ProHits-viz

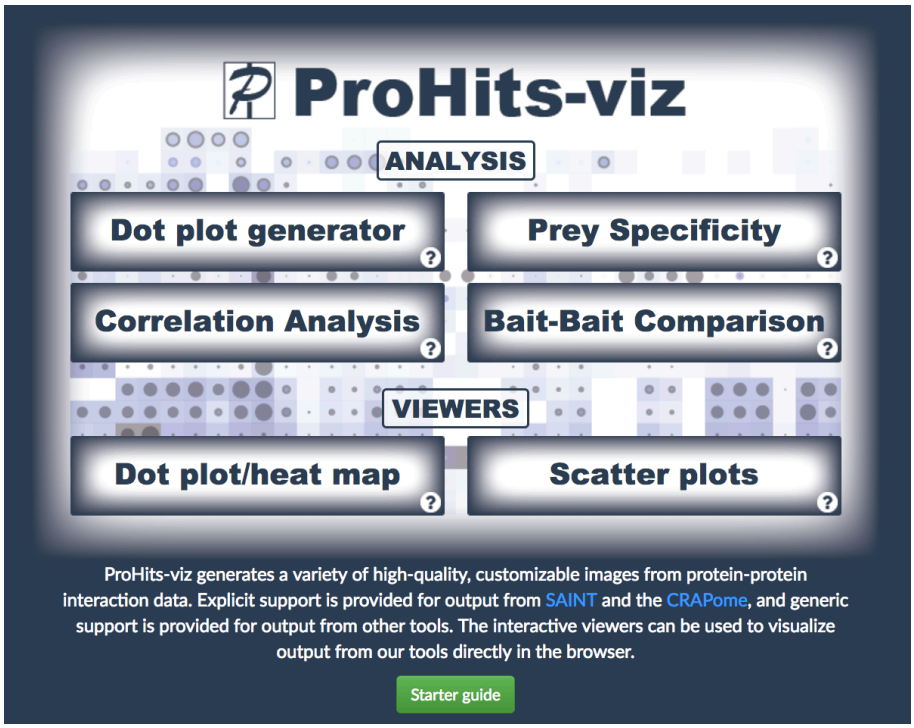
Supplementary Protocol

Table of contents

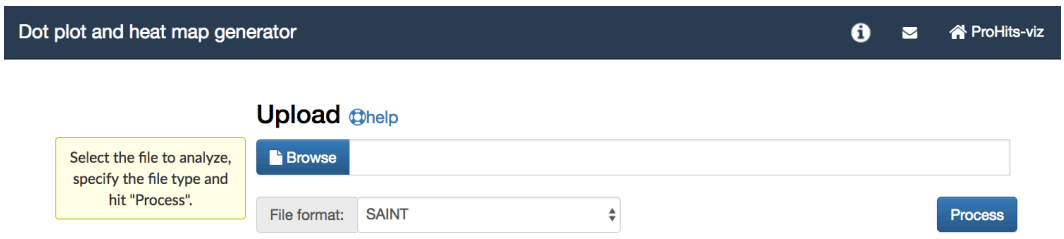
ProHits-viz

Starter guide

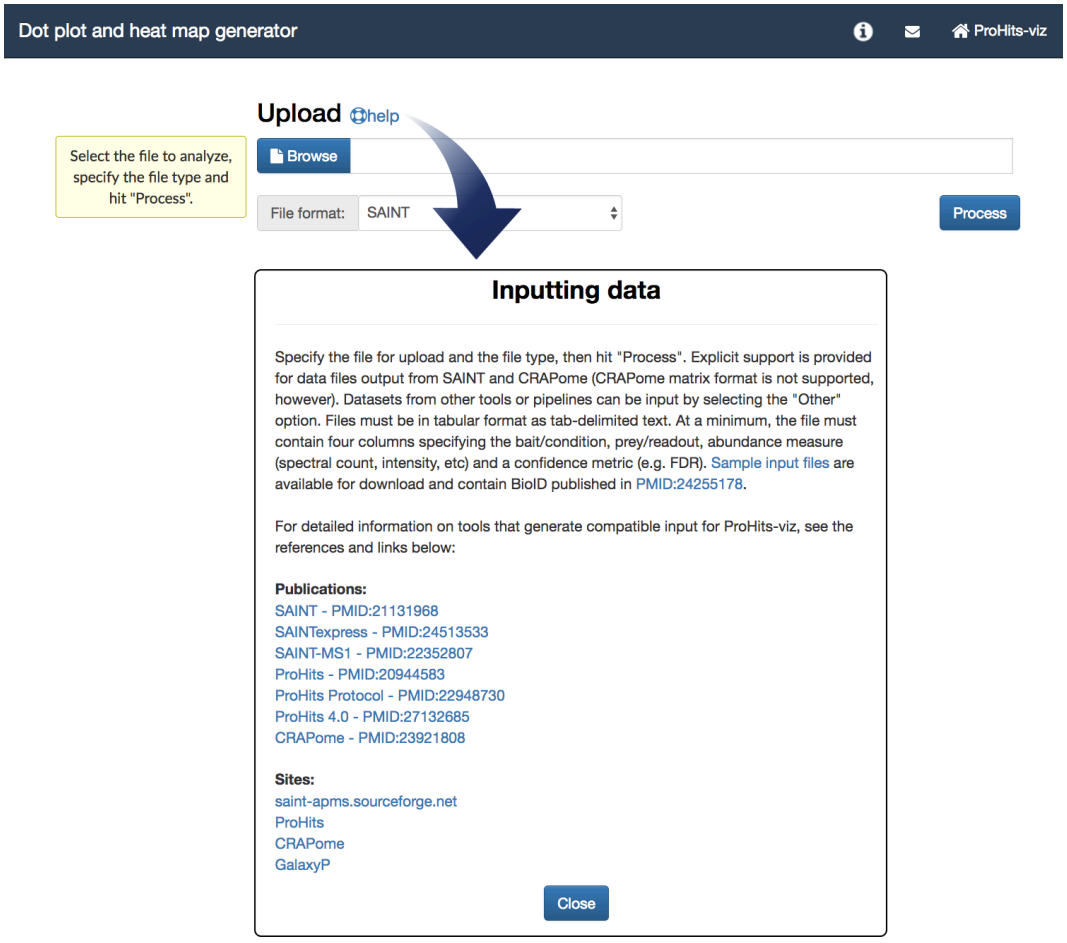
ProHits-viz consists of six tools with intuitive user interfaces for inputting and processing data quickly and easily. Four are designed for processing protein-protein interaction data directly from tools such as SAINT and the CRAPome or more generally from any tabular data. Click the help icon (?) on any tool on the main page for a brief description.



From the homepage of any tool, click the information icon (i) at the top to get a more detailed description of its purpose. The yellow boxes at the left give step-by-step instructions on how to use the tool, with further information available about each step of the process from the help links at the top of each section.



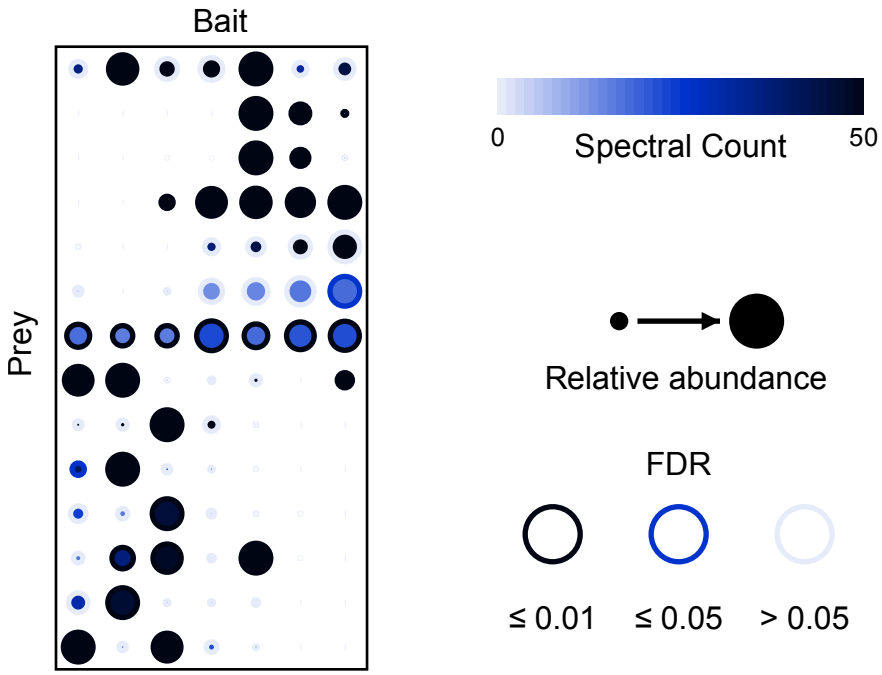
Sample files are available for use with our tools by clicking the help link for the **Upload** section. Click the Sample input files hyperlink to download a test file.



ProHits-viz

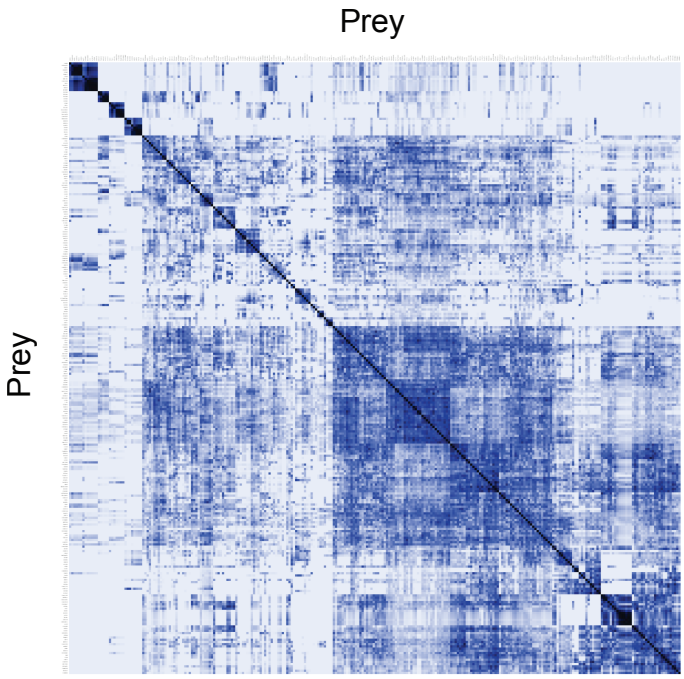
Dot plot Generator

This tool is designed to take a file with quantitative information on bait-prey interactions and visualize those interactions as either a dot plot (shown below) or a heat map. Dot plots have the advantage over heat maps in that they take up the same amount of space but visualize more information. In addition to raw quantitative values being displayed via coloured circles, dot plots display the relative abundance of preys between baits via circle size and confidence in the interaction via coloured edge. Heat maps, however, are better for presenting very large data sets. Regardless, a static dot plot and heat map will be generated for all input datasets.



Distance matrices are a by-product of generating clustered bait vs prey images and we use these distance matrices to create bait vs bait and prey vs prey heat maps as a way of showing similarity between them (as an alternative to comparing columns or rows in the bait vs prey images). We also generate Cytoscape compatible files for these matrices as network diagrams offer another alternative way of visualizing this similarity.

Dot plot generator



Static images, files compatible with our interactive viewer and Cytoscape formatted files for the bait-bait and prey-prey distance matrices will be available upon completion of the task. Alternatively, interactively images can be directly viewed by clicking on the appropriate link. Although a direct link is not provided for an interactive heat map, this can be viewed by selecting the dot plot, and changing the image type to "Heat map" under the display options of the interactive viewer.

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

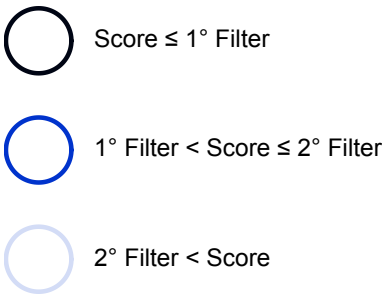
Data columns

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

There are several options available for processing and visualizing the data. Defaults are suggested on the input page but these can be changed as needed.

1. Primary filter: All preys that satisfy this score filter for at least one bait will be displayed in the dot plot. If a prey satisfies this filter for at least one bait, all the quantitative values for this prey – even those that did not satisfy the cutoff in particular bait-prey pairs - will be returned. Prey confidence be indicated by the circle edge as shown below:



2. Secondary filter: Interactions that do no pass the primary score filter but pass this secondary filter will be marked with an intermediate intensity in the dot plot. Interactions that do not pass either filter will be marked with a low intensity edge. The secondary filter can be adjusted depending on the dataset to allow a greater or lesser number of interactions into this “medium” confidence range.

3. Minimum abundance value: In addition to the above filter requirement, a prey must have an abundance value (e.g. spectral count) above this minimum to be included in the analysis. As with the primary filter cutoff, once a prey passes this threshold for one bait, all the values for this prey are returned for all baits, whether or not they pass the threshold for this specific pair.

4. Maximum abundance value: Any preys with an abundance value (e.g. spectral count) above this cutoff will be capped at this maximum in the output images. This is to give greater visual dynamic range for lower values when outlier preys with very high abundance are present. This cutoff will be dependent on the instrument and interaction method used and should be selected based on the data set.

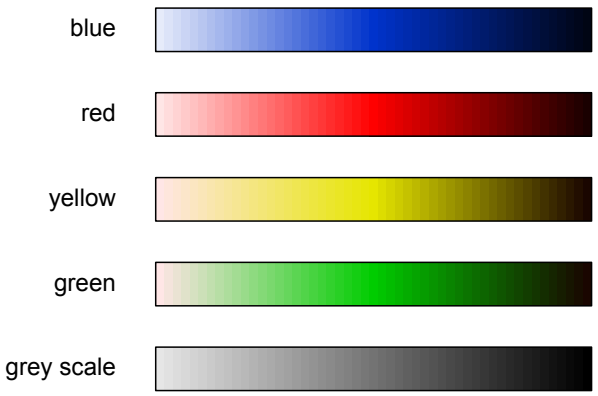
5. Control subtraction: The average value of a prey across control samples will be subtracted from the detected value for the bait if this is set to "yes". The quantitative value

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.

6. 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.

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. Color scale (default blue-black): Various options are available for the fill and edge color, and can be applied independently to each.



Clustering options

1. Agglomerative Hierarchical (Hierarchical): This is the default clustering option and is executed using R. There are several options for calculating the distance metric and for the clustering criterion. Canberra is the default distance metric and Ward's method the default clustering type. From our own experience we have found Canberra to be a very good metric to use for protein-protein interaction data, although other options are available and may produce more desirable results. The defaults clustering type is Ward's, which acts to minimize variance within clusters, although many of the types available will produce comparable results.

> Available distance metrics: binary, Canberra, Euclidean, Manhattan, maximum and Minkowksi
>
> Available clustering criteria: average, centroid, complete, McQuitty, median, single and Ward's

2. Nested Clustering (Biclustering): Another and more sophisticated clustering approach is available within ProHits-viz. This is a probabilistic biclustering approach termed "Nested Clustering", details of which can be found here: PMID: 20571534. This approach first clusters baits based on the similarity of their prey interaction profiles, then pools preys with

similar abundances within these clusters to form a nested cluster. Note that this clustering approach will take significantly longer than the hierarchical clustering option, especially for large data sets. Data sets must have at least 3 baits to use this clustering option.

3. No-clustering (Clustering: none): The user can generate dot plots without clustering if desired. In this case, a list of bait and prey genes in the desired display order must be supplied in the text boxes. Only baits and preys entered in the text box will be included in the dot plot. Bait and prey names must be entered as they appear in the input file name and are case sensitive. In some cases the user may want to control which baits are shown in the dot plot and their ordering, while wanting to show and cluster all preys, or vice versa. This can be done as well by selecting “cluster all” under the bait or prey options menu.

Output

After the data has been processed, the user will be prompted to download a Dotplot_results.zip file that contains the results in three 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 dot plot and heat map images, as well as a legend. The output bait vs bait and prey vs prey images are the clustered distance matrices visualized as heat maps. In contrast to similar images generated by our Correlation Tool, baits (or preys) in these images have not been subjected to correlation analysis prior to clustering. Instead, the metric used is the distance metric calculated using the clustering option specified above. These 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: This folder contains plain text files for importing to Cytoscape. The file 'bait-prey_cytoscape.txt' contains only bait-prey pairs that have passed the primary filter, and with the abundance equal to its post-processing value. For example, if control subtraction and normalization have been selected, the abundance value in this file will be the value after control subtraction and normalization have been performed. This allows users to import only desired bait-prey pairs into Cytoscape with an abundance value suitable for edge-weighting that has beed processed according to the user's need. Two additional files are also present in this folder that contain all bait-bait and prey-prey pairs from the distance matrices with an additional column containing the distance metric itself that can be used for filtering which pairs will be displayed in Cytoscape. These files are generated so that users can view bait-bait and prey-prey information in a network format as an alternative the heat map format we provide here.

4. OtherFiles: This folder contains bait-bait and prey-prey dendrograms and will contain additional files generated by the Nested Clustering option, if selected.

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

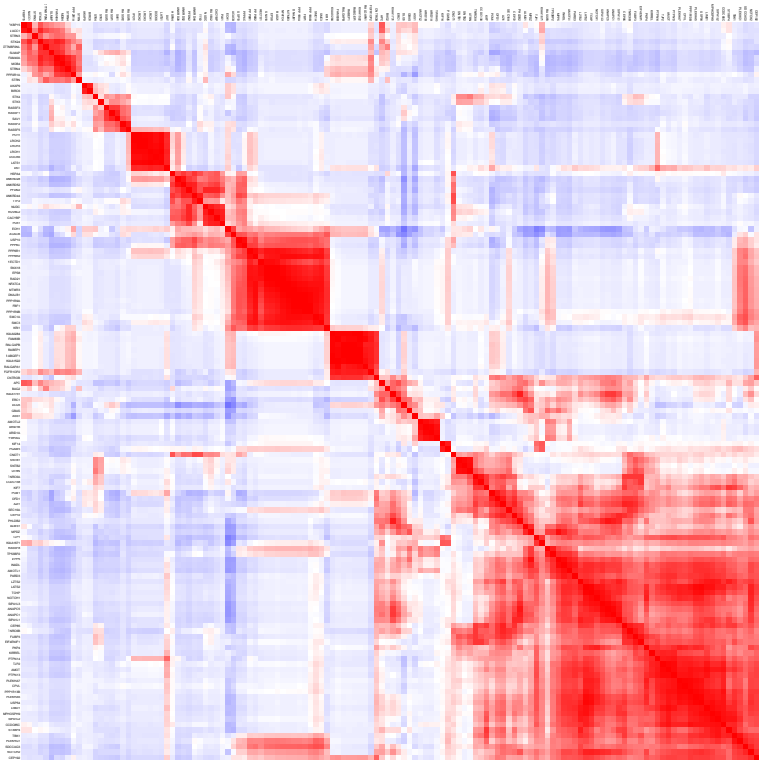
Problems generally result from errors in the input file format, and we encourage users to compare their input files against the samples provided on the web page.

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

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.

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

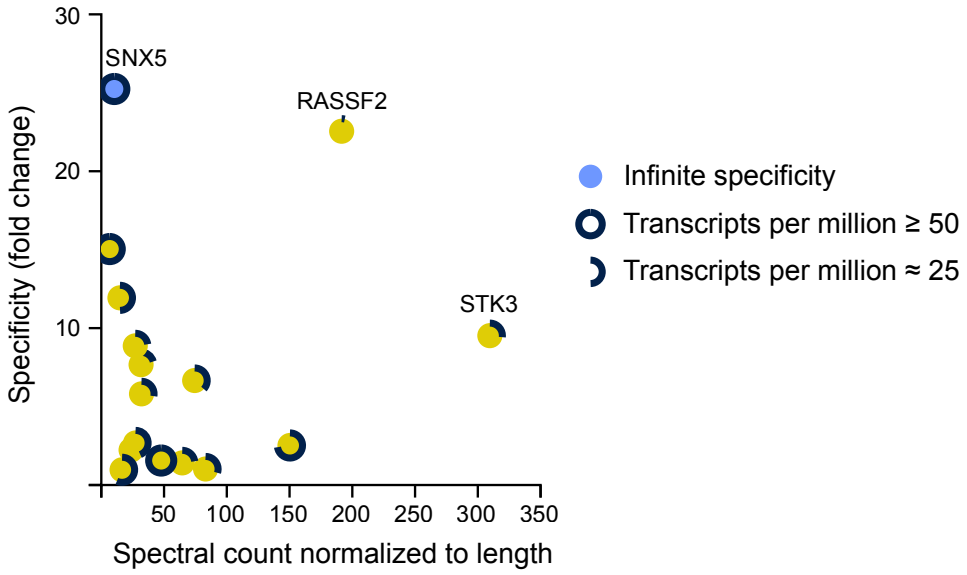
Problems generally result from errors in the input file format, and we encourage users to compare their input files against the examples provided on the web page.

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

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:

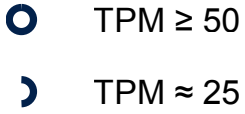
$$s_{i,j} = (N-1) \cdot \frac{x_{i,j}}{\sum_{k=1,k\neq i}^{N} x_{k,j}}$$

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.



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

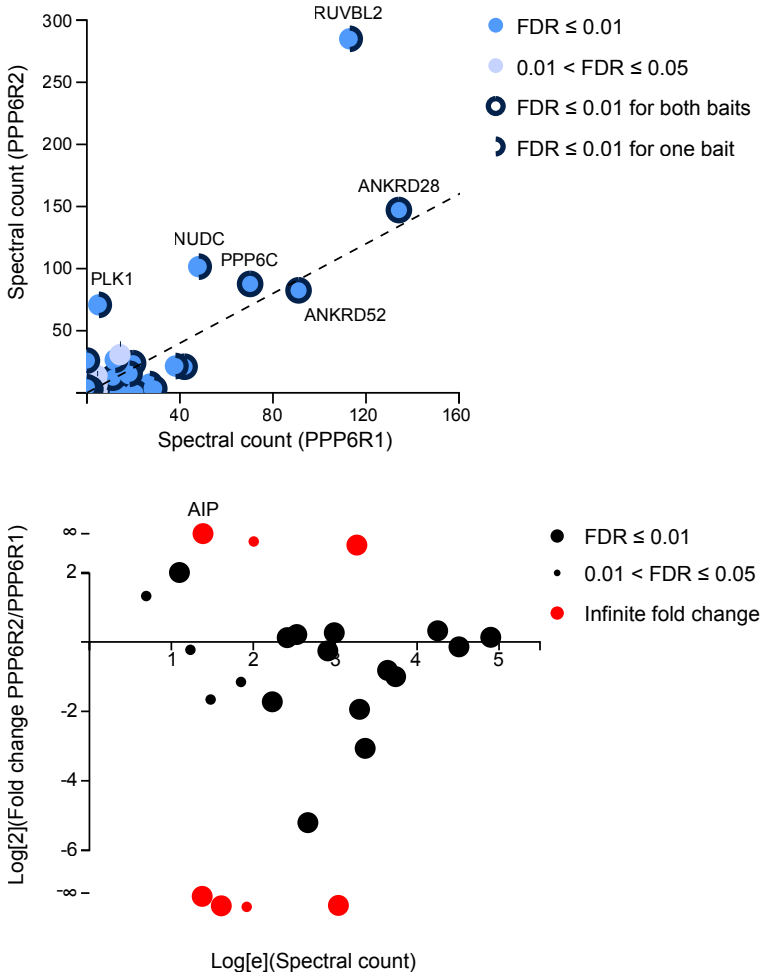
Problems generally result from errors in the input file format, and we encourage users to compare their input files against the samples provided on the web page.

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

ProHits-viz

Bait vs bait comparison

This tool will take quantitative protein-protein interaction data for two baits and generate a scatterplot summarizing the spectral counts, relative change and confidence information. The image produced can be in one of two formats, as shown below. The first (called versus) displays the prey abundances for each bait relative to each other, while the second shows the prey fold change of one bait relative to a reference bait. These types of images are particular helpful for a pair of baits when there are a large number of preys between them, as the information is displayed in a much smaller area than what would be generated with a dot plot or heat map.



A static image and a file compatible with our interactive viewer will be available upon completion of the task. Alternatively, 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).

Bait selection

Two baits will be compared against one another and they should be specified here. The reference bait would be the control/wild-type/untreated bait against which the other bait (secondary) will be compared.

Parameters

1. Output type: Bait-bait comparisons can be output in two different types. Select the type here. By default, images will be output as the "versus" plot type.

2. Primary filter value: All preys that satisfy this cutoff for the control bait will be displayed in the scatter plot. For the "fold-change" plot type, these preys will be displayed at the largest circle size, while in the "versus" plot type, they will be shown in blue.

3. Secondary filter value: All preys that satisfy this filter will also be displayed on the scatter

plot. For the "fold-change" plot type, these preys will be shown at a smaller size than those that passed the primary filter, while in the "versus" plot type, they will be shown in light blue.

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. 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.

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

Output

After the data has been processed, the user will be prompted to download a BaitvBait_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. baitvbait.pdf: This is the scatter plot which 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. legend.pdf: This is an FDR legend for the scatter plot and it can be opened and edited in Adobe Illustrator or a similar program.

3. BaitVsBait_data.txt: This file is a tab-delimited table that contains information for each point on the scatter plot. Often in the case of crowded scatter plots, data points and labels may overlap. This table is provided to assist the user in finding preys of interest in such cases, and to help resolve issues as to which points and labels belong to one another.

4. baitvbait_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.

5. 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.

ProHits-viz

Interactive dot plot/heat map viewer

This manual provides usage information for the interactive dot plot and heat map viewer at prohits-viz.lunenfeld.ca. This tool is designed to interactively display dot plots and heat maps generated from other tools on Prohits-viz.

Browser compatibility

Optimal performance of this tool is achieved with Chrome (tested in versions 47 - 55) although it is fully compatible and supported in Firefox (tested in versions 43 - 50) and Safari (tested in versions 7 - 9). For optimal performance with very large images, close unnecessary applications and browser tabs. In some cases fonts may appear too large in Firefox due to Firefox setting a hard cap on the minimum font size. This can be disabled by following the instructions here.

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

Sample input files

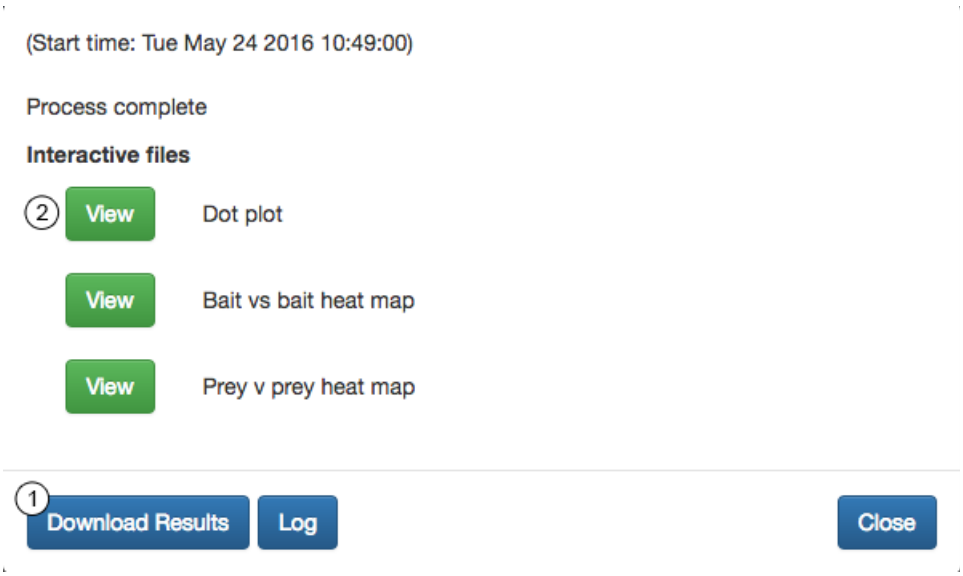
Input files are available for download from the "sample input files" hyperlink on the main page. These files were generated using the Dot plot Generator and Correlation tool using BioID data from the following paper:

Couzens, A.L., Knight, J.D.R., Kean, M.J., Teo, G., Weiss, A., Dunham, W.H., Lin, Z.Y., Bagshaw, R.D., Sicheri, F., Pawson, T., Wrana, J., Choi, H., Gingras, A.-C. (2013) Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. *Sci Signal,* **6:** rs15. PMID: 24255178

Inputting data

Currently, compatible output is produced from the following tools: Correlation Analysis and the Dot plot Generator. After running these tools, results can be downloaded (**1**) and interactive files for this viewer will be found in the folder *InteractiveFiles*. Alternatively, results can be passed directly from these tools by clicking on the *Interactive files* links (**2**) after the task is complete.

Figure 1



The input interactive file is simply a plain tab-delimited text file with between four to five columns ordered first by row and then by column as shown in **Figure 2**. The *score* column is optional for heat maps, while the column names *row*, *column*, *value* and *params* are required in all cases. Entries in the *params* column will dictate how the viewer displays and handles the input data. This column will be generated automatically by the tools on Prohits-viz but can be edited as desired. The first row in *params* must be one of the available display types, either "dot plot" or "heat map". The second row specifies the type of filter that should be applied, with "0" indicating a score that functions as a "less-than" filter, i.e. the lower the score the better (e.g. FDR), while "1" indicates a score that functions in the opposite way, a "greater-than" filter. The third and fourth rows specify the primary and secondary filters to use for coloring the circle edge on dot plots. The fifth column indicates the name for the "score" being used, while the sixth column indicates the type of "value", in this example "Spectral count". These names are used simply for labeling the legend. For heat maps, only the first and sixth rows need to be specified.

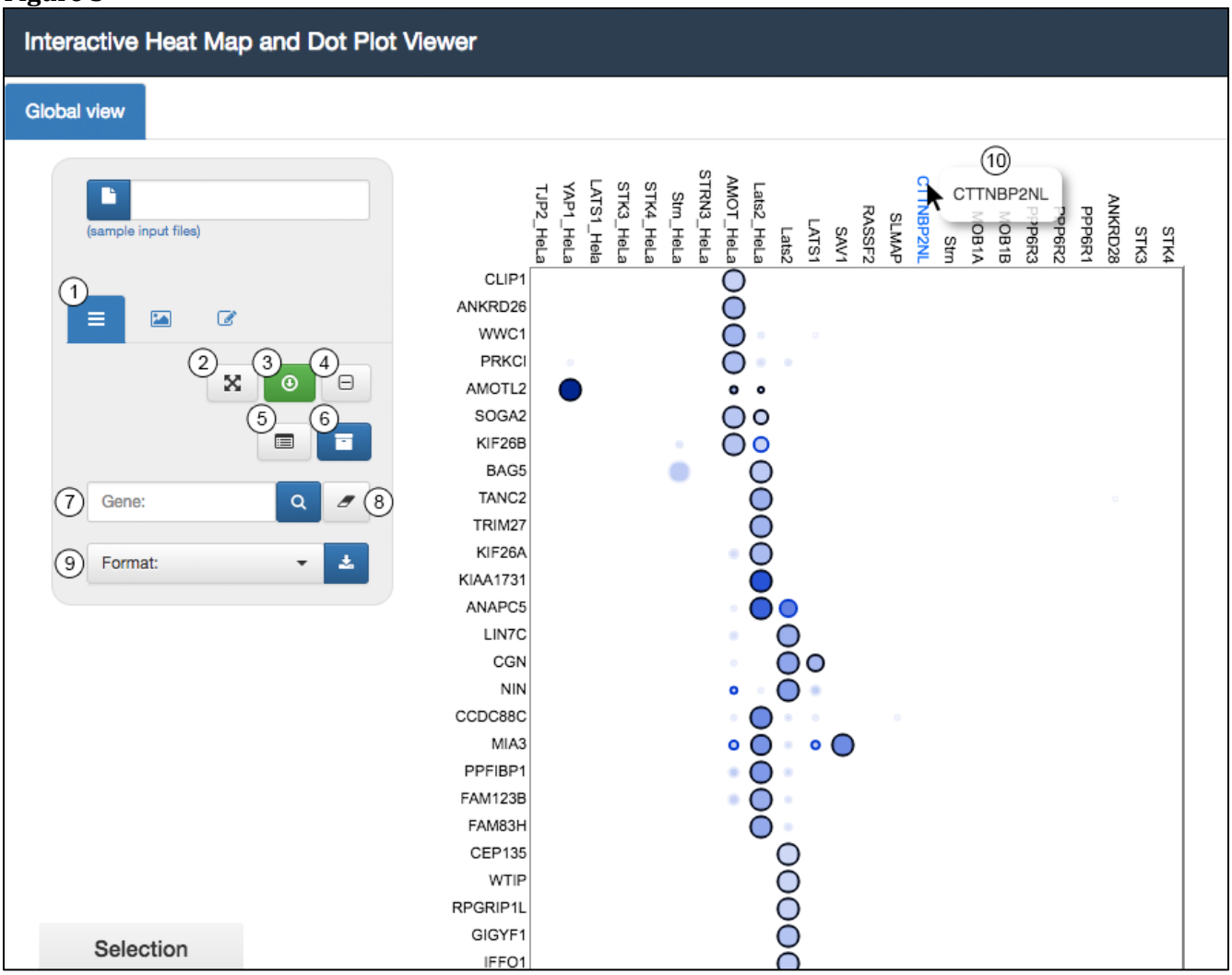
Figure 2

row	column	value	score	params
gene A	gene X	4	0.06	dot plot
gene A	gene Y	5	0.06	0
gene A	gene Z	76	0.01	0.01
gene B	gene X	34	0	0.05
gene B	gene Y	26	0	FDR
gene B	gene Z	13	0.02	Spectral count
gene C	gene X	0	1	
gene C	gene Y	41	0	
gene C	gene Z	91	0	

Global view

After loading an image into the viewer there are several navigation features available from the side panel *Options* tab (☰ **1**). If axis labels are too small, ⤧ (**2**) can be enabled to enlarge them as shown in **Figure 4**. As the user scrolls down the page the x-axis will follow the scroll by default, although this can be disabled by de-selecting ⓓ (**3**). Select ⊟ (**4**) to remove empty rows and columns on zoomed areas, i.e. rows and columns that contain no data points. The legend can be toggled with ▤ (**5**). Images on ProHits-viz will only be stored temporarily. Clicking the archive ▀ (6) button will store the image in our database with a permanent URL. The input at (**7**) is used for gene searches. Searches are case insensitive. Matches will be highlighted with a green circle on the appropriate axis (both if applicable) and these can be removed by ◢ (**8**). Finally, the image can be exported using (**9**) either as an SVG or PNG. When saving very large images (heat maps with more than 200K cells), we recommend saving first as a PNG before attempting an SVG export as this operation can often be too memory intensive. Long gene names will be truncated in the axes labels, but on hovering the full name will be displayed (**10**). The full name will be shown in an exported file.

Figure 3

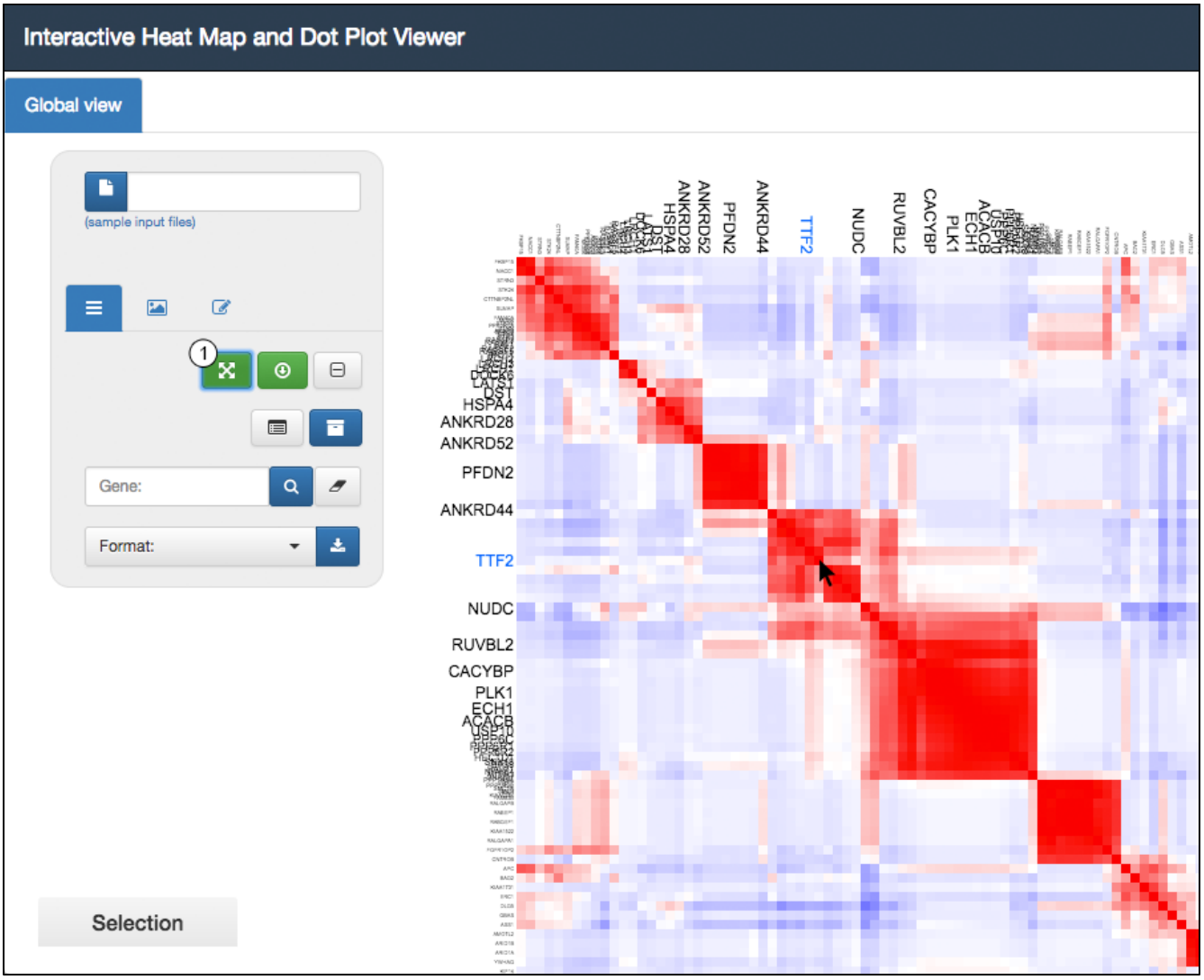


Interactive viewer

Image navigation

Axis labels that are too small can be selectively enlarged relative to the cursor position by enabling ⤧ (**1**). The cell over which the cursor is positioned will have its labels colored blue.

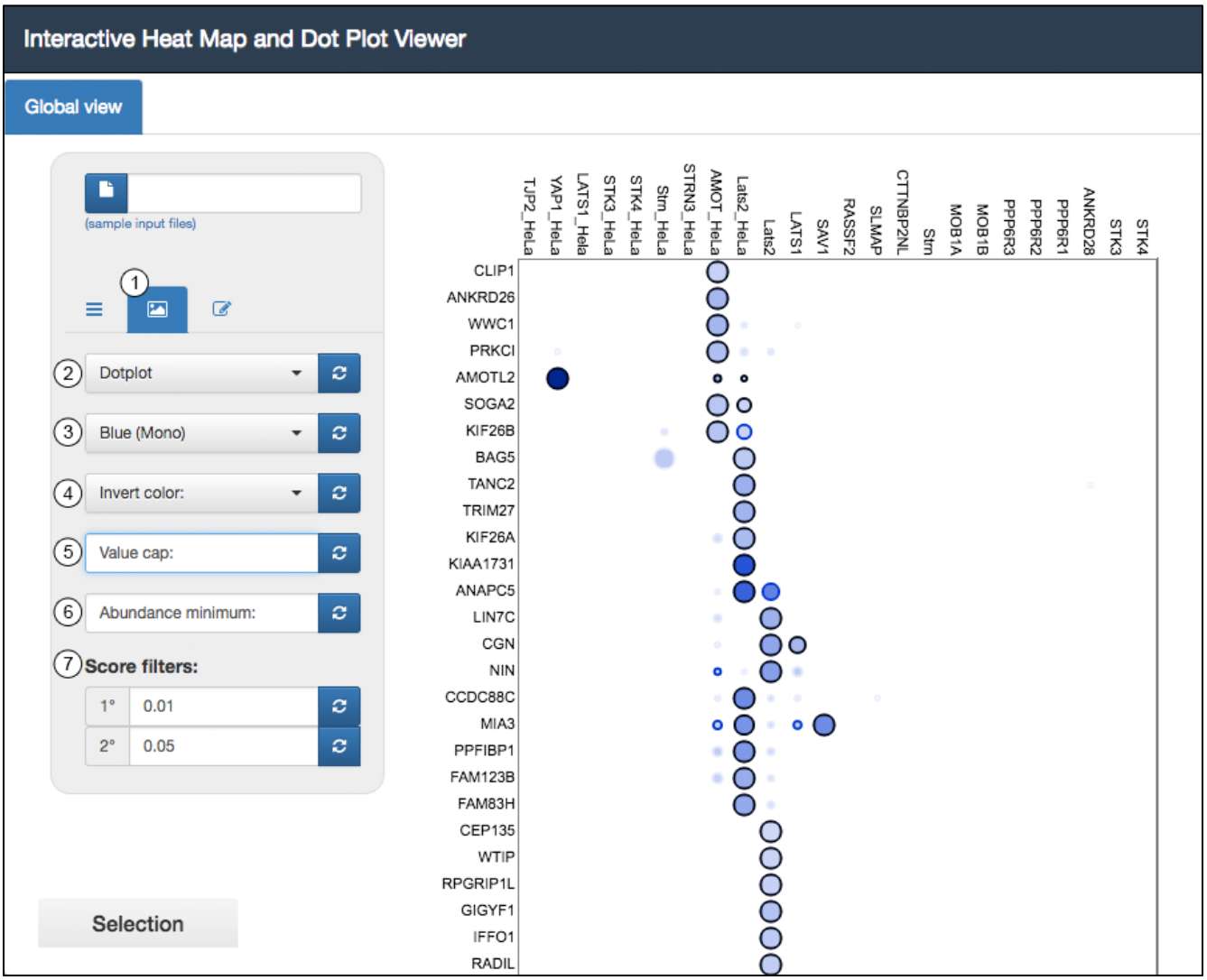
Figure 4



Display options

The image that is shown on the *Global view* tab is static in that rows and columns cannot be re-ordered, however superficial display changes can be made from the *Display options* side panel tab (**1**). The image can be switched between a dot plot and heat map by using the select box at (**2**) and hitting refresh. When an interactive file is loaded into the viewer, a dot plot will be displayed by default if an *score* column is present. If a *score* column is not present and the user switches from a heat map to a dot plot, it will be assumed that all entries have passed the 1° filter and will be displayed accordingly. A variety of color schemes are available from (**3**) although monochromatic (*Mono)* color schemes will only be available for data with values ≥ zero. The color scheme can be inverted with (**4**). For example, for data with positive and negative values a color scheme of red-green can be selected where red would be used for negative values and green for positive, and the *Invert color* option allows this to be reversed. The *Value cap* (**5**) specifies the color display cap and by default is set to the maximum value in the dataset or 50, whichever is lower, although the user can select a different value. Any values above this cap will be displayed in the image with the same value as the cap. This is too ensure there is sufficient dynamic display range for lower values in the dataset. Data points below a certain minimum can be removed from the image with (**6**). The coloring of edges in dot plots is controlled by the filter settings in (**7**). The 1° filter dictates the darkest edge color, while the 2° filter dictates the medium intensity edge color; values above this secondary cutoff will be displayed with the lightest edge color.

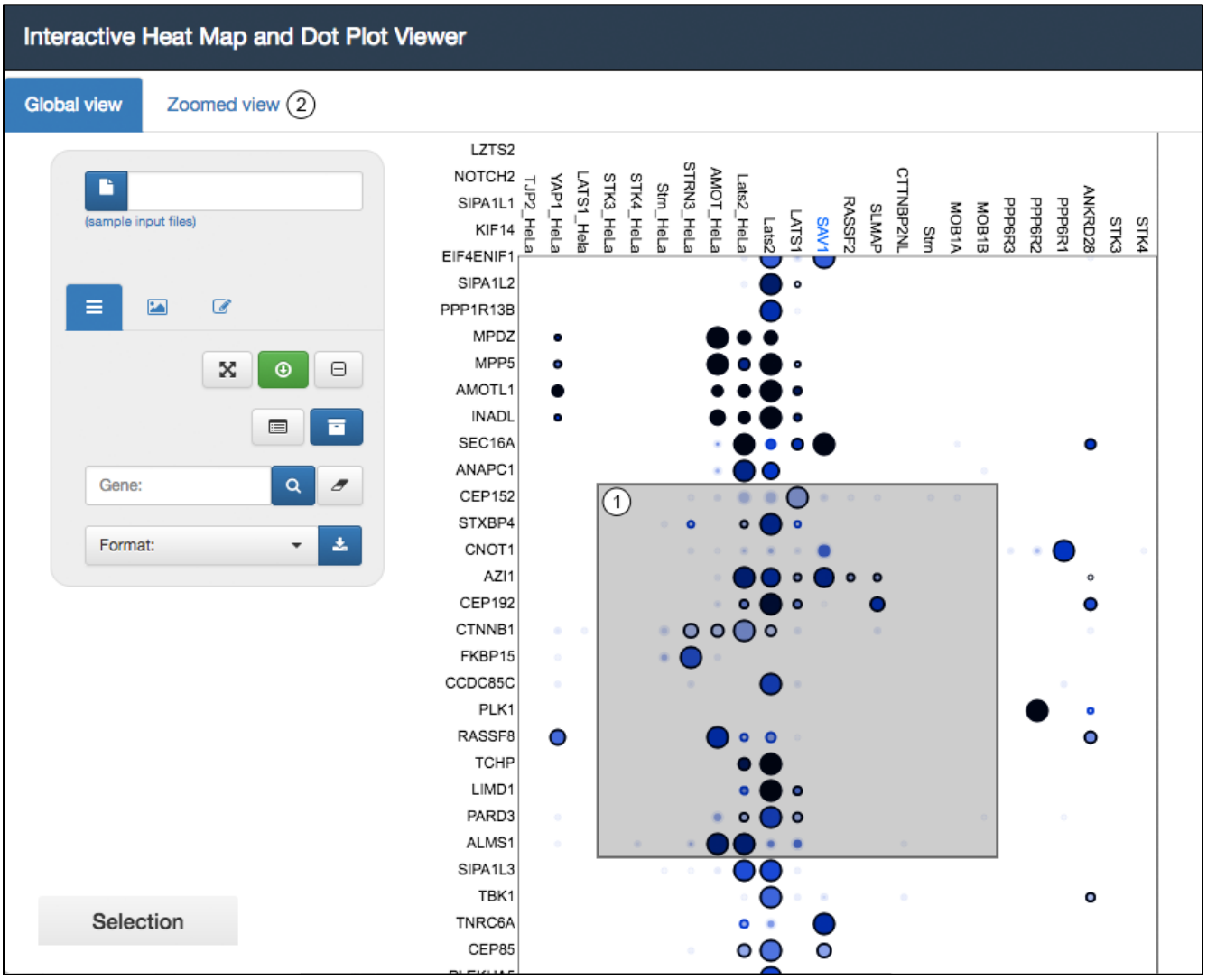
Figure 5



Selections

From the *Global view* tab regions can be selected for zooming and further analysis by dragging and selecting a region with the mouse (**1**). The selected region will be passed to the *Zoomed view* tab (**2**).

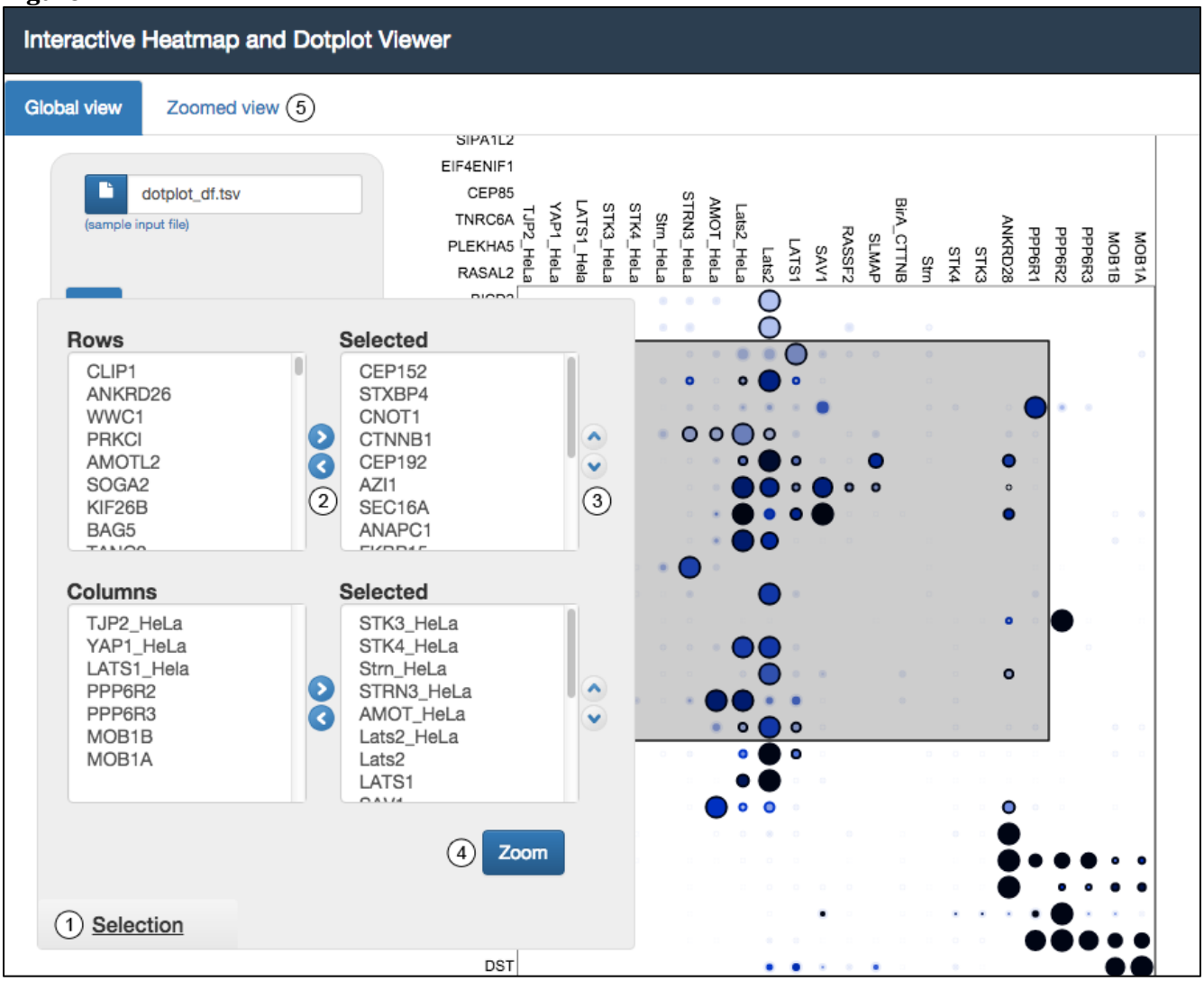
Figure 6



Selections continued

Alternatively, genes or regions can be selected use the *Selection* panel (**1**), shown in **Figure 7**. Here, row and column names will be listed based on their order in the input file. If clustering has been performed using the Dot plot generator, the order will reflect the order post-clustetering. They can be chosen and passed to the *Selected* boxes using the arrows shown at (**2**). The arrows at (**3**) can be used to reorder the genes. On hitting *Zoom* (**4**), the selected genes will be displayed in the chosen order on the *Zoomed view* tab (**5**). Note: if rows have been selected but no columns, all columns will be selected by default when hitting *Zoom*, and vice versa.

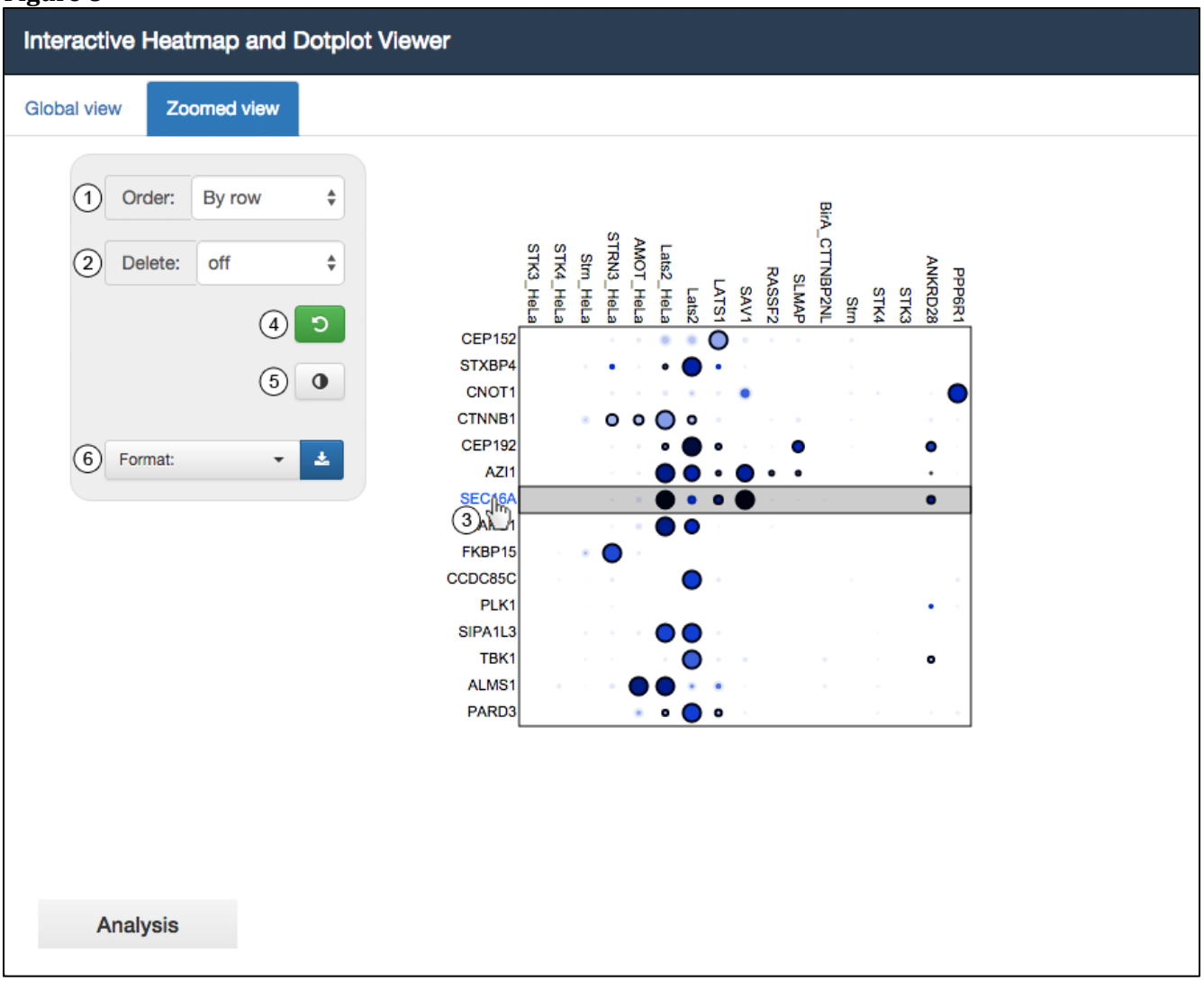
Figure 7



Zoomed view

The *Zoomed view* tab is used for exploring zoomed regions, generating custom images and performing analysis. The *Zoomed view* is a dynamic image in that rows and columns can be re-ordered and/or deleted. To re-order or delete a row or column, select the appropriate entry (from (**1**) or (**2**)). When re-ordering, clicking on a row or column in the image (**3**) will highlight it while dragging will reposition it. Deleting a row or column is done by simply clicking on it after selecting the desired action from the "Delete" box (**2**). The most recent action can be undone by pressing ↺ (**4**); pressing this button again will undo the next most recent action and so on. On dot plots, the circle size indicates the relative abundance of a row gene across the columns. On the *Zoomed view* tab this size will still be in reference to the entire data set by default. However, ◑ (**5**) can be toggled to make this size relative to the currently selected region. Finally, the image can be exported as a PNG, SVG or PDF using 🖫 (**6**).

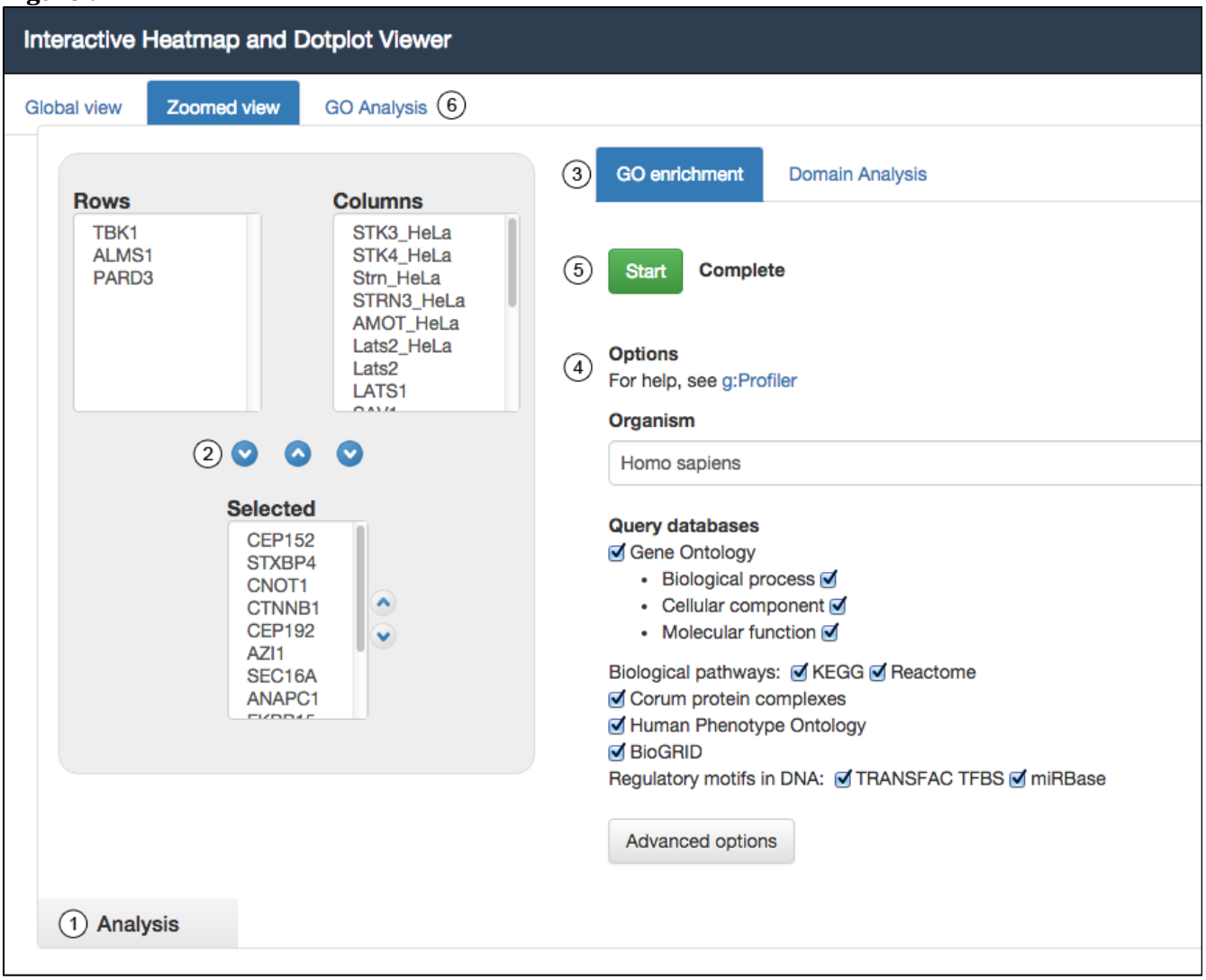
Figure 8



Analysis

Genes displayed on the *Zoomed view* tab can be passed for analysis by selecting the *Analysis* panel (**1**) in **Figure 9**. Genes highlighted in the *Rows* and *Columns* lists can be passed to the *Selected* list using the arrows at (**2**). Available analysis options are displayed as tabs (**3**). Currently, three analysis options exist and more will be added on an ongoing basis. To learn more about the current analysis options, follow the hyperlink (**4**) next to "For help". In this example, GO enrichment is performed through the site at biit.cs.ut.ee/gprofiler/ where the user can find help information and details about the analysis tool. Analysis begins by pressing *Start* (**5**) and the results will be available as a new tab (**6**) when complete.

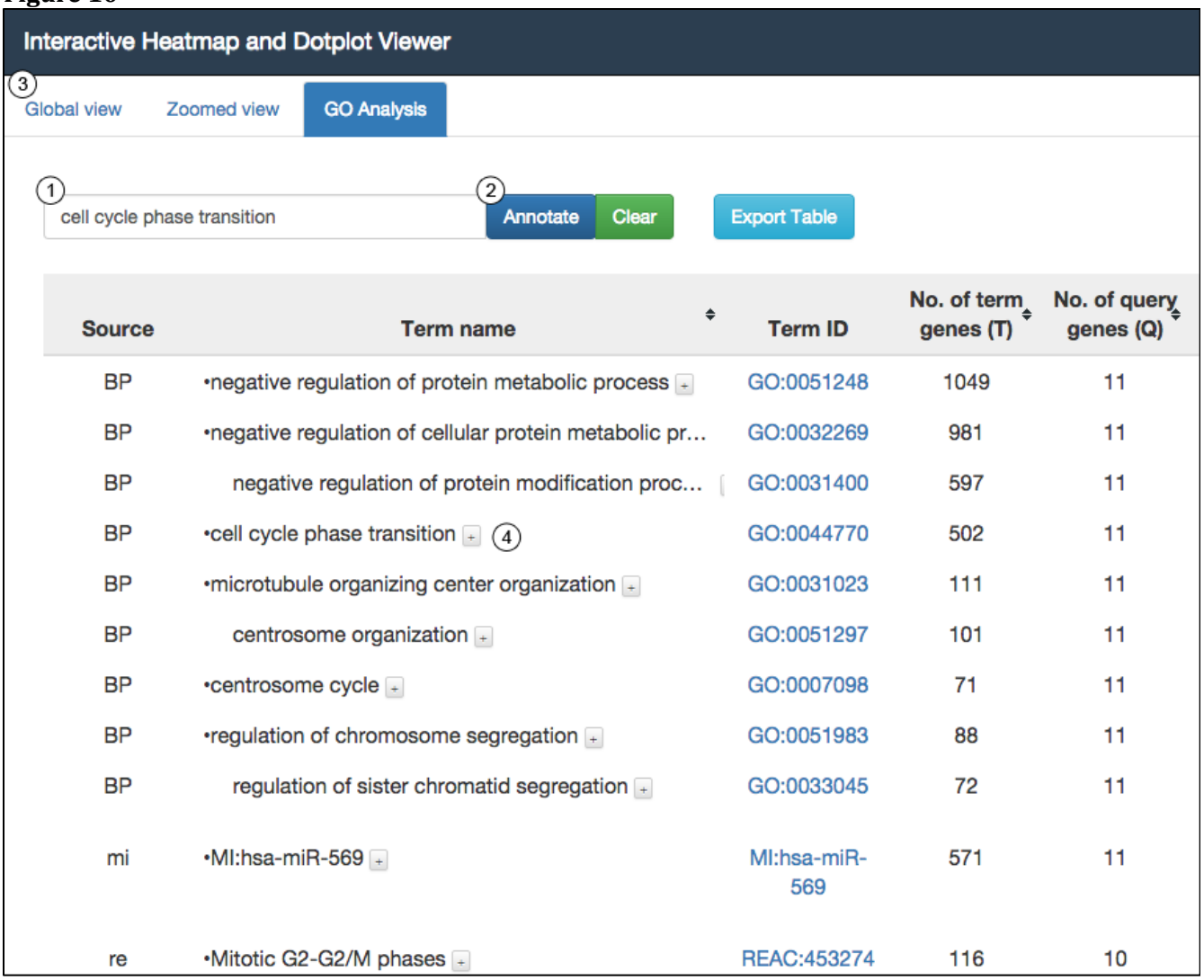
Figure 9



Analysis tabs

On an analysis tab, the user can view the results of their inquiry. At the top of every analysis tab is located an annotation text box (**1**) and an *Annotate* button (**2**). From here the user can annotate the original selection made on the *Global view* tab (**3**). For example, if the user thought "cell cycle phase transition" best characterized his/her chosen selection, that term could either be typed in the annotation box or passed directly to it by pressing the + button (**4**) beside the term. This term will be displayed over the selection on the *Global view* tab (**Figure 11**) after pressing Annotate (**2**).

Figure 10

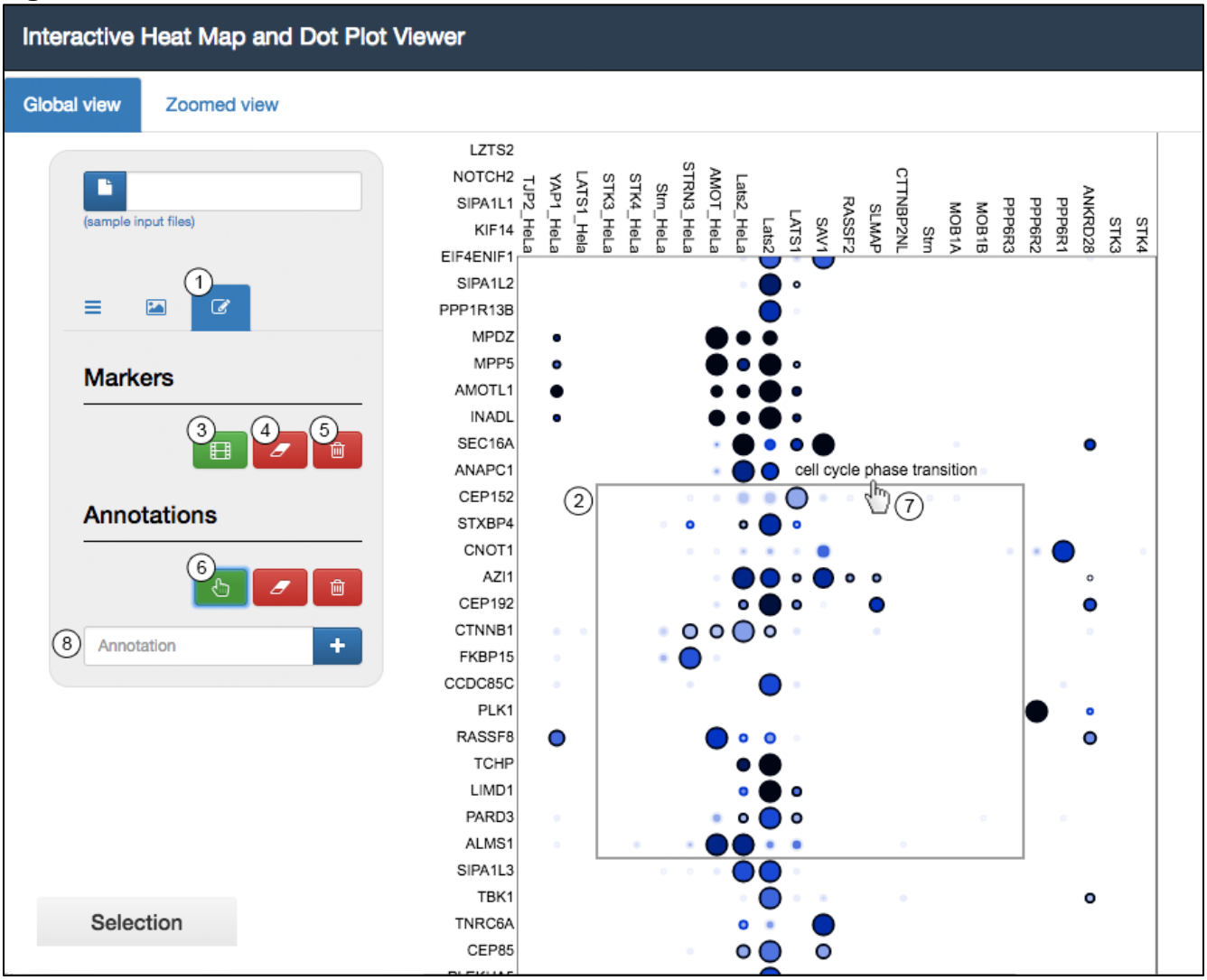


Source	Term name	Term ID	No. of term genes (T)	No. of query genes (Q)
BP	•negative regulation of protein metabolic process	GO:0051248	1049	11
BP	•negative regulation of cellular protein metabolic pr…	GO:0032269	981	11
BP	negative regulation of protein modification proc…	GO:0031400	597	11
BP	•cell cycle phase transition	GO:0044770	502	11
BP	•microtubule organizing center organization	GO:0031023	111	11
BP	centrosome organization	GO:0051297	101	11
BP	•centrosome cycle	GO:0007098	71	11
BP	•regulation of chromosome segregation	GO:0051983	88	11
BP	regulation of sister chromatid segregation	GO:0033045	72	11
mi	•MI:hsa-miR-569	MI:hsa-miR-569	571	11
re	•Mitotic G2-G2/M phases	REAC:453274	116	10

Annotations

After annotating a selection from an analysis tab and returning to the global view, further customizations are available from the side panel *Annotations* tab (**1**). All selected regions will by default be delimited by a border (**2**). This automatic recording of selections can be enabled by (**3**). If selections are being recorded, they can be removed individually (in the reverse order from which they were added) by (**4**) or completely removed by (**5**). Text annotations will be placed in the middle of the selection box by default but these can be moved by selecting (**6**) and dragging the annotation to the desired position (**7**). Manual annotations can be added with (**8**). Similar removal options are available for annotations as were described for selection boxes. If any adjustments to the image are made on the side-panel *Display options* tab (as shown in **Figure 5**), for example changing from a dot plot to a heat map or changing the color scheme, all annotations and selection borders will remain on the image. If the image is exported as an SVG, all annotations will be editable in Adobe Illustrator. PDF images will also be editable, although text elements may be grouped in unexpected ways.

Figure 11



Interactive scatter plot viewer

This manual provides usage information for the interactive scatter plot viewer at prohits-viz.lunenfeld.ca. This tool is designed to interactively display scatter plots generated from other tools on Prohits-viz.

Browser compatibility

Optimal performance of this tool is achieved with Chrome (tested in versions 47 - 55) although it is fully compatible and supported in Firefox (tested in versions 43 - 50) and Safari (tested in versions 7 - 9). For optimal performance with very large images, close unnecessary applications and browser tabs. In some cases fonts may appear too large in Firefox due to Firefox setting a hard cap on the minimum font size. This can be disabled by following the instructions here.

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

Sample input files

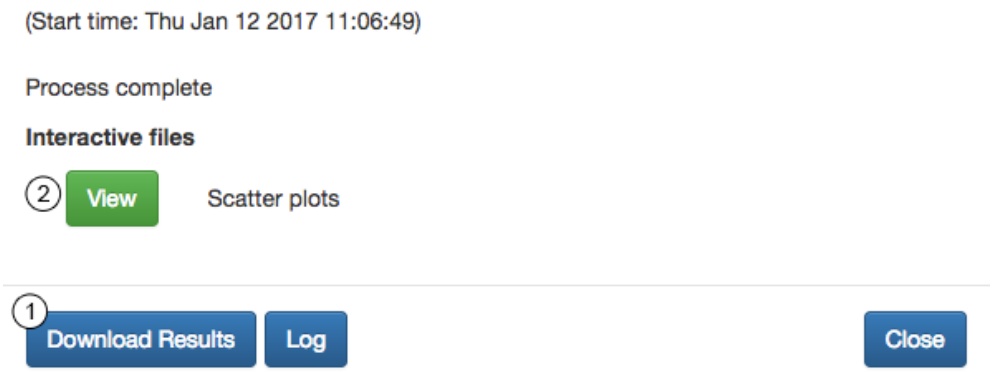
Input files are available for download from the “sample input files” hyperlink on the main page. These files were generated using the Prey Specificity and Bait-Bait Comparison tools using BioID data from the following paper:

Couzens, A.L., Knight, J.D.R., Kean, M.J., Teo, G., Weiss, A., Dunham, W.H., Lin, Z.Y., Bagshaw, R.D., Sicheri, F., Pawson, T., Wrana, J., Choi, H., Gingras, A.-C. (2013) Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. *Sci Signal,* **6:** rs15. PMID: 24255178

Inputting data

Currently, compatible output is produced from the following tools: Prey Specificity and Bait-Bait Comparison. After running these tools, results can be downloaded (**1**) and interactive files for this viewer will be found in the folder with the extension _df.tsv. Alternatively, results can be passed directly from these tools by clicking on the *Interactive files* link (**2**) after the task is complete.

Figure 1



The input interactive file is simply a plain tab-delimited text file with six columns as shown in **Figure 2**. The first line begins with the word "entry" to signify the start of information on a scatter plot and is followed by the plot title, the x-axis label, the y-axis label, the plot type (anything can be entered here as this is a custom parameter for generating legends specific to the tools at ProHits-vis) and the name for the score parameter in the dataset (optional). This line is followed by lines for each data point, beginning with the gene name, x and y coordinates, circle radius in pixels, circle colour (#hex value) and text information on the point that will be displayed when the point is moused-over. Multiple scatter plots can be included in a single file, with each beginning with the "entry" line as described above.

Figure 2

entry	plot title A	x-axis label	y-axis label	plot type	score name
gene A	x coordinate	y coordinate	circle radius	circle color	point information

Interactive scatter plot viewer

After loading an image the title of the current scatter plot will be displayed at (**1**). If more than one scatter plot is available to view, the others can be selected from this same dropdown. If the plot has been zoomed or translated, its original view can be reset with (**2**). Labels can be added to the plot with (**3**). Individual points can also be labelled selectively by clicking on them (**4**). The legend can be toggled with (**5**). By default, plots will be drawn to fill the browser window but square plots (x- and y- axis equivalent lengths) can be made by selecting (**6**). The image can be exported using (**7**) either as an SVG or PNG image. Images on ProHits-viz will only be stored temporarily. Clicking the archive (8) button will store the image in our database with a permanent URL. The input at (**9**) is used for gene searches. Searches are case insensitive and currently require a complete match. Matches can be removed by (**10**). To display information about a point, simply hover over it (**11**). Finally the image can be zoomed along both axes by positioning the cursor over the plot and using the scroll wheel (**12**), or zooming can be restricted to a single axis by placing the cursor over that axis (**13**).

Figure 3