. R Package Version 2.0.9.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al.* (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* (2000). Gene ontology: tool for the unification of biology. Nat Genet 25.

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001). Controlling the false discovery rate in behavior genetics research. Behavioural Brain Research 125.

Carlson M (2019). org.Hs.eg.db: Genome wide annotation for Human. R Package Version 3.8.2.

Chawade A, Alexandersson E, Levander F (2014). Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. J Proteome Res 13.

Cheung K, Senese S, Kuang J, Bui N, Ongpipattanakul C, Gholkar A, Cohn W, Capri J, Whitelegge JP, Torres JZ (2016). Proteomic Analysis of the mammalian katanin family of microtubule-severing enzymes defines katanin p80 subunit B-like 1 (KATNBL1) as a regulator of mammalian katanin microtubule-severing. Mol Cell Proteomics: MCP 15.

Chien C, Bartel P, Sternglanz R, Fields S (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. Proc Natl Acad Sci USA 88.

Choi H, Larsen B, Lin Z-Y, Breitkreutz A, Mellacheruvu D, Fermin D, Qin ZS, Tyers M, Gingras A-C, Nesvizhskii AI (2011). SAINT: probabilistic scoring of affinity purification-mass spectrometry data. Nat Methods 8.

Choi H, Liu G, Mellacheruvu D, Tyers M, Gingras A-C, Nesvizhskii AI (2012). Analyzing protein–protein interactions from affinity purification-mass spectrometry data with SAINT. Bioinformatics 39, 8.15.1–8.15.23.

Consortium GO (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 49.

Consortium TGO (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 47.

Fay C, Rochette S, Guyader V, Girard C (2021). A Framework for robust shiny applications. https://github.com/ThinkR-open/golem.

Fields S, Song O (1989). A novel genetic system to detect protein–protein interactions. Nature 340.

Firat-Karalar EN, Stearns T (2015). Probing mammalian centrosome structure using BioID proximity-dependent biotinylation. Methods Cell Biol 129.

Garcia YA, Velasquez EF, Gao LW, Gholkar AA, Clutario KM, Cheung K, Williams-Hamilton T, Whitelegge JP, Torres JZ (2021). Mapping proximity associations of core spindle assembly checkpoint proteins. J Proteome Res 20.

Gingras A, Gstaiger M, Raught B, Aebersold R (2007). Analysis of protein complexes using mass spectrometry. Nat Rev Mol Cell Biol 8.

Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Ruepp A (2019). CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res 47.

Go CD, Knight JDR, Rajasekharan A, Rathod B, Hesketh GG, Abe KT, Youn J-Y, Samavarchi-Tehrani P, Zhang H, Zhu LY, *et al.* (2019). A proximity-dependent biotinylation map of a human cell. Nature 595, 120–124.

Guo X, Ramirez I, Garcia YA, Velasquez EF, Gholkar AA, Cohn W, Whitelegge JP, Tofig B, Damoiseaux R, Torres JZ (2021). DUSP7 regulates the activity of ERK2 to promote proper chromosome alignment during cell division. J Biol Chem 296.

Hubner NC, Bird AW, Cox J, Splettstoesser B, Bandilla P, Poser I, Hyman A, Mann M (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J Cell Biol 189.

Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics: MCP 4.

Jafari M, Ansari-Pour N (2019). Why, when and how to adjust your *P* values? Cell Jl 20.

Knight JDR, Choi H, Gupta GD, Pelletier L, Raught B, Nesvizhskii AI, Gingras A-C (2017). ProHits-viz: a suite of web tools for visualizing interaction proteomics data. Nat Methods 14.

Kuenzi BM, Borne AL, Li J, Haura EB, Eschrich SA, Koomen JM, Rix U, Stewart PA (2016). APOSTL: An interactive galaxy pipeline for reproducible analysis of affinity proteomics data. J Proteome Res 15.

Lu H, Zhou Q, He J, Jiang Z, Peng C, Tong R, Shi J (2020). Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. Signal Transduction Targeted Therapy 5.

Lundgren DH, Hwang S-I, Wu L, Han DK (2010). Role of spectral counting in quantitative proteomics. Expert Rev Proteomics 7.

Michnick S, Remy I, Campbell-Valois F, Vallée-Bélisle A, Pelletier J (2000). Detection of protein–protein interactions by protein fragment complementation strategies. Methods Enzymol 328.

Nesvizhskii AI (2012). Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. Proteomics 12.

Ong S-E, Mann M (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). Nat Protocols 1.

Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, *et al.* (2019). The BioGRID interaction database: 2019 update. Nucleic Acids Res 47.

Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C (2010). Transient protein–protein interactions: structural, functional, and network properties. Structure 18, 1233–1243.

Remnant L, Booth DG, Vargiu G, Spanos C, Kerr ARW, Earnshaw WC (2019). In vitro BioID: mapping the CENP—a microenvironment with high temporal and spatial resolution. Mol Biol Cll 30.

Sears RM, May DG, Roux KJ (2019). BioID as a tool for protein-proximity labeling in living cells. Methods Mol Biol 2012.

Selvin P (1995). Fluorescence resonance energy transfer. Methods Enzymol 246.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13.

Singh S, Hein MY, Stewart AF (2016). msVolcano: a flexible web application for visualizing quantitative proteomics data. Proteomics 16.

Student (1908). The probable error of a mean. Biometrika 6, 1-25.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, *et al.* (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research 47.

Teo G, Liu G, Zhang J, Nesvizhskii AI, Gingras A-C, Choi H (2014). SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. J Proteomics 100.

Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AKA, Hamon C (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 75.

Torres JZ, Summers MK, Peterson D, Brauer MJ, James Lee SS, Gholkar AA, Lo Y-C, Lei X, Jung K, Anderson DC, *et al.* (2011). The STARD9/Kif16a kinesin associates with mitotic microtubules and regulates spindle pole assembly. Cell 147.

Välikangas T, Suomi T, Elo LL (2018). A systematic evaluation of normalization methods in quantitative label-free proteomics. Briefings Bioinformat 19.

Wei R, Wang J, Su M, Jia E, Chen S, Chen T, Ni Y (2018). Missing value imputation approach for mass spectrometry–based metabolomics data. Sci Rep 8.

Yugandhar K, Gupta S, Yu H (2019). Inferring protein-protein interaction networks from mass spectrometry–based proteomic approaches: a mini-review. Comput Struct Biotechnoly J 17.

Zhang G, Ueberheide BM, Waldemarson S, Myung S, Molloy K, Eriksson J, Chait BT, Neubert TA, Fenyö D (2010). Protein quantitation using mass spectrometry. Methods Mol Biol 673.

# Supplemental Materials

## Molecular Biology of the Cell

Velasquez *et al*.

**SUPPLEMENTAL MATERIAL**

# CANVS: an easy-to-use application for the analysis and visualization of mass spectrometry-based protein-protein interaction/association data

Erick F. Velasquez[1], Yenni A. Garcia[1], Ivan Ramirez[1], Ankur A. Gholkar[1] and Jorge Z. Torres[1-3*]

[1]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, USA

[2]Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA

[3]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA

***Corresponding author:** Jorge Z. Torres

Email: torres@chem.ucla.edu

# CANVS Instructions

## Getting stared with CANVS

The goal of this document is to get you started with CANVS as quickly as possible.

CANVS is an open-access easy-to-use pipeline for studying protein-protein interaction/association data. CANVS was created so that scientists with no computational background can quickly analyze mass spectrometry data from affinity-based and proximity-based protein purifications, with an emphasis on identifying biologically interesting protein interactions/associations that can be further validated and explored. The pipeline can be divided into the following steps:

1. Uploading data
2. Cleaning data
3. Analysis with system biology databases/Network visuals

The tabs in the application reflect these steps, where the network visuals steps can be done with or without filtering the results with GO terms. Each step is described in more detail below.

### 0. Download and Setup

CANVS is packaged as an R Shiny dashboard app but since it includes data from large databases, the application is deployed locally. Currently the local version of the app can only be used in a windows systems. To install CANVS, download the CANVS.zip file from (https://sourceforge.net/projects/canvs/files/). Unzip the file in a location where you want to save the application. Open the folder and right click on the CANVS.bat file, then select create a shortcut. Name the shortcut CANVS and then drag the shortcut to the desktop. Then double click on the new shortcut and CANVS will open in a web browser window. Please make sure that there are no antivirus software blocking application connections to ports in the computer and that firewall does not block use of ports.

### 1. Uploading Data

CANVS accepts multiple csv or tab delimited files as long as each data table has the following information, along with these exact column names;

- *Accesion* - UNIPROT accession ID.
- *Description* - Description of the protein identified.
- *File* - Unique filename for the purification.
- *Bait* - The bait or protein that was affinity tagged. This is usually the protein of interest that is being studied.

If the user chooses to carry out a semi-quantitative qualitative analysis the user should also include a score column that represents a quantitative value at the protein level.

- *Score* - a quantitative score at the protein level.

Data is uploaded in the "upload" tab, however it's important to remember that the pipeline will only run if the above columns are included in the data.

Instructions of acceptable data structures are summarized in the first box *"1. Instructions"*. Data is uploaded in the second box *"2. Upload"*. Multiple files can be uploaded in each of the control and experimental sections. Once the *Upload Data* button is pressed, the results will be rendered as data tables in the final two boxes, *"3. Control"* and *"4. Experimental"*. Each box can be exapanded in order to view the data tables with the uploaded control and experimental data. Additionally, the user has the option of uploading a single file for either the control or experimental as long as it is formatted with the columns above and replicates are indicated with unique file names in File column.

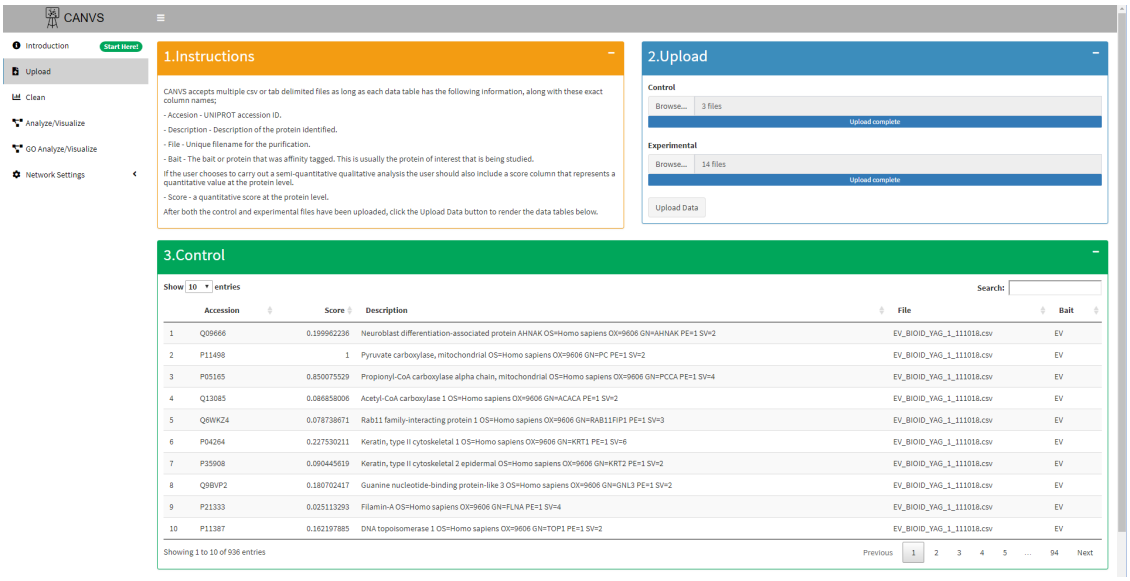


Figure S1: Upload Tab

## 2. Cleaning Data

Once the data has been uploaded, the user has to set up the analysis method for the pipeline to run. There are three main settings that the user has to consider:

- Perform Analysis - whether to perform analysis or not.
- Protein Replicates - Filter proteins to consider only proteins that appear in n replicates.
- Normalization - Normalize score values by the median across purifications.

CANVS can perform log2fold change and a two-tailed t-test. Coupled together it will render a volcano plot, however the user also has the option of just using log2fold change as a metric to filter data. Once the analysis is complete the user can change settings like the log fold value to filter by, the P value to consider or whether to look at UP or DOWN regulated proteins. The user also has the option of not performing any analysis or using already processed data as input

Each item is described in more detail below.

### Statistical Analysis

Once the data is filtered by replicates and normalized, the user has the option of applying statistical tests to proteins shared in the control and experimental proteins.

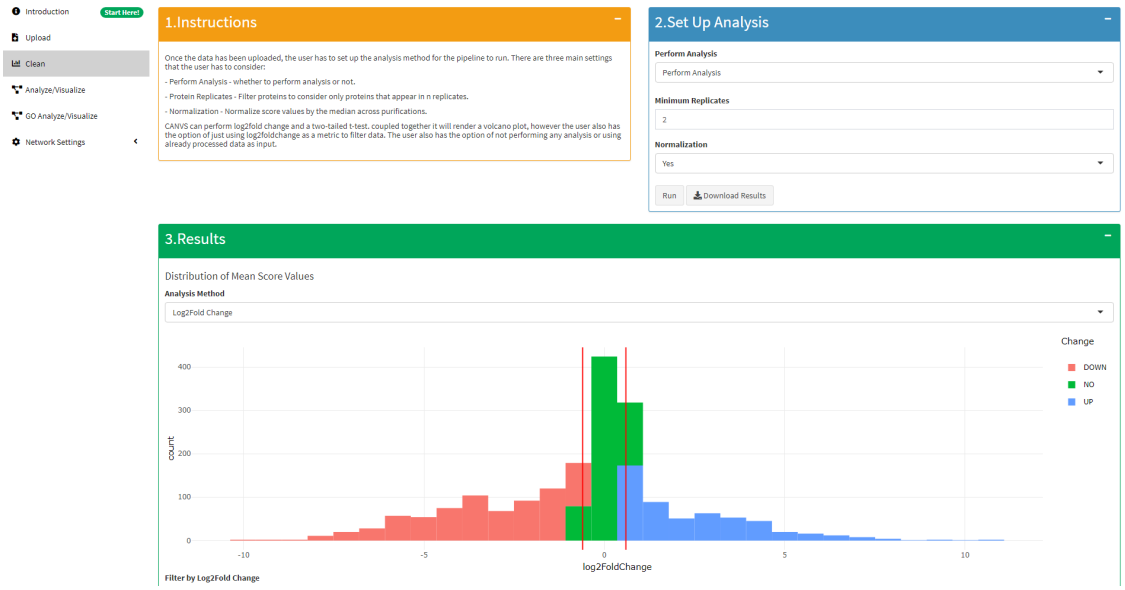


Figure S2: Clean Tab

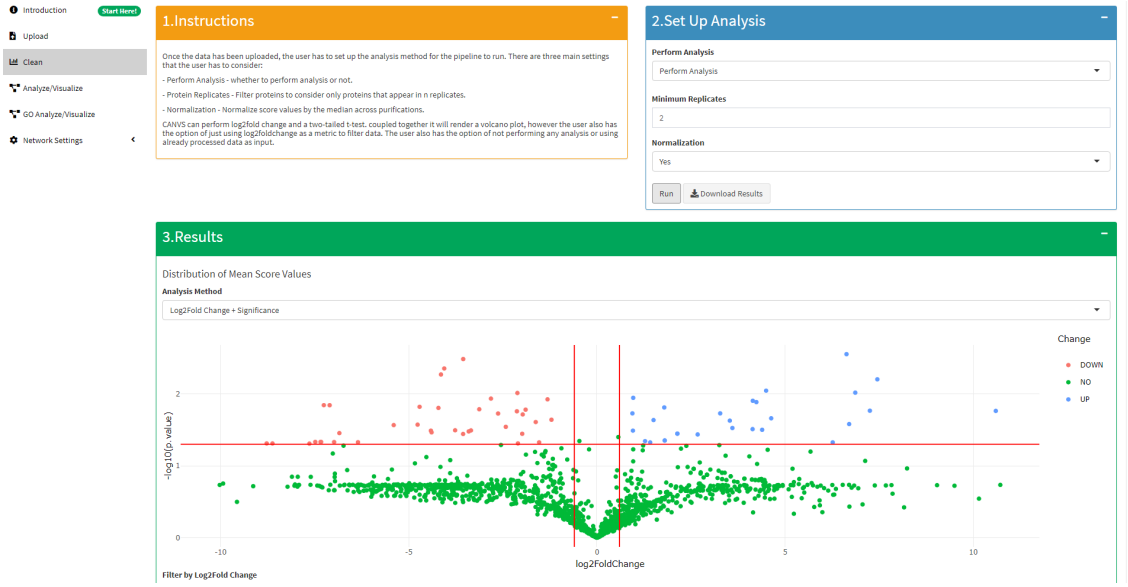


Figure S3: Clean Tab

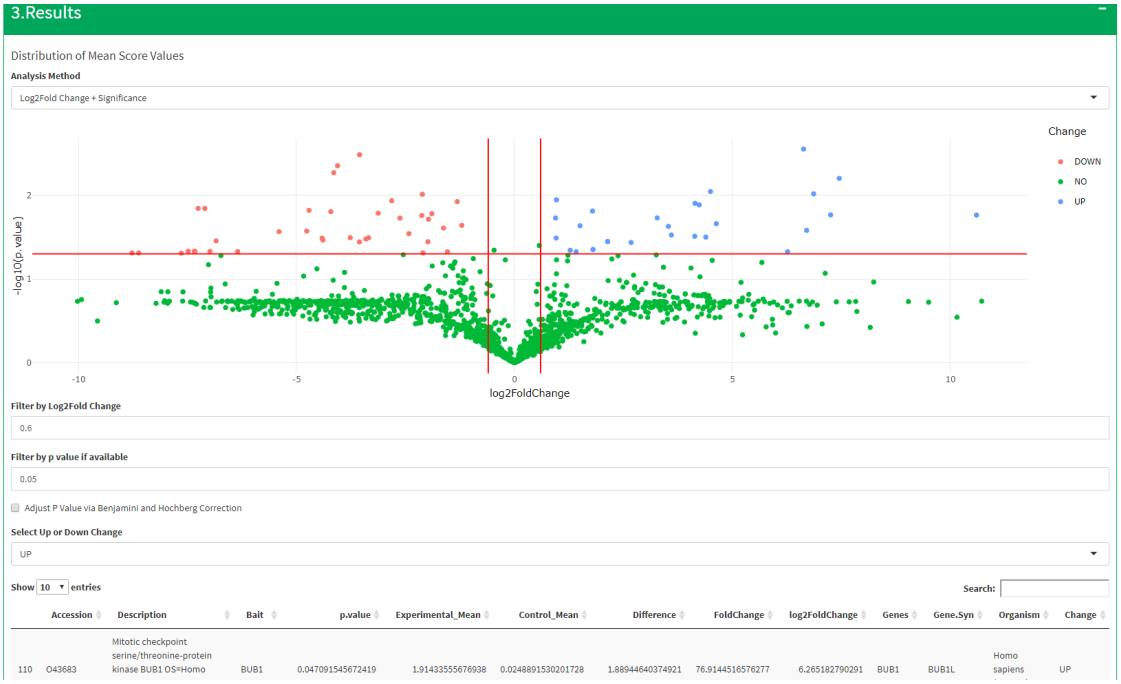


Figure S4: Clean Tab

A. Perform Analysis - CANVS first calculates the difference between the logarithmized mean protein intensities between experimental and control purifications. The log base-2 of the fold change is then calculated since it is beneficial to represent the distribution around zero, a value that indicates no change for a protein between the experimental and control. If a large number of replicates (we recommend three biological replicates and two technical replicates) are used in the analysis, the user has the option of also calculating the significance in the log2fold change using a two-tailed t-test. Comparing the negative log of the P value to the log2fold change creates a volcano plot where background proteins cluster at zero.

B. No Analysis - All proteins in the experimental are considered hits.

For the first option experimental and control data sets must be supplied, however if the user chooses not to carry out a statistical analysis only experimental data is needed. No analysis might be useful if the user has done a quantitative analysis somewhere else and wants to use CANVS for analysis/visual purposes.

### Filtering proteins by replicates

The user has the option to filter proteins by how many times they where identified in multiple replicates. This is particularly helpful to experimental set ups that included multiple replicates and want to increase confidence by only including proteins found in n replicates in the results. Default setting is set to 2 but can be changed using the drop down menu.

### Normalizing

Currently, CANVS supports normalization by the median. Briefly, CANVS scales samples so that each purification has the same median value.

### Results

Results can then be viewed in the box below and the user can download them in a csv format by clicking the Download Results button. The result table will also be used in the rest of the pipeline to create nework visuals. The table and analysis are reactive to user inputs and so if the user chooses to analyze data in a different method, the user would first update results in the analysis tab and then run the rest of the pipeline. The filtering tools allow the user to change the conditions by which to consider hits. The user has the option of filtering either by log2fold change or log2fold change in combination with a given P value. Additionally the user can correct P values by using the Benjamini and Hochberg correction. Briefly, adjusting the P value controls the false discovery rate and therefore corrects for the expected amount of false positives among all positives which rejected the null hypothesis.

## 3. Analysis/Visualization

The application features 2 final tabs that provide the user with visual network representations of the results. The first tab creates networks for all identified proteins while the second tab features a Gene Ontology based filtering tool that allows the user to create visual representations of proteins with certain GO terms. Both tabs feature annotated data in the form of protein-protein and protein-complex networks. To incorporate protein-protein information we integrated the Biological General Repository for Interaction Datasets (BioGRID v. 3.5). To incorporate protein-complex information, we integrated the Comprehensive Resource of Mammalian Protein Complexes (CORUM v. 3.0).

### Protein Networks Considering All Hits

The first visualization tab features 4 different ways to analyze results:

- Bait-Bait Interaction/Association Networks
- Protein-Protein Interaction/Association Networks
- CORUM Complex Protein-Protein Interaction/Association Networks
- BioGRID Protein-Protein Interaction/Association Networks

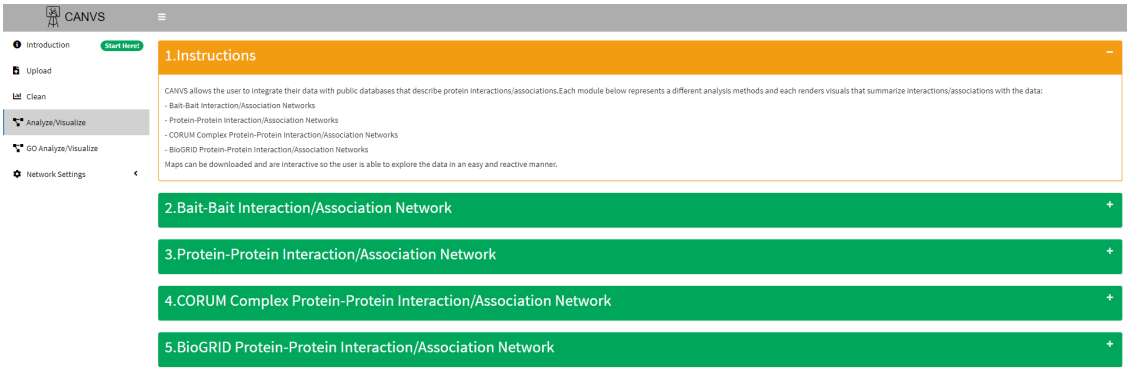


Figure S5: Main Analyze/Visualize Tab

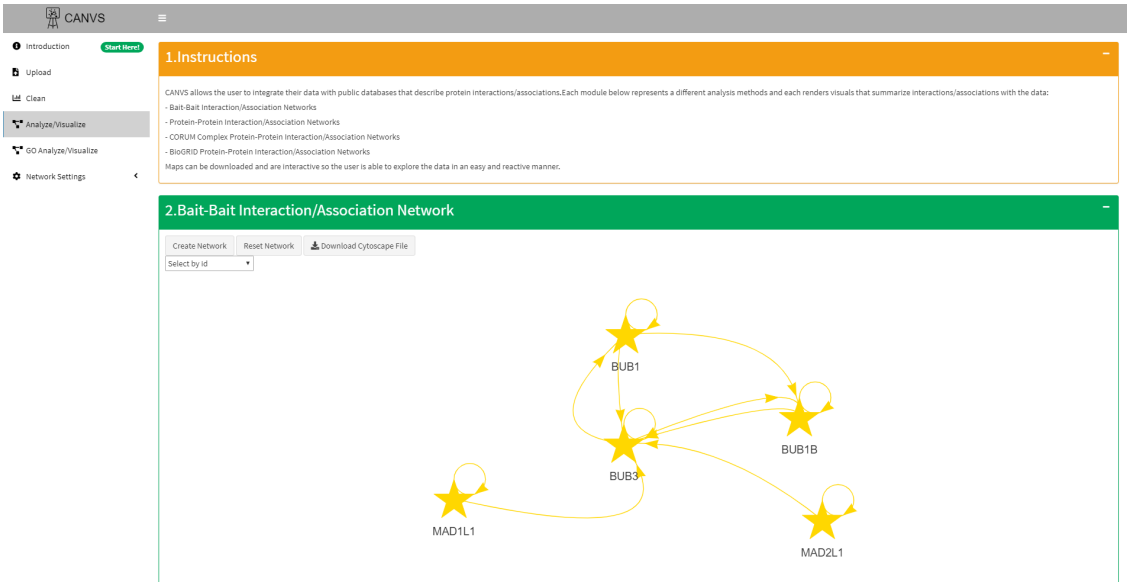


Figure S6: Main Analysze/Visualize Expanded Tab

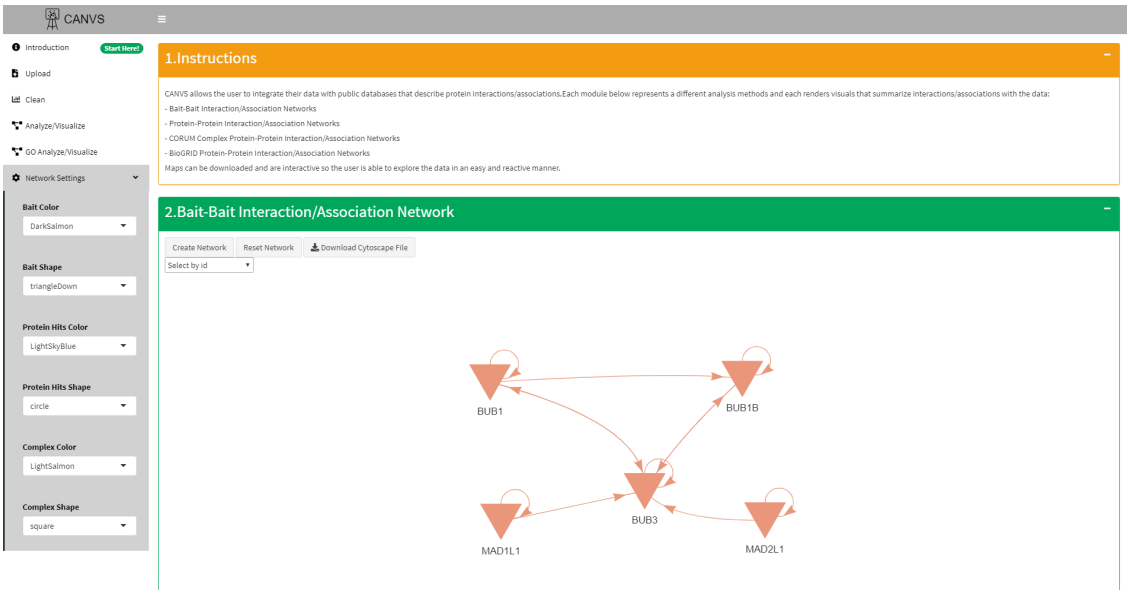


Figure S7: Change Color and Shape in Networks

Users can create networks corresponding to the network type of that box by pressing the + button and expanding each box. For example, a bait-bait network diagram is shown below after expanding the first box.

All networks are customizable (in terms of color and shape of nodes) by clicking the Network Settings link in the sidebar. For example, the color and shape of the nodes in the previous network are changed from gold to salmon and from star to down triangle by changing the color and shape of baits in the sidebar, resetting the network, and then creating the network again. All networks carry the same logic and settings in terms of color and shape to make the analysis consistent throughout tabs.

To reset a network after applying any change: 1. Click Reset Network 2. Click Create Network

Furthremore, the user has the option of downloading Cytoscape network files that can be used to create network visuals in Cytoscape. To do so press the Download Cytoscape File button in each network and then upload the file to Cytoscape as a network.

### Filter by GO Terms

CANVS also allows the user to filter protein results by GO terms associated with the proteins in the results. To do so, first search GO Terms by a keyword which will be used to search for any GO terms associated with the results that have that keyword. The terms that have the keyword and the subterms, or child terms in Gene Ontology, will be selected. These will be summarized in a table and the association of how the terms might be related to each other can also be analyzed in a network. The network will color code GO terms based on whether they are biological processes, molecular functions, or cellular components (the three categories Gene Ontology uses to describe GO terms).

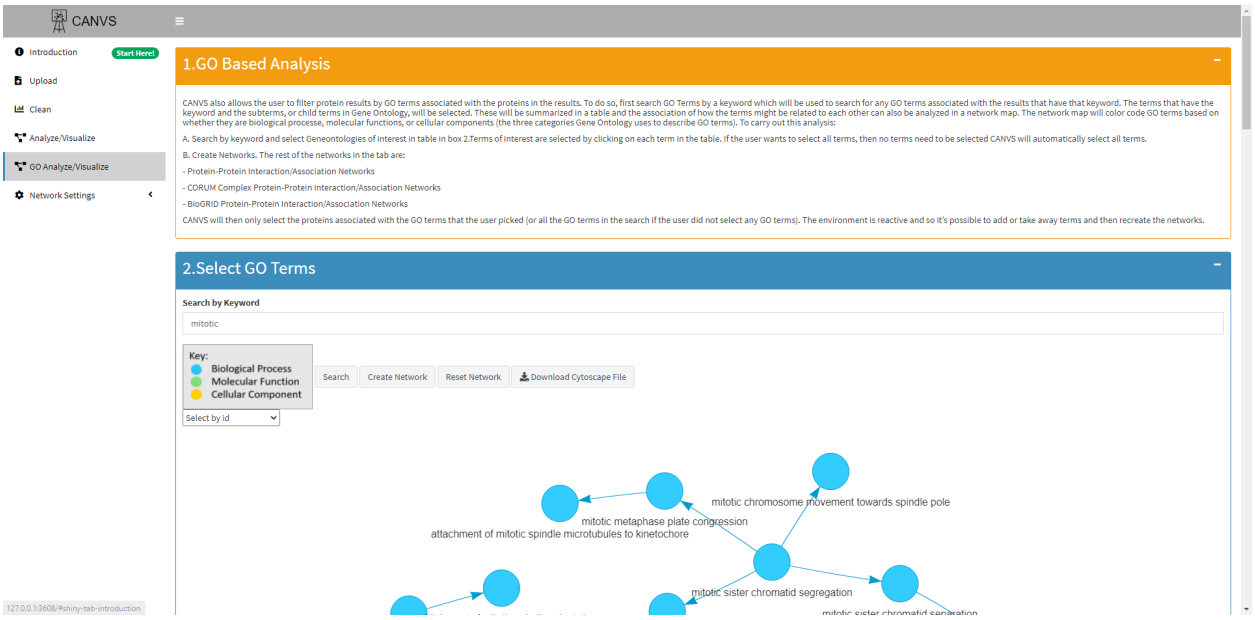


Figure S8: GO Analyze/Visualize Tab

Terms of interest are selected by clicking on each term in the table. If the user wants to select all terms, then no terms need to be selected CANVS will automatically select all terms. The rest of the networks in the tab are:

- Protein-Protein Interaction/Association Networks
- CORUM Complex Protein-Protein Interaction/Association Networks
- BioGRID Protein-Protein Interaction/Association Networks

CANVS will then only select the proteins associated with the GO terms that the user picked (or all the GO terms in the search if the user did not select any GO terms). The environment is reactive and so it's possible to add or take away terms and then recreate the networks.

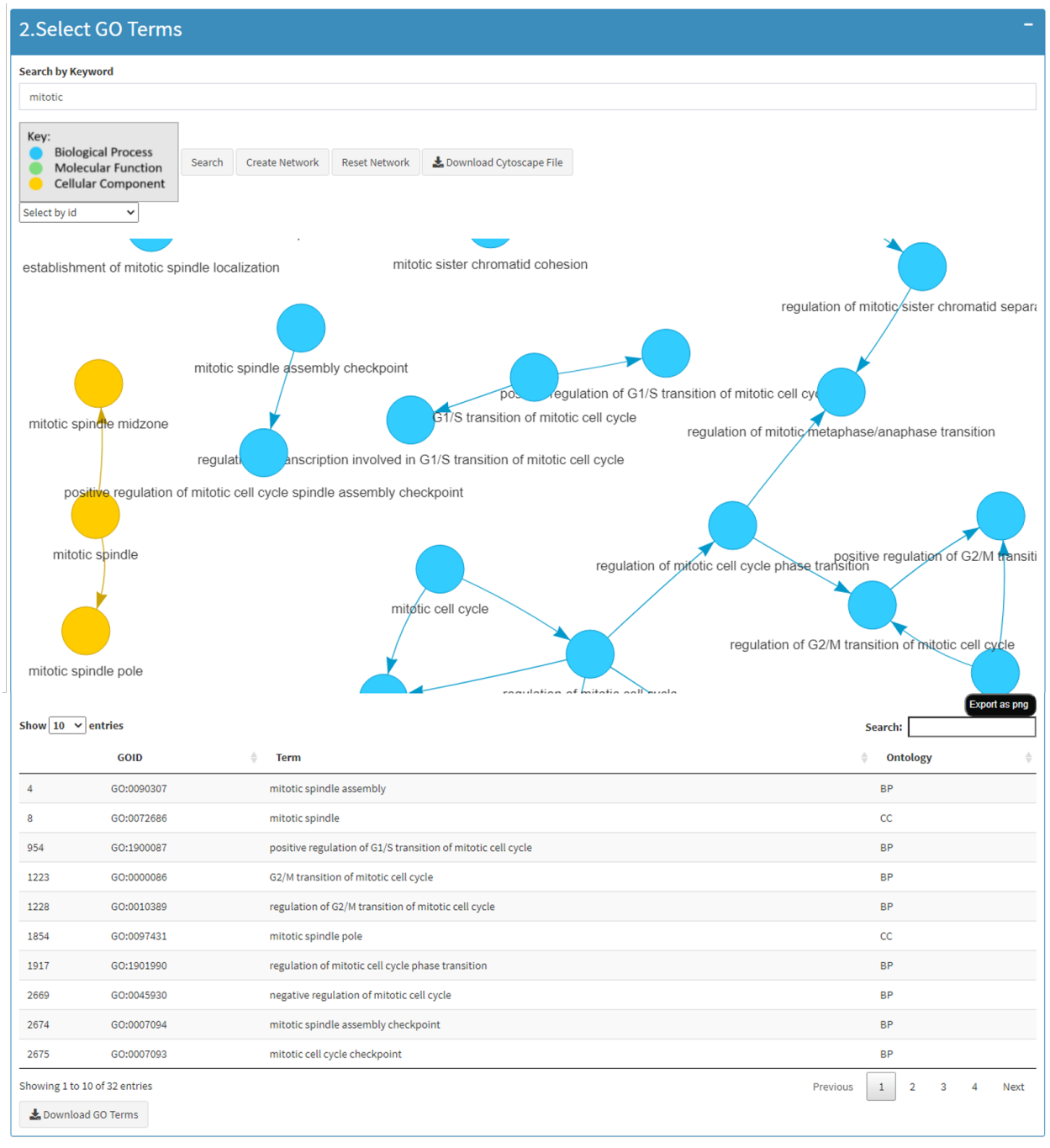


Figure S9: Select GO Terms

**CORUM Networks**

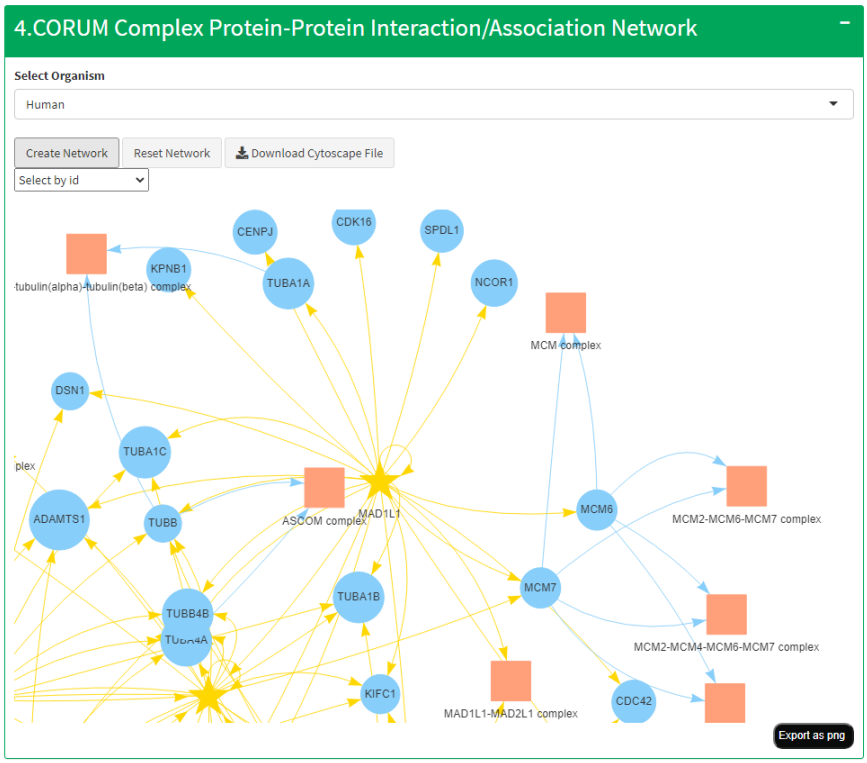


Figure S10: CORUM Example

The user can create CORUM (CORUM v. 3.0) based networks with proteins from the general results or proteins pertaining to particular GO Terms. To expand the CORUM Network box, select an organism of interest and click create network. This will render a network with PPI/As and their respective complexes under the organism chosen. To create an informative network, only complexes that are shared by two or more proteins are considered when making the network. Currently CORUM supports the following organisms:

- Human
- Mouse
- Pig
- Bovine
- Rat
- Mammalia
- Rabbit
- Dog
- Hamster
- Mink

The user has the option of changing the original search terms or updating the results, which results in a different network that the user can render by first clicking reset network and then creating the network again. Below is an example:

## BioGRID Networks

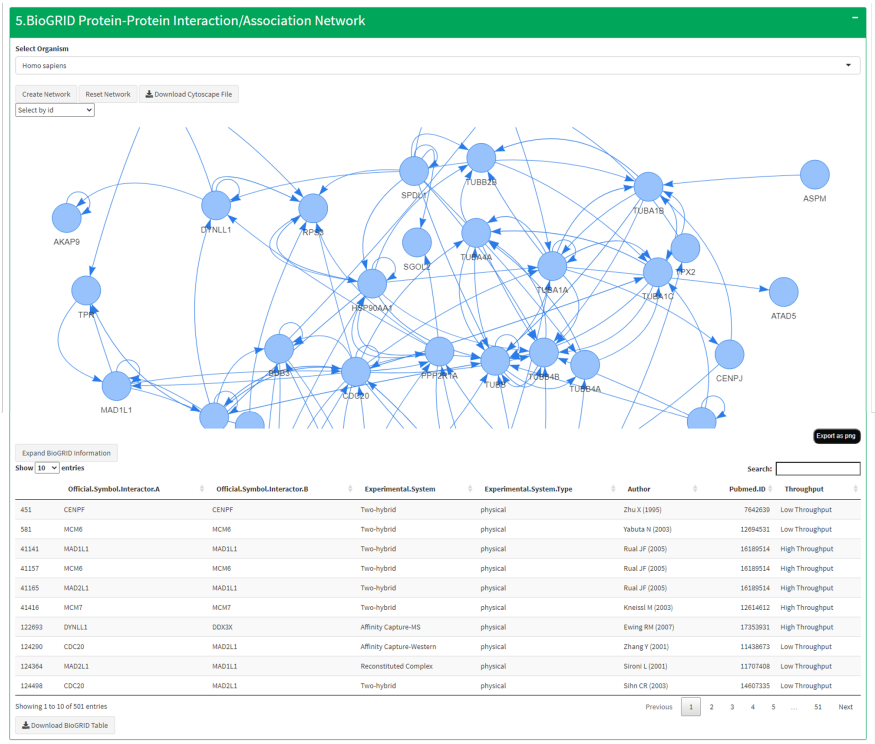


Figure S11: BioGRID Example

To contextualize the results of PPI/A data with previously known interaction data, CANVS integrates BioGRID (BioGRID V. 3.5). To create these networks, click on the BioGRID network box and search for an organism of interest. Click on create network which will render a network showing how the results are related in the context of BioGRID interactions. The user also has the option to expand BioGRID information and create a datatable that summarizes each interacion and offers the user information about it's nature and where it was initially reported. This datatable can also be downloaded. BioGRID currently supports the following organisms:

- Homo sapiens
- Drosophila melanogaster
- Caenorhabditis elegans
- Saccharomyces cerevisiae S288C
- Schizosaccharomyces pombe 972h-
- Mus musculus
- Rattus norvegicus
- Canis lupus familiaris
- Arabidopsis thaliana
- Bos taurus
- Gallus gallus
- Escherichia coli BW2952
- Escherichia coli str. K-12 substr. MG1655
- Neurospora crassa OR74A
- Dictyostelium discoideum AX4
- Human papillomavirus type 16
- Escherichia coli str. K-12 substr. W3110

- Ustilago maydis 521
- Candida albicans SC5314
- Aspergillus nidulans FGSC A4
- Bacillus subtilis subsp. subtilis str. 168
- Anopheles gambiae str. PEST
- Pediculus humanus corporis
- Selaginella moellendorffii
- Escherichia coli K-12
- Mycobacterium tuberculosis H37Rv
- Chlorocebus sabaeus
- Oryza sativa Japonica Group
- Human gammaherpesvirus 8
- Plasmodium falciparum 3D7
- Human betaherpesvirus 6B
- Human betaherpesvirus 6A
- Vitis vinifera
- Tobacco mosaic virus
- Simian immunodeficiency virus
- Human immunodeficiency virus 2
- Human immunodeficiency virus 1
- Hepacivirus C
- Simian virus 40
- Human gammaherpesvirus 4
- Human betaherpesvirus 5
- Human alphaherpesvirus 3
- Human alphaherpesvirus 2
- Human alphaherpesvirus 1
- Vaccinia virus
- Cavia porcellus
- Cricetulus griseus
- Oryctolagus cuniculus
- Sus scrofa
- Equus caballus
- Pan troglodytes
- Macaca mulatta
- Meleagris gallopavo
- Xenopus laevis
- Danio rerio
- Strongylocentrotus purpuratus
- Apis mellifera
- Zea mays
- Solanum tuberosum
- Nicotiana tomentosiformis
- Solanum lycopersicum
- Ricinus communis
- Glycine max
- Chlamydomonas reinhardtii

The user has the option of changing the original search terms or updating the results which would result in a different network which the user can render by first clicking reset network and then creating the network again. Below is an example of the box:

## 4. Conclusion

CANVS is meant to be used as an interactive platform to explore how proteins from PPI/A experiments might be associated with each other. Proper experimental validation of PPI/As is then required. Furthermore, as proteoinformatic databases expand to include data from more organisms, CANVS can be extended to include those organisms. Nonetheless, CANVS empowers any scientist to carry out a simple straight-forward bioinformatic analysis of PPI/A data without having the technical/coding skills to do so.