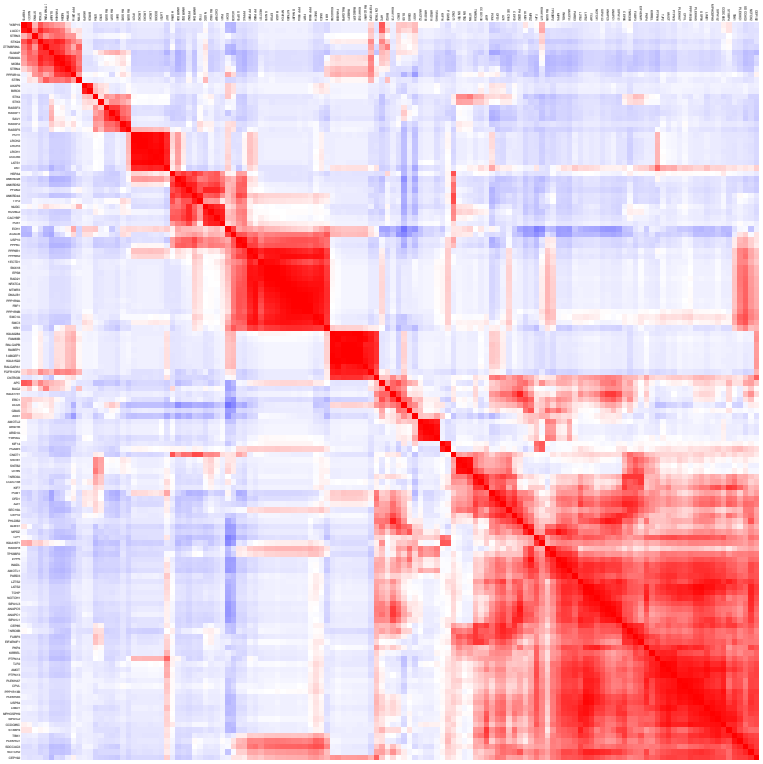
ProHits-viz

Correlation Tool

The correlation tool takes a file with quantitative information on bait-prey interactions, performs both prey-prey and bait-bait correlation analysis, and visualizes the results as clustered heat maps. The particular value of this tool is the prey-prey correlation it performs, which identifies preys that are seen in a correlated manner between baits, indicating they may co-localize and/or be part of a complex. This offers an alternative prey-centric way of looking at data, as opposed to more traditional analysis performed by our dot plot tool for example. An example image is shown below.



We also generate Cytoscape compatible files for the correlation matrices as network diagrams offer another alternative way of visualizing the relationships shown in these heat maps.

A bait-prey heat map is output from this tool that shows preys along the x-axis in the same order as they appear in the prey-prey heat map, while baits are clustered along the y-axis based on their prey-abundance profiles. The purpose of this image is to assist in showing the baits that drive the prey-prey correlation. This image is only available in a static form via the downloadable "Results" folder as it has little value in its own right and should always be viewed alongside the prey-prey correlation image.

Static images, files compatible with our interactive viewer and Cytoscape formatted files for the bait-bait and prey-prey correlation matrices will be available upon completion of the task. Alternatively, interactively images can be directly viewed by clicking on the appropriate link.

Correlation tool

Help is available directly on the page by selecting the "help" links next to the section headers. Any additional questions jknight@lunenfeld.ca.

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. Score filter for prey correlation: Only preys that pass this cutoff will be used for prey-prey correlation analysis. If a prey satisfies this cutoff for at least one bait, all quantitative values for this prey will be used across all baits, even those that did not satisfy the cutoff in particular bait-prey pairs.

2. Abundance cutoff for prey correlation (default 20): In addition to the above score filter requirement, a prey must have an abundance value above this parameter for at least one bait to be included in prey-prey correlation analysis. A high value helps to ensure that a prey is a suitably abundant protein easily detectable by the protocol and mass spectrometer used. This is prevents low abundance proteins from skewing the analysis. This parameter should be set to a lower value for small datasets.

3. Minimum bait requirement (default 1): Preys must pass the above filter and cutoff for this number of baits to be used in analysis. A higher cutoff prevents outlier baits from skewing the clustering.

4. Score filter for bait correlation: Only preys that pass this cutoff will be used for bait-bait correlation analysis. If a prey satisfies this cutoff for at least one bait, all quantitative values for this prey will be used across all baits, even those that did not satisfy the cutoff in particular bait-prey pairs. This should be set to a less stringent value than the cutoff for prey correlation to give more data points for correlation. As the number of baits and detected preys in the dataset increases, this parameter can be made more stringent, although this will be dataset dependent.

5. Abundance cutoff for bait correlation (default 20): In addition to the above filter requirement, a prey must have an abundance value above this parameter for at least one bait to be included in bait-bait correlation analysis.

6. Add back preys (default yes): If a prey passes the filter and minimum bait requirements described above then all data points for this prey across all baits will be included when performing the correlation.

7. Add bait counts (default no): SAINT automatically filters out peptides detected for the bait protein but this value can be simulated if this parameter is set to "yes", in which case, the highest detected prey value satisfying the prey filter cutoff will be used for the bait. The absence of a bait gene can negatively affect correlation analysis for smaller clusters making it necessary in cases to simulate this value when not available.

8. Ignore source genes in correlation: (default no): By default all spectral count information is used when calculating the correlation between gene X and Y. However, there are cases when it may be appropriate to ignore spectral counts for the source genes (X and Y) themselves, particular when either gene was used as a bait. Setting this option to "yes" will cause the source gene spectral counts to be ignored during their pairwise correlation calculation.

9. Use replicates (default yes): If multiple enrichments/replicates have been performed for bait proteins, these replicates can be treated as separate entries for each bait when calculating the correlation or averaged to yield a single entry per bait. If replicates are present and this parameter is set to "yes", replicates will be treated as separate entries for correlation. Replicate information must be specified as a pipe-separated list in the abundance column.

10. Control subtraction (default yes): 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.

11. Normalization (none by default): No normalization 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.

12. 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.

13. Correlation cutoff for Cytoscape output (default 0.7): Text files compatible with direct input to Cytoscape PMID:21149340 containing prey-prey and bait-bait pairs will be output from this tool to assist in generating network diagrams based on the prey-prey and bait-bait correlation profiles respectively. This parameter determines which gene pairs will be output, i.e. if a pair of genes has a correlation value above or equal to the default value of 0.7, it will be classified as an "interacting" pair and will be found in this file, and hence will have a connecting edge in Cytoscape. The text files will contain the gene pair and the correlation value.

Correlation and clustering options

• *Correlation*

Correlation tool

Correlation method (default Pearson): Pearson, Kendall and Spearman are the available correlation methods.

• *Clustering*

After correlation, genes will be clustered via the agglomerative hierarchical approach. There are several options for calculating the distance metric and for the clustering criterion. The available distance metrics are: binary, Canberra, Euclidean, Manhattan, maximum and Minkowksi. The available clustering criteria are: average, centroid, complete, McQuitty, median, single and Ward's.

Output

After the data has been processed, the user will be prompted to download a Correlation_results.zip file that contains the results in several subfolders. By default the parent unzipped folder will be named "Results" but this name can be modified as desired. The subfolders are:

1. PDFs: This folder contains the correlation images, one for prey vs prey and one for bait vs bait. A third bait vs prey image is generated to assist in analyzing the prey vs prey image. This image is a heat map showing preys along the x-axis in the same order as they appear in the prey vs prey image, while baits are clustered along the y-axis based on their prey-abundance profiles. The cell values are the prey abundances normalized to 1 across baits. This image helps in determining which and how many baits are driving the prey clusters. All of the PDF files 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. InteractiveFiles: Contains plain text files formatted for our interactive viewer found at ProHits-viz that allows navigation and analysis of images. These files should be loaded directly into that tool without modification. Their extension is _df.txt.

3. CytoscapeFiles: Two plain text files that list prey-prey or bait-bait pairs that passed the correlation cutoff defined above. In simple terms, if a pair of genes passes a strict enough correlation cutoff, they are likely part of a complex or found in a sub-structure. These will be listed as pairs in this file and when imported into Cytoscape (PMID:21149340) will be shown as connected nodes.

4. Treeview: Java TreeView (PMID:15180930) compatible files for the three images generated by this tool.

Finally, the "RESULTS" folder will also contain a process.log file the 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 examples provided on the web page.

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