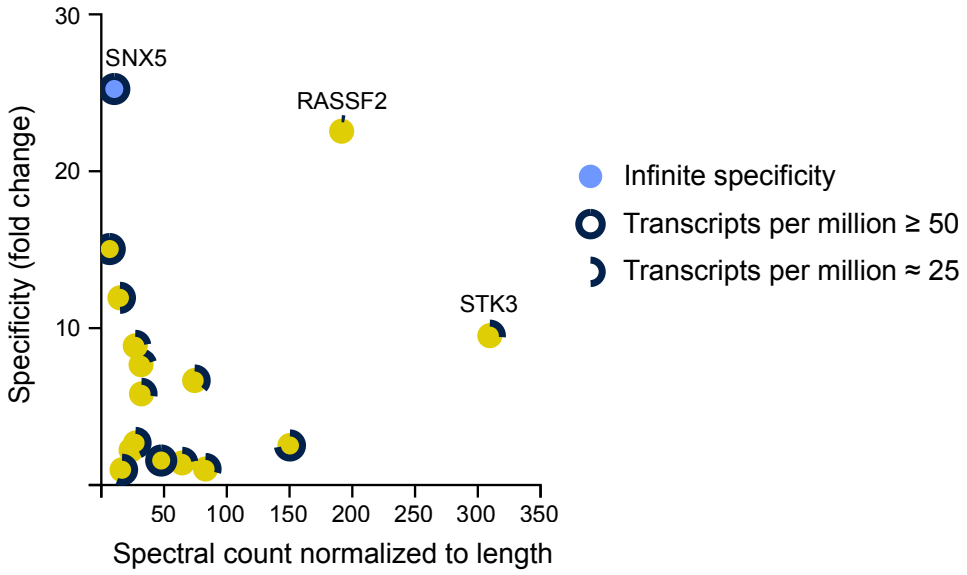
ProHits-viz

Prey specificity

This tool takes a file with quantitative information on bait-prey interactions and calculates prey specificity scores for each bait in the dataset relative to all baits in the dataset. An example output is shown below where specificity scores for each prey are plotted relative to their abundance (in this case spectral count). Only preys that pass the specified filter will be plotted, while blue points indicate preys with infinity specificity.



The number of baits in the input dataset and their prey-profile similarity will heavily influence the specificity score. As such, specificity scores should only be interpreted relative to the input data.

Static images, a file compatible with our interactive viewer and a file containing the specificity scores will be available upon completion of the task. Alternatively, the interactively images can be directly viewed by clicking on the appropriate link.

Help is available directly on the page by selecting the "help" links next to the section headers. Any additional questions jknight@lunenfeld.ca. For more information, see PMID: 25422071.

Inputting data

Specify the file for upload and the file type, then hit "Process". Explicit support is provided for data files output from SAINT and CRAPome (CRAPome matrix format is not supported, however). Datasets from other tools or pipelines can be input by selecting the "Other"

option. Files must be in tabular format as tab-delimited text. At a minimum, the file must contain four columns specifying the bait/condition, prey/readout, abundance measure (spectral count, intensity, etc) and a confidence metric (e.g. FDR). Sample input files are available for download by clicking the "help" link and contain BioID published in PMID: 24255178.
For detailed information on tools that generate compatible input for ProHits-viz, see the references and links below.

References: SAINT - PMID:21131968, SAINTexpress - PMID:24513533, SAINT-MS1 - PMID: 22352807, ProHits - PMID:20944583, ProHits Protocol - PMID:22948730, ProHits 4.0 - PMID:27132685, CRAPome - PMID:23921808

Links: SAINT-APMS.sourceforge.net, ProHits.com, CRAPome.org, GalaxyP

Data columns

Up to five columns will be used when processing the input file: a column listing the 1) baits, 2) preys, 3) abundance measure (spectral count, intensity or simply a non-negative number), 4) score for filtering (for unknown filter types, the direction of the filter must be specified, i.e. is a smaller number better, or vice versa), and 5) control values for performing control subtraction, i.e. subtracting the average control value from the observed prey abundance (this is optional and can either be a pipe-separated list of values or a single averaged number).

Parameters

1. Specificity metric: There are several options for the specificity metric. The first is a simple fold enrichment score calculated for each prey and the bait it was detected with, relative to the entire dataset:

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

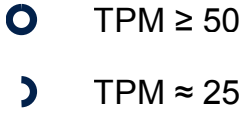
where $x_{i,j}$ is the spectral count for prey j relative to bait i and N is the number of baits.
The other scores are implemented as described by the Comparative Proteomic Analysis Software Suite (CompPASS). We would refer the user to the tutorial page for detailed descriptions.

CompPASS tutorial

Z-score: a prey's Z-score indicates the number of standard deviations away it is from the mean.

S-score: the S-score reflects the abundance of a prey adjusted by the frequency with which it is found across baits (lower frequency = higher score). Unlike the fold-enrichment and Z-scores, prey abundance will affect comparisons between preys, for example if two preys are equally frequent, the one with the higher abundance will receive a higher score.

Prey specificity



D-score: the D-score is calculated in the same was as the S-score, except reproducibility is incorporated into it, i.e. a reproducibly found prey will score higher than one that isn't. This score should only be selected when abundance information is available for two or more replicates. This abundance column must contain the replicate values as a pipe-separated list. See the "Spec" column from the example SAINT file to see how this should be formatted.

WD-score: the WD-score is a weighted D-score, that attempts to adjust the D-score to better recover/score frequently found proteins that show behavior typical of true interactors. Like the D- and S-scores, prey abundance affects comparisons between preys.

2. Score filter: All preys that satisfy this score cutoff will be displayed in the scatter plot. Note: specificity scores will be calculated for all preys; this cutoff is only for display purposes.

3. Points to label (default 10): The number of points to label on the plot beginning with the highest specificity score and moving downwards.

4. Control subtraction: The average value of a prey across control samples will be subtracted from the detected value for the bait if this is set to "yes". The quantitative value for the prey becomes the value above and beyond what is seen across the control samples. Specify the column to use for controls in the adjacent "Control column" field.

5. Adjust abundance to protein length (no by default): The spectral count/abundance value of each prey can be normalized to its protein length if a column with protein length is available in the input file. This normalization will not affect specificity scores. It can be used to adjust the x (abundance) dimension on the output scatterplots so that it is weighted relative to protein length. The multiplcation factor used to normalize a prey's abundance is calculated as the median of the length of all significant preys (those passing the cutoff) divided by the prey's length.

6. Normalization between samples (none by default): No normalization across baits is applied by default, but when baits in the same dataset have been run on instruments with varying sensitivity or dynamic range, normalization should be applied. The options for normalization are to divide by the total abundance for all proteins identified in the run or normalize based on a specific prey. Normalization will be applied after control subtraction if both are specified.

7. Log transform (default no): If desired, data can be log-transformed by base 2, base 10 or natural. Log transformation will be performed after control subtraction and/or normalization if these are also specified.

8. Mark expression level on node (default no): The RNA expression level of a gene can be drawn on nodes by selecting this option. You must specificy a cap for high expression in transcripts per million (TPM, default 50) and specify the cell line. Expression information is taken from The Human Protein Atlas. Expression level will be indicated on a node as the edge length. Nodes with expression ≥ the specified cap will have a edge length equal to the

complete node circumference, while nodes with levels of expression less than that will be shown with an edge length proportional to their expression divided by the cap, as shown in this example:
9. Remove contaminants (default no): If you wish to omit plotting of preys considered to be contaminants (or for other reasons), you can select this box and specify a list of them in the text area to the right. Gene names must be entered one per line and are case sensitive.

Output

After the data has been processed, the user will be prompted to download a PreySpecificity_results.zip file that contains the results in a folder. By default the unzipped folder will be named "RESULTS" but this name can be modified as desired. There are four files in this folder.

1. specificity_plots.pdf: This file contains scatterplots displaying prey specificity vs spectral count for all baits in the input data set. Blue points on the scatterplot indicate preys with infinite specificity score. The file can be opened and edited in Adobe Illustrator or a similar program. In some cases the following warning may appear on opening the file in Illustrator: "The font AdobePiStd is missing. Affected text will be displayed using a substitute font." If this occurs, the image will not display correctly.

To fix this issue on a Mac, copy the file AdobePiStd.otf from /Library/Application Support/Adobe/PDFL/*Current Version*/Fonts/ and transfer it to the folder / Library/Fonts/. The *Current Version* folder refers to your version of Adobe. On Windows, the font file is located in C:\Program Files\Common Files\Adobe\PDFL *Current Version*\Fonts\ and needs to be placed in C:\Program Files\Adobe \Adobe Illustrator CS5\Support Files\Required\Fonts\. If the AdobePiStd.otf file is missing, it can be freely downloaded from a number of sites on the web in either Mac or Windows format.

2. specificity.txt: This is the original file that was input to the tool with an additional column appended to it containing the specificity score and expression information if that option was selected. This file can be directly opened with tools such as Cytoscape for network visualization of your dataset.

3. specificity_df.tsv: Plain text file formatted for our interactive scatter plot viewer found at ProHits-viz that allows navigation of images. This file should be loaded directly into that tool without modification.

4. process.log: This file contains information on the input parameters that were selected for the user's future reference (e.g. to assist with writing the Methods section for a manuscript).

Troubleshooting

Problems generally result from errors in the input file format, and we encourage users to compare their input files against the samples provided on the web page.

Any other issues should be sent to jknight@lunenfeld.ca.