Supplementary Methods

Additional information and data included in this manuscript

The **Supplementary Protocol** contains a quick start guide to familiarize users with site navigation as well as detailed instructions for each of the six tools, and **Supplementary Table 1** contains sample BioID data[1] for testing with our tools.

File input for the analytical tools

The file input system supports any file in tabular format. The file itself must contain four columns at a minimum that specify the bait, prey, abundance of the prey (spectral count, intensity, etc.) and the confidence score. The abundance measure can be any sort of metric providing it is a non-negative number. Any confidence score can be used, but the user will have to specify how it works if the score is not automatically recognized by ProHitz-viz (i.e. is a higher score better, or *vice versa*). Suggested columns will be selected by default if the input file has been generated by supported tools, including SAINT[2-4], SAINTexpress[5], SAINT-MS1[6] and the CRAPome[7]. Alternatively, for unsupported input formats, the user simply needs to select the desired columns from dropdowns generated after file upload (**Fig. 1a**, **Supplementary Fig. 1** and **Supplementary Protocol**). As all ProHits-viz tools perform comparisons, a minimum of two baits is required in the input file. Please note that several input files in the same format can be simultaneously loaded into ProHitz-viz (e.g. for analyzing datasets jointly); all that is required is that these files are located within the same folder and have the same format.

Output from the analytical tools

All tools will generate a downloadable folder that contains images in editable PDF format, text files with the results of the analysis, input files for outside tools such as Cytoscape[8] and Treeview[9] (when relevant), and a log file that contains all selected parameters for analysis. In addition, files (ending in _df.tsv) are generated that can be used as inputs for the interactive visualization tools.

Dot plot analytical tool

Two “score” filters must be supplied for generating dot plots. The “primary filter” is used for selecting preys to display on the dot plot: once a prey passes this primary filter with at least one bait, quantitative values for the prey across all baits will be used for visualization, even if they did not pass the score cutoff in every particular bait-prey pair. The “secondary filter” is used purely for display purposes on the output image, i.e. if a prey has not passed the primary filter but has passed the secondary, its node edge will be colored accordingly. Data generated through SAINTexpress (the system currently employed by most of our users) will have these filters set to 1% and 5% by default, respectively. Additional parameters can also be customized. A “minimum abundance” value can be set to restrict the preys that will be included in the dot plot (preys detected with a lower abundance across all baits will not be visualized, irrespective of their score; once a prey passes the abundance filter with at least one bait, all its quantitative values will be displayed provided it also passes the primary filter). A “maximum abundance” value can also be set so that any preys with an abundance value equal to or greater than the selected value will have their color capped on the output image. Additional options include the ability to subtract prey spectral counts found in controls from bait counts prior to any other calculations (performed by default if control values are available), log transform spectral counts (different bases are available) and perform normalization using total spectral counts/abundance or using a specific prey for normalization (an affinity tag for example). Hierarchical clustering is performed using R. Bait and prey distances are measured using the Canberra distance as default (binary, Euclidean, Manhattan, maximum and Minkowski are also supported and accessed through a drop-down menu from the “Advanced” options) and clustering is performed using Ward’s linkage method as default (average, centroid, complete, McQuitty, median and single are also available). Biclustering is available using the nested clustering package we have previously described[10]. It is also possible to turn clustering off: baits and preys will then be ordered as specified in the corresponding input boxes. Detailed help for these parameters can be found directly from help links on the input page of the dot plot tool at prohits-viz.lunenfeld.ca.

Correlation tool

The correlation tool is best suited for medium to large datasets. Correlation and clustering of correlated variables is performed using R with the Pearson method set as default (Kendall and Spearman are also offered). After correlation scores are calculated, the distance between variables is measured and clustering performed, using the Euclidean distance and the complete linkage method as defaults respectively (Canberra, binary, Manhattan, maximum and Minkowski are also offered as distance metrics, and average, centroid, McQuitty, median, single and Ward's for the linkage method). Filters can be set at the prey level to restrict the observations used for both bait and prey correlation. For example, SAINTexpress analyzed data by default will set a requirement of preys to pass a 1% FDR and have a spectral count of at least 20 for at least one bait to be included in analysis. As for dot plots, control subtraction, log transformation and normalization options are available. Additional options include the ability to simulate bait spectral counts (which are filtered out by some analysis tools, and can here be imputed as the value of the highest abundance prey), use replicate information if available (default) and ignore source genes when calculating correlation (i.e., should spectral counts/abundance for genes X and Y be ignored when determining the correlation between X and Y).

Specificity tool

The specificity tool is useful to determine which of the preys detected with a bait are specifically enriched with this bait in relation to the other baits in the dataset. Prey specificities can be calculated using several metrics. A simple fold change measurement we offer reflects the amount of prey found with a bait relative to the average amount found across all other baits:

$$s_{i,j} = (N-1) \cdot \frac{x_{i,j}}{\sum_{k=1,k\neq i}^{N} x_{k,j}} \qquad (1)$$

where $x_{i,j}$ is the spectral count/abundance of prey j for bait i and N the number of baits. This measure can be coarse for low abundance preys or sparse datasets. The other specificity scores available in ProHits-viz are taken directly from CompPASS[11] and include Z-Score, S-score, D-score and WD-score. As above, users can choose to perform control subtraction, normalization and log transformation of prey data. They can also adjust the spectral count/abundance of a prey to its length. This is done by taking the median length of all significant preys and normalizing to that, so a prey with a length half the median will have its spectral counts doubled. This may be useful to highlight relatively small baits with comparatively high abundance, although this change does not affect the specificity score. RNA-seq data from The Human Protein Atlas[12] (www.proteinatlas.org) can be mapped onto the node border. If this option is selected, then genes with a read count of 50 transcripts per kilobase million (TPM) or greater will be shown with a full 360° edge, indicating high expression. This cutoff can be adjusted as desired. Genes with an RNA-seq value less than the specified cap will be shown with an edge length relative to that, *i.e.*

$$border\ length\ = \frac{read\ count}{read\ count\ cap} \times 360° \qquad (2).$$

RNA-seq data is available for 56 different cell lines that can be selected through a dropdown menu.

Bait-bait comparison tool

Oftentimes, a user may want to compare the recovery of individual preys across only two baits (e.g. the same bait gene following some perturbation, or a mutant variant of the same bait gene). Bait comparisons can be plotted showing either the prey fold-change for one bait relative to a reference bait on the y-axis, with the prey abundance for the reference bait on the x-axis, or as what we term a “versus” plot (default) with the prey abundance of each bait shown on its own axis. Filters can be applied to control the preys displayed on the output image, and as above, control subtraction and normalization of data can be performed. Log transformation of data can be done for versus plots but not fold-change plots as the axis for the latter are already log scaled by default. To indicate which preys have passed the selected score cutoff for both baits on the versus plot, the edge attribute is used: a full circle indicates that the cutoff was met in both cases, while a 180° circle indicates that the selected cutoff was met with only one bait.

Interactive viewers

The analytical tools described above are complemented by two interactive viewers: one for dot plots and heat maps generated by the dot plot and correlation tools, and one for the scatter plots generated by the specificity and bait-bait comparison tools. These can open the results from the analytical tools directly, or through re-upload of the _df.tsv file generated by the tools. Images can be output in SVG and/or PNG format. Note that interactive images can be archived, and re-accessed at a later time using a web link. Image interactivity is created using a mixture of D3[13] (d3js.org) and custom JavaScript.

The interactive viewers are complemented by additional analytical tools accessed dynamically through API. GO analysis of genes is performed using g:Profiler[14], mirroring their interface. Domain analysis is done by first mapping gene names to UniProt IDs (www.uniprot.org) and then retrieving domain information from Pfam[15] (pfam.xfam.org). Networks are created using reported protein-protein interactions from BioGRID[16] (thebiogrid.org) mapped using a force-directed layout.

Data set analysis

BioID data from[1, 17] was run through each of the tools using default settings. RNA-seq data from HEK 293 cells was used for node borders on the specificity plot.

Code availability

Server side code that performs underlying data analysis is available by contacting the authors. ProHits-viz and all its tools will always be freely accessible and we intend to maintain it for the foreseeable future.

Data availability

Data used for generating images is available at prohits-viz.lunenfeld.ca/data.

References

1. Couzens, A.L. et al. Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. *Sci Signal* **6**, rs15 (2013).
2. Breitkreutz, A. et al. A global protein kinase and phosphatase interaction network in yeast. *Science* **328**, 1043-1046 (2010).
3. Choi, H. et al. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat Methods* **8**, 70-73 (2011).
4. Choi, H. et al. Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT. *Curr Protoc Bioinformatics* **Chapter 8**, Unit8 15 (2012).
5. Teo, G. et al. SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *J Proteomics* **100**, 37-43 (2014).
6. Choi, H., Glatter, T., Gstaiger, M. & Nesvizhskii, A.I. SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J Proteome Res* **11**, 2619-2624 (2012).
7. Mellacheruvu, D. et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* **10**, 730-736 (2013).
8. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).
9. Saldanha, A.J. Java Treeview--extensible visualization of microarray data. *Bioinformatics* **20**, 3246-3248 (2004).
10. Choi, H., Kim, S., Gingras, A.C. & Nesvizhskii, A.I. Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data. *Mol Syst Biol* **6**, 385 (2010).
11. Sowa, M.E., Bennett, E.J., Gygi, S.P. & Harper, J.W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389-403 (2009).
12. Uhlen, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
13. Bostock, M., Ogievetsky, V. & Heer, J. D(3): Data-Driven Documents. *IEEE Trans Vis Comput Graph* **17**, 2301-2309 (2011).
14. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**, W193-200 (2007).
15. Finn, R.D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-285 (2016).
16. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-539 (2006).
17. Gupta, G.D. et al. A Dynamic Protein Interaction Landscape of the Human Centrosome-Cilium Interface. *Cell* **163**, 1484-1499 (2015).