SIGNAL user guide
v1.0 February 2021

Sections:
Titles as they appear on the sidebar menu:

Text as you click on each

1. Sample dataset:

The sample dataset uses the data from the *Sun et al.* siRNA study measuring the TNF transcriptional response to LPS in Human Macrophages [1].

1. To run the analysis in SIGNAL, begin by accessing the SIGNAL interface in your web browser https://signal.niaid.nih.gov (Chrome is the recommended browser) and downloading the dataset to your local device.
2. Click on the "Browse" icon in the left panel to locate where you have saved the sample file. Select the sample file and click "choose".
3. A dropdown bar with the title "Cutoff Type" will appear,

There are three ways that a dataset can be split into high-confidence, medium confidence, and low confidence/non hits sets. The sample dataset can be analyzed using either of these approaches.

A. Using a single criterion:
    a. Select "Zscore" from the "Cutoff Type" dropdown menu.
    b. In the field "High Confidence Cutoff Value" enter the value -2, in the field below it ("Medium Confidence Cutoff Value") enter the value -1.5.
    c. Leave the "Add an Additional Criteria" unchecked.

B. Using two criteria:
    a. select "Zscore" from the "Cutoff Type" dropdown menu.

b. In the field “High Confidence Cutoff Value” enter the value -2, in the field below it (“Medium Confidence Cutoff Value”) enter the value -1.5.
c. Click on the “Add an Additional Criteria” option.
d. Select “OffTarget_pValue” from the “Column to Use for Secondary Criteria” dropdown menu.
e. Select “≥” from the “Direction” dropdown menu.
f. In the field “Value” enter 0.05.

C. Assigning a distinct value to each gene before uploading to SIGNAL and using the assigned value to separate the tiers:
   a. select “assigned.value” from the dropdown menu.
   b. In the field “High Confidence Cutoff Value” enter the value 1, in the field below it (“Medium Confidence Cutoff Value”) enter the value 0.5.
   c. Leave the “Add an Additional Criteria” unchecked.

4. Click “Analyze my data”.

Note: The rest of the guide will follow the analysis using option A (Zscore).

A progress bar will appear at the bottom right of the browser window. Once it is complete the window will shift to the Enriched Pathways tab listing the enriched pathways identified in the analysis. Gene IDs that were assigned as high confidence hits in the upload file are in blue. Gene IDs that were categorized as medium confidence hits in the input file are in red.

The names of the enriched pathways can be clicked on to open a new window with a KEGG pathway map overlaying the hits identified in the analysis.

5. Click on the “Proteasome” pathway box. A warning window will popup saying you are leaving an NIH website. Click OK. A new tab will open in your browser with a schematic of the proteasome. The SIGNAL identified hits from the screen are highlighted in blue (high confidence) and in red (medium confidence).
6. Return to the SIGNAL tab in your browser and click on the “Gene Hits” tab. The table with the “SIGNAL Gene Hits” lists all the hits identified by SIGNAL from the uploaded study. Additional columns show which of the other identified hits from the dataset the listed candidate has predicted interactions with and what enriched pathways the interacting genes are members of.
7. Click on the subtab labeled “Graph: Gene Hits by Iteration”. A table and graph show the number of high confidence and medium confidence hits selected as candidates through each iteration of SIGNAL. Iteration 0 corresponds to the settings of the upload file with 453 hits described as high confidence. After one iteration a number of the high confidence hits are dropped out and a number of the

medium confidence hits are added. The cycle repeats until iteration 3 & 4 that show no changes and thus the iterative analysis terminates.

8. Click on the “High Confidence Hits Not Selected by SIGNAL” subtab. All the hits that were designated as high confidence in the upload file but were *not* selected as hits by SIGNAL are listed in this table. This helps the user review any of the candidates that were not selected by SIGNAL that the user might want to manually add to the set of selected hits.
9. Click on the “Pathway Enrichments” subtab. The enriched pathways with statistics and gene names are shown in a table. The last column includes an “Enrich score” sorting the enrichment score from highest to lowest, shows a strong enrichment for glycan related pathways.

In addition to the lists of selected hits and pathway enrichments that SIGNAL identifies the platform also enables the exploration of selected hits and possible ‘missing links’ between enriched pathways. Using a unique and interactive network display, the user can explore predicted interactions between hits associated with specific enrichments and hits outside of the group to infer novel regulatory links.

10. Click on the “Network” tab.
11. Select (by checking the adjacent box) Pathways 1: Proteasome, 5: Toll-like receptor signaling pathway.
12. Click “>> Create Network Graph” at the top of the page. A progress bar will show the progress of the graph. Once complete the “Network Path” tab will open.
13. The image can be enlarged by zooming in on the browser image (Command + for Mac, or Ctrl scroll for PC). Adjust the text size using the sliders on the right to make the image more legible and aesthetic.
14. Hover with your cursor over the names or nodes of different genes to see their interactions and more information in the window at the top right of your browser.
15. Click on the gene “PSMC3” that is in the red Proteasome group in the top right of the graph. The predicted interactions will appear, as well as further information in the top right window.
16. Hover over the highlighted gene “TNFRSF1A” which is on the left of the graph in the green group. The information window will fill with information about this particular interaction based on the information from the STRING database.
17. Click on the “TNFRSF1A” label, the gene ID “MAP3K7” in the Toll-like receptor signaling pathway group (blue) will highlight among others. Click on “MAP3K7”.

    Note: if at any point an unintended or incorrect click was made, the click can be undone by clicking “Revert Click” at the right side of the browser window. To restart the process, choose “Click to reset”.

18. Click “Highlight Clicked Pathways”. The graph will show the interactive pathway between Proteasome hits and Toll-like receptor signaling pathway hits going through TNFRSF1A.

19. Click on the “Clicked Pathways Table” subtab. The pathway clicked through above is presented in table format.
20. Click on the “Download” tab. All the analysis generated are available in CSV format and can be download in a zip folder. (Note: the “Clicked Pathway Table” resets every time the clicked pathway is reset in the graph. To save a specific clicked table before trying a different path click on the download folder to download the files before clicking “Click to reset” to save the most recent clicked table.) For data security, once the browser window is closed the analysis is deleted from the server.

References

1. Sun, J., et al., *Genome-wide siRNA screen of genes regulating the LPS-induced TNF-alpha response in human macrophages.* Sci Data, 2017. **4**: p. 170007.

2. Preparing Your Data:

There are three requirements for a file to be successfully uploaded for analysis by SIGNAL:

1. The file must be in a .csv, .txt, or .xlsx format
2. The file must contain a correctly labeled column with gene IDs (either NCBI EntrezID or HGNC GeneSymbol)
3. The file must contain a column with numeric values that can be used to separate high confidence hits, medium confidence hits, and non-hits in the data set.

A description of these requirements are detailed below:

'images/SampleInput.png'

File format: To upload your data for analysis by SIGNAL first ensure that your document is in .csv, .txt. or .xlsx format.

IDs: The file must contain one of the following:

1. A column exactly titled “GeneSymbol” that has HGNC gene symbols in all the rows

   or, alternatively,

2. A column exactly titled “EntrezID” with NCBI EntrezID in all the rows.

Both ID columns, however, do not need to be included in the upload file. If only one is included, SIGNAL will identify the missing ID type and generate a column of the other IDs based on the provided IDs. If both are provided, SIGNAL will use both columns as provided by the input file.

Assigning confidence: The upload file must include a column with the numeric values to be used for selecting high and medium confidence hits. The name of this column is up to the user.

The numeric values can be either continuous values (such as a range of *p* values or a range of Zscores) or assigned values (such as assigning a value of 1 to all IDs that should be considered "high confidence". A value of 0.5 to all IDs that should be considered "medium confidence" and a value of 0 for all IDs that should be considered "non-hits"). A more detailed description of how to assign the value and cutoffs to different types of datasets are described (*Appendix A: How to segment data into confidence tiers)* Each gene ID should only have one value associated with it and be listed in the document only once.

If the user wishes to use a secondary filtering criteria (such as p-value in addition to fold change), that column must have a different name than the primary column used for assigning confidence cutoffs.

Additional columns can be included in the document and they will be skipped over by the analysis but preserved in the main download file (*[name of your upload file]_SIGNALhits.csv*)

3. Running an analysis:

'images/SideBar_map.png'

Begin by loading the SIGNAL website in your browser using the web address https://signal.niaid.nih.gov/

A panel of dropdown menus will be on the left side of the browser window. Select the parameters that best describe your data and the database setting you want to use for your analysis.

'images/StartAnalysis_page.png

Some of the parameters come with default options, others require an input from the user. Following is a step-by-step description of the parameter definitions and settings:

Select your organism:

Under this dropdown menu select “Human” or “Mouse” based on the gene IDs used in the dataset.

Default setting: *Human.*

Select a Database for Enrichment Analysis:

For the enrichment analysis portion, SIGNAL uses the pathway designations curated by the Kyoto Encyclopedia of Genes and Genomes (KEGG). (for information about the KEGG database, see https://www.kegg.jp/kegg/pathway.html)

Under the **Select a Database for Enrichment Analysis** dropdown, the user can select whether to use only the pathways from the KEGG database that describe biological processes by selecting the *KEGG: Biological Processes* option, or to use only the pathways associated with disease descriptions by selecting the *KEGG: Disease Pathways* option. To use both types of pathways the user can select *KEGG: All Pathways* which includes all of the pathways curated in the KEGG database.

Default setting: *KEGG: Biological Processes.*

Select Interactions for Network Analysis:

For the network analysis component, SIGNAL uses predicted protein-protein interactions from the STRING database (mapped back to the associated gene name). The default setting is *Experimental & Database* which incorporates interactions from STRING that have an evidence source in other experimental and curated databases. An additional option is to select *Advanced Options* from the drop down, once selected a list of six evidence criteria will appear and the user can manually select which interactions to include or exclude based on the evidence criteria of origin. (For a further explanation of each of the possible criteria see http://string-db.org/cgi/help.pl?sessionId=Z6t6X5gizxo0)**.**

Default setting: *Experimental & Database*

Interaction Confidence for Network Analysis:

The STRING database assigns to each predicted interaction a confidence score ranging from 0 to 1.

The user can select what confidence score (as defined on the STRING database) to consider for the network interactions used in the analysis. Only interactions whose confidence score cross the selected confidence threshold will be considered in the network analysis portion of the analysis. For simplicity, the provided cutoff options are broken into three groups. *Low (> 0.15)*, *Medium (> 0.4), High (> 0.7).* The higher the cutoff, the less predicted interactions are included in the set. The lower the cutoff, the

higher the likelihood of false positive interactions being included. Medium confidence is recommended for most cases.

Default setting: *Medium (> 0.4)*

Choose an input file to upload:

Upload your .csv file by clicking on the "Browse" button and locating the file in your computer. A progress bar will inform you when the upload is complete. The data from the uploaded file will appear in a table under the "Input" tab.

Cutoff Type:

Once the file is successfully uploaded a dropdown menu appears under the title "Cutoff Type". This is where the numeric column to be used for separating the IDs into different hit confidence groupings will be assigned. The dropdown menu consists of a list of the column names in the uploaded document. Select the column that contains the numeric values to be used for the high confidence/medium confidence cutoffs of your targets.

High Confidence Cutoff Value & **Medium Confidence Cutoff Value:**

Type in the numeric values to be used as a cutoff for high confidence and medium confidence hits from your screen. Based on the difference between the high confidence cutoff value entered and Med-confidence cutoff value entered SIGNAL will infer whether the values should be taken as "greater than or equal to" or "less than or equal to".

(For example if 0.01 is assigned to the *High Confidence Cutoff Value* and 0.05 is assigned to the *Med Confidence Cutoff Value* field SIGNAL interprets the input as any ID with a value of 0.01 or less in the selected column should be considered high confidence, any ID with an associated value between 0.01 and 0.05 should be assigned medium confidence, and any ID with an assigned value greater than 0.05 will be assigned as a "non-hit". Alternatively, if 2 is assigned to the *High-conf Cutoff Value* and 1 is assigned to the *Med-conf Cutoff Value* field SIGNAL interprets the input as any ID with a value of 2 or more in the selected column should be considered high confidence, any ID with an associated value between 1 and 2 should be assigned medium confidence, and any ID with an assigned value less than 1 will be assigned as a "non-hit".

Add an Additional Criteria:

Checking this option allows the user to add an additional condition that all hits must satisfy. This is useful for when an additional metric is to be used, such as *p value* or *cell count*, to filter hits in addition to the criteria selected in the "Cutoff Type" menu.

Column to Use for Secondary Criteria:

If the “Add an Additional Criteria” option is selected a drop down menu with all the column names from the uploaded document appears (minus the column name selected under “Cutoff Type” for the first criteria). Select the column that contains the numeric values that will have to meet the assigned cutoff.

Direction & **Value:**

Select the direction (≥ or ≤) and value for the cutoff to be applied to the secondary criteria. All high-confidence hits and medium confidence hits from the primary criteria that *don’t* meet the cutoff of the secondary criteria are reassigned as “non-hits”.

Add genome background:

Checking this option adds genes to the list that aren’t included in the upload file to be used as a background for statistical enrichment analysis.

This feature is recommended for when the upload file doesn’t include a genome-scale set of “no confidence hits” (as in when only a partial list of the targets are available and not a list of all targets in the genome). In order to be able to run enrichment analysis statistics on the group of assigned hits it is critical to have a robust background of “non-hits” against which to measure it. In cases where those IDs are not included, the “Add genome background” option provides a pseudo genome-scale background for the dataset. If the upload file includes a robust genome-scale background than the “Add genome background” can remain unclicked and only the IDs in the uploaded file will be used.

When selected, the added background “genes” will not appear as suggested hits by the SIGNAL analysis, the background genes are only used as a means to have more robust statistics on the enrichment of pathways. The added background genomes use only the known protein coding genes of the selected organisms (source: Human, HGNC. Mouse, MGI.) that are not in the upload file. SIGNAL uses the difference between the size of selected hits and the number of known protein coding genes for that organism as the number of “non-hits” for enrichment statistics.

Default: *Unselected.*

Optional: Enter gene IDs that should be kept as high confidence hits throughout the analysis:

This text box can be used to enter any gene IDs (either separated by comma or in separate lines) that should not be filtered out by SIGNAL and should be kept as hits independent of what the iterative analysis finds. This is recommended for when the study identifies genes in a pathway that hasn’t yet been annotated by the pathway database or when the user has a particular interest or knowledge about the relevance of these

candidates. Adding the gene IDs in this text box will place them in an “assigned hit” category and their enrichments and network interactions with other hits will be shown in the results.

It is important here to select to correct gene ID type (EntrezID or GeneSymbol) and to ensure that all the entries are in the correct format.

Analyze my data:

Once the parameters for that analysis have been selected and entered, click the “Analyze my data” icon and the analysis will begin. A progress bar at the bottom right corner of your browser window will indicate the progress of the analysis.

Reset:

This tab allows the user to reset all the selections and restart the analysis from the start with default settings. This can be done at any point in the analysis. Clicking the reset tab will remove the uploaded file and all the analysis files generated up to that point. If trying multiple analysis or different settings it is recommended that the analysis be reset between each run by clicking the “Reset” icon or refreshing the browser.

3. Reading SIGNAL Results:

Once the analysis is complete a range of interactive and downloadable tables are generated to suggest hit selection sets and facilitate exploration of the uploaded data.

SIGNAL provides analysis output on three levels: selected hits, enriched pathways, and generated networks. The output and data presentation is designed to connect between the three levels of outputs such that, as an example, findings in the selected hits section can be used to prioritize the enriched pathway section, and findings from the enriched pathways data exploration can be used to filter the enriched network, and findings from the generated network can be used to identify critical enrichments.

The output results are separated into five of the nine visible tabs:

Enriched pathways: An interactive table that lists all the pathway enrichments identified in your data by SIGNAL.

Gene Hits: A number of tables that list the hits selected by SIGNAL and their predicted and associated enrichments.

Network: A table to select up to three pathways from the list of enriched pathways to generate a pathway-network graph.

Network Graph: An interactive figure of the pathway-network graph. These figures will only populate after networks have been selected and plotted from the '*Network'* tab.

Download: A list and "Download" button to locally save all the tables generated in the analysis.

A description and guide for how to use the information in each tab is below.

Enriched Pathways:

When the analysis is complete, a list of enriched pathways appears in a table under the "Enriched Pathways" tab. The list includes all pathways that have a *p* value of 0.05 or less in the completed SIGNAL analysis of the uploaded data. The table also includes a list of the Gene IDs from the pathway that are also hits in the screen. The Gene IDs that were designated as high confidence in the input file are highlighted in blue, Gene IDs that were designated as medium confidence in the input file are highlighted in red.

'images/EnrichPathway_map.png'

The table includes columns providing the following information:

Pathway: This column includes the names of the enriched pathways. Clicking on the pathway name will open a new tab from the KEGG database showing a schematic of the genes in the pathways with the gene hits from the analysis highlighted. Genes that were marked as high confidence at the start of the analysis are highlighted in blue and those marked as medium confidence are highlighted in red.

'images/KEGGPathway_map.png'

pVal: The p-values based on a hypergeometric test for the enrichment of each pathway are listed.

pValFDR: The p-values with added correction for False Detection Rate are listed.

pValBonferonni: The p-values with the Bonferroni corrections for multiple testing are listed.

TotalGenes: The total number of genes in the pathway.

HitGenes: The number of hit genes as selected by the SIGNAL analysis that are in the pathway.

HitGeneNames: The HGNC Gene Symbols of the SIGNAL hit genes in each pathway are listed. Genes that were marked as high confidence at the start of the analysis are highlighted in blue and those that were marked as medium confidence are highlighted in red.

Gene Hits

The Gene Hits tab is a guide for hit selection from the uploaded data based on the analysis by SIGNAL. The Gene Hits tab is separated into five different tabs

SIGNAL Gene Hits: *A table of all the hits selected by SIGNAL from the high and medium confidence hits in the input file.*
Gene Hits by Iteration: *A table that lists all the hits in the input file along with what category of confidence they were assigned to at each iteration of the SIGNAL analysis.*
Graph: Gene Hits by Iteration: *A table and figure of the number of hits selected at each iteration from the input high and medium confidence categories.*
High Confidence Hits Not Selected by SIGNAL: *A table of hits that were assigned as high confidence in the input file but were not selected as high confidence hits by SIGNAL.*
Pathway Enrichments: *A table of pathway enrichments as assigned by SIGNAL.*

SIGNAL Gene Hits: This table provides a list of hits selected by the SIGNAL analysis. The table also includes supporting information on interacting genes (based on the network criteria that were selected at the start of the analysis) and membership in enriched pathways.

'images/SIGNALHits_map.png'

The table's columns and what they include are:

EntrezID: NCBI EntrezID identifier.
GeneSymbol: HGNC official Gene Symbol.
InputCategory: A designation of "HighConf" or "MedConf" based on the original category it was assigned to at the start of the analysis (based on the user provided cutoffs).
SIGNALhit: "Yes" indicates a hit selected by the SIGNAL analysis.
Pathway: Names of enriched pathways from the analysis which the associated gene is a member of. (This column only lists pathways that were selected as significantly enriched by the analysis. The individual gene may also be part of other pathways not listed in this table due to the pathway as a whole not being significantly enriched in the final dataset.)
InteractingGenes: List of other genes also selected as hits by the SIGNAL analysis that have predicted protein interactions with the assigned gene. The predicted interactions are selected based on the network criteria that was set by the user at the start of the analysis.

NetworkGenePathways: List of enriched pathways that the network of interacting genes are individually part of. Number in parenthesis indicates number of interacting genes that are members of each pathway.

Gene Hits By Iteration: This table includes all the columns from the input document with appended columns each iteration of the SIGNAL analysis. The added columns list the confidence category assigned to each gene ID at each analysis step. Columns are designated as either PathwayEnrichment or Network Enrichment followed by the iteration number. Genes counted as high confidence hits in an iteration are indicated as "HighConf", genes counted as medium confidence are indicated as "MedConf". The "SIGNALhit" column indicates the gene hits that are counted as final hits in SIGNAL. The top row highlighted in orange indicates the total number of hits considered high confidence at the end of each iteration.
(The table also includes columns with information about the pathways, interactions with other hits, and the pathway membership of the interacting genes as in the SIGNAL Gene Hits table.)

Graph: Gene Hits By Iteration: This tab includes a table and graph. The table lists the number of high confidence and medium confidence hits that are selected as hits by SIGNAL at each iteration. (The table starts at iteration 0, corresponding to the input settings. At iteration 0 only the input high confidence hits are considered high confidence, and none of the medium confidence hits are reassigned as high confidence. In subsequent iterations of the analysis some high confidence hits are dropped out and some medium confidence of hits are reassigned as high confidence. The total numbers are listed in the table.
The graph shows the above information in graphical form.

'images/IterationHits_map.png'

High Confidence Hits Not Selected By SIGNAL: This table is a list of all the hits that were not selected as hits by SIGNAL yet were assigned as high confidence hits in the input file. This table can be used to review the hits dropped out by SIGNAL to see if any of them should be manually added to the list of selected hits by SIGNAL based on the user's knowledge and discretion.

Pathway Enrichments: This table lists the enriched pathways from the analysis with the associated *p* values and gene members. This table includes all the information from the Enriched Pathways tab as well as additional columns and details. The pathway enrichment table can be used to further prioritize different subsets of the SIGNAL hit selection and to explore imputed enrichments in the data. An added feature, that can be helpful in the exploration of the data, is the Enrich Score column. The Enrichment Score provides an additional way to prioritize different enrichments beyond the provided *p values.* The enrichment score represents how strong the enrichment is (that is how many

of the genes in the pathway are also in the set of selected hits) and how much of that enrichment is driven by high scoring genes (i.e. of the genes in the hit selection set that drive the enrichment of a particular pathway, how many of them were assigned as “high confidence” in the input file). The table can be sorted by decreasing Enrichment Scores, the search bar can also be used to look for specific pathway keywords or genes.

'images/PathwayEnrichments_map.png'

The columns in the Pathway Enrichments table are:

Pathway: Name of the enriched pathway as it is labeled by the KEGG database.
pVal: The p-values based on a two-tailed fisher’s exact test for the enrichment of each pathway are listed.
pValFDR: The p-values with added correction for False Detection Rate are listed.
pValBonferonni: The p-values with the Bonferroni corrections for multiple testing are listed.
Genes: The total number of genes in the pathway.
HitGenes: The number of hit genes as selected by the SIGNAL analysis that are in the pathway.
HighScoreGenes: The number of hit genes as selected by SIGNAL that were also categorized as high confidence based on the user provided cutoff at the start of the analysis
HighScoreGeneNames: The HGNC Gene Symbols of the high score hit genes in each pathway.
MedScoreGeneNames: The HGNC Gene Symbols of the medium score hit genes in each pathway.
EnrichScore: A calculation representing the robustness of the pathway enrichments by the number of genes represented in the SIGNAL dataset and how many of them are high scoring. The EnrichScore is calculated from $\left(\frac{HitGenes}{GenesInPathway} + \frac{HighScoreGenes}{HitGenes}\right) / 2$

Network

SIGNAL can generate a unique interactive version of an integrated pathway and network graph. The integrated pathway and network figure provides an additional way to explore the enrichments and hits identified by SIGNAL. The user can select up to three pathways from the list of enriched pathways and SIGNAL will generate an interactive graph of all the SIGNAL hits (in the *Network Graph* tab; see below) that are part of the selected pathways as well as all the SIGNAL hits that have predicted interactions with the ‘pathway’ genes. This configuration is designed to facilitate an exploration of the enrichment data that goes beyond the curated annotations of the enrichment databases. The interactive and tracking features of the figure can be used to identify possible “missing links” between

different enrichments identified in the dataset. The analysis can also be used to identify novel pathway associations by identifying which targets from the study have predicted interactions with multiple members of an enriched pathway. The various features of this analysis can be used to further prioritize subsets of hits for validation and develop hypotheses for testing.

'images/SelectNetwork_map.png'

The names of the pathways are listed in a table. The pathways of interest can be selected by clicking the box beside the name. After the number of pathways have been selected, clicking the “Create Network Graph” icon at the top will generate the graph and load the next page.

Network Graph

The network graph page generates an interactive graph of enriched pathways and interacting hits based on the selections from the user in the preceding “Network” tab. The visual parameters of the graph can be adjusted by the user. The tension of the interaction edges, the text size of the gene IDs, the size of the nodes, and the text of the legend, can all be adjusted using the separate sliders on the right side of the graph.

'images/NetworkGraph_map.png'

The design uses specific groupings, color coding, and filtering to make it easier to view and interpret. A more detailed description of these settings is below:

Grouping and colors of hits: The graph separates the groups by color. Gene hits that are members of the selected pathways are shown in blue, red, and brown for the first, second, and third selected pathways, respectively. Gene hits that are annotated as members of more than one pathway selected by the user are placed in separate groups with different colors. Gene hits that are members of pathway 1 and 2 are placed between pathway 1 and pathway 2 genes and highlighted in olive (Hex Color Code #a09d01). Gene hits that are members of pathway 2 and pathway 3 are placed between the pathway 2 and pathway 3 grouped genes and are highlighted in orange (Hex Color Code #ce6702). Gene hits that are members of pathway 1 and pathway 3 are placed to the right of pathway 3 grouped genes and highlighted in saturated dark orange (Hex Color Code #6b5b3e). Gene hits that are annotated as members of all three pathways selected by the user are placed to the left of pathway 1 grouped genes and highlighted in purple (Hex Color Code #6d1c8e). Gene hits that are not annotated as members of any of the three selected pathways, yet have predicted protein-protein interactions with at least one of the hits in one of the pathways, are grouped in the lower half of the graph and highlighted in green. A legend at the top left side of the figure indicates all the relevant color labeling.

Filtering: For ease of reading the interactive network graph removes some of the total network and pathway information to make the novel and connecting relationships between hits from different pathways and hits outside the selected pathways easier to highlight. To that end, the network graph does not show predicted interaction between gene hits that are annotated as members of the same pathway. The graph also removes pathway gene hits that don't have any predicted interactions with any of the hits that are outside of the pathway. This configuration is designed to make it easier to explore the interactions suggesting novel connections between enrichment groups and hits from the analysis not annotated within the selected enrichment group.

The figure includes interactive features that are revealed and tracked based on the user's activity as described below:

Highlighting interactions: Hovering with a cursor over a specific gene ("node") highlights all the predicted interactions ("edges") of that gene.

Show only highlighted interaction: Clicking on a gene ID or node 'fixes' the highlighted interactions, so that the user can then click on one of the predicted interactions to observe the interactions from the second node.

Evidence source and confidence score of interactions: A panel at the side of the graph provides information about the highlighted interaction, such as the evidence source for the interaction and its confidence score from the STRING database. The panel also lists the number of interactions a node has within the graph. The panel also shows which confidence level the gene ID was categorized in according to the settings at the start of the analysis.

Clicking through the graph to map a novel pathway: In addition to highlighting different interactions, specific interactions can be clicked and resultant pathways mapped.

'images/HighlightNetwork_map.png'

After clicking through a string of interacting genes the user can click the "Highlight Clicked Pathway" icon and all the genes clicked through in the exploration, together with their predicted interactions, are highlighted. The pathways they are members of are also listed in a separate table that can be downloaded with the rest of the analysis at the *Download* tab.

To drive further exploration of the data, the SIGNAL platform makes it possible to view which of the gene hits identified by the SIGNAL analysis that are not known members of specific gene sets ("Novel" genes) have predicted interactions with known members of the gene sets ("pathways")

For ease of viewing the network graph is generated with two viewing options and two information tables.

1st Degree Network: A circular graph showing the hit genes from each selected pathway (separated by group and highlighted by different colors) and only the "novel" genes from the analysis that have predicted interactions with any of the "pathway genes". Hovering with the cursor over a gene name will highlight its interaction path and the name of the gene(s) it is predicted to interact with. As described above, this feature can be used to generate exploratory hypothesis of novel mechanisms and interactions and to identify "missing links" between predicted biological processes generated by the analysis to be further validated by subsequent research.

2nd Degree Network: A circular network like the 1st Degree Network that also includes "novel" genes that don't show 1st degree connectivity with genes in the pathways of interest but show 2nd degree connections to the pathways via predicted interactions with other novel genes identified by SIGNAL that have predicted direct interactions with the pathways of interest. This feature can be used for more complex exploration of possible network reconstruction and to broaden the targets for further exploration.

Network Table: The Network table provides the tabulated information for the pathway-based network query selected by the user in the "Network" panel. The Network table lists all the SIGNAL hit genes that have primary or secondary predicted interactions with SIGNAL hits in the pathways (selected by the user) with the following additional information:

'images/NetworkTable_map.png'

Group: Which of the (up to) three pathways selected by the user the gene is a member of. If none, the gene is grouped as a "Novel" interactor with these pathways.
Pathway: Which of all of the enriched pathways in the SIGNAL analysis the gene is a member of.
Allnet.count: the number of other SIGNAL hit genes the gene has predicted interactions with (based on the user set criteria at the start)
Ntwrk.all: The gene names of the genes that have predicted interactions with the individual gene.
NtwrkCount.[Name of first selected pathway]: The number of genes in the first selected pathway that are hits by SIGNAL that have predicted interactions with the specific gene target.
Ntwrk.[Name of first selected pathway]: The gene names of the genes from the first selected pathways that are hits by SIGNAL that have predicted interactions with the specific gene.

NtwrkCount.[Name of second selected pathway]: The number of genes in the second selected pathway that are hits by SIGNAL that have predicted interactions with the specific gene target.
Ntwrk.[Name of second selected pathway]: The gene names of the genes from the second selected pathways that are hits by SIGNAL that have predicted interactions with the specific gene.
NtwrkCount.[Name of third selected pathway]: The number of genes in the third selected pathway that are hits by SIGNAL that have predicted interactions with the specific gene target.
Ntwrk.[Name of third selected pathway]: The gene names of the genes from the third selected pathways that are hits by SIGNAL that have predicted interactions with the specific gene.
Total_Path_Hits.net.count: The total number of hits by SIGNAL that are members of all of the selected pathways that have predicted interactions with the specific gene.

Clicked Pathways Table: The clicked pathways table highlights the gene ("nodes) and interactions ("edges") that the user followed and clicked on in the interactive network graph.

'images/SelectedPathway_map.png'

Name1: Gene Symbol of the clicked node.
Node1: Node title (group name followed by Gene Symbols).
Parent1: Name of the group the node is in (Pathway name or "Novel" if it is outside one of the selected pathways.)
Interactions1: Number of predicted interactions between the node and other hits selected by the analysis.
Screen.Input: The confidence category this gene was assigned to at the start of the analysis.
Name2, Node2, Parent2, Interactions2: Gene symbol, node title, group, and number of predicted interactions of the second node.
Weight: Assigned confidence weight for the predicted interaction in the STRING database.
Source: Evidence source for the predicted interaction in the STRING database.

Both graphs can be viewed as interactive HTMLs and screen grabbed for saving the image. .csv files of the Network Table and Clicked Pathway Table can be downloaded from the "Download" tab.

Note: the "Clicked Pathway Table" resets every time the clicked pathway is reset in the graph. To save a specific clicked table before trying a different path click on the download

folder to download the files before clicking "Click to reset" to save the most recent clicked table.

5. Saving and securing your analysis:

The Download tab contains a list of all the files generated by the analysis and available to download. The "Download all files" icon downloads a zipped folder of all the listed files in .csv format.

As more analysis are added within the same session, the analysis files are added to the zipped folder for download. Once the session ends the analysis files are deleted.

All the analysis download files begin with the same file name as the file uploaded by the user at the beginning of the analysis with the different titles appended to the end.

'image/Download_map.png'

The file names and their contents are:

[name of input file]_SIGNALhits.csv: A table listing all the hits by SIGNAL analysis with enriched pathway and interacting network information.

[name of input file]_SIGNALenrichment.csv: A table of enriched pathways identified by the analysis with their associate p-values, FDR values, Bonferroni values, and high confidence and medium confidence hit genes included.

[name of input file]_ nonSIGNALhits.csv: a table listing all the genes that were assigned as high confidence hits at the start of the analysis (by the user provided criteria) but were not selected as hits by the end of the SIGNAL analysis.

[name of input file]_SIGNALnetwork_[Name(s) of selected pathways].csv: A table of all the hit genes selected by SIGNAL analysis with their predicted membership or interactions with genes in the pathways selected under the "Network" tab.

[name of input file] _Clicked_Pathways.csv: A table of all the genes clicked on by the user in the "Highlight Clicked Network" feature under the network tab.

Unmapped_rows: A table of all the rows with GeneSymbols that couldn't be mapped to EntrezIDs. (This document only shows up when the upload file only included a GeneSymbol column and EntrezID mapping was done by SIGNAL).

Appendices:
Appendix A: How to segment data into confidence tiers
Appendix B: Data security
Appendix C: Running SIGNAL with alternative or bespoke databases

Appendix A: How to segment data into confidence tiers

There are three ways that a dataset can be split into high-confidence, medium confidence, and low confidence/non hits sets. Different datasets require different approaches and different ratios.

While there's no rule as to what a good number of candidates is to be assigned to each of the data tiers, in our tests we have found that a 1:2 ratio of high confidence hits to medium confidence hits (with all other candidates serving as background/non-hits) work best (i.e. ~400 high confidence hits, ~800 medium confidence hits).

(The pathway enrichment results from the analysis can serve as 'gut check' for whether some of the expected enrichments show up. The rapid speed with which each analysis is completed on the platform makes it possible to try different cutoffs and different approaches to see which gives a recognizable yet informative result.)

Data segmentation can be achieved using a single readout, dual measurements, or multiple assays.

Data can be segmented into confidence scores either by using the normalized readout from an assay or by combining readouts and studies.

Below is a guide for how to set cutoffs and assign confidence to tiers using three different approaches:

A. Using a single readout or value:

This approach is most recommended when using readouts like Z score or Expression that have already been normalized and corrected for outliers.

'images/Segment_1criteria.png'

To segment data into three tiers based on a single readout, simply assign those values to a specific column in your input file. Choose a "stringent" cutoff for the

"High Confidence Cutoff Value" in SIGNAL and a more "lenient" cutoff for the "Medium Confidence Cutoff Value" in SIGNAL.

B. Using two readouts or values:

This approach is most recommended when using values like Fold Change where an additional readout such as the replicate p Value is helpful to increase confidence in the selected genes.

'images/Segment_2criteria.png'

To segment data into three tiers based on two readouts, choose the readout that will be broken into a "stringent" cutoff and a "lenient" cutoff (Fold Change for example) assign those values to a specific column in your input file. Choose a "stringent" cutoff for the "High Confidence Cutoff Value" in SIGNAL and a more "lenient" cutoff for "Medium Confidence Cutoff Value" in SIGNAL.
Assign the second set of values to a different column in your input file. Choose a cutoff which all high confidence and medium confidence hits must meet and enter those values into the fields under "Add an Additional Criteria".

C. Assigning a distinct value to each gene before uploading to SIGNAL and using the assigned value to separate the tiers:

This approach is most recommended when using multiple criteria or when combining the results of multiple studies. While the SIGNAL sidebar menu doesn't have a direct option to combine these studies the information can be combined outside of SIGNAL (such as in Excel or in R) and then run as a SIGNAL analysis.

'images/Segment_3criteria.png'

In Excel or in R, create a new column in the file you upload. In the new column, assign a value of 1 to gene candidates that meet the high confidence criteria (i.e. the gene candidate is a hit in more than one of the studies or in all the studies being compared). Then assign a value of 0.5 in the same column to all the gene candidates that meet the criteria for medium confidence (i.e. the gene candidate is a hit in just one or two of the studies being compared.). Assign a value of 0 in the column to all the other gene candidates that don't meet any of the criteria you've set.

A similar approach can be used when selecting hits based on multiple guides in a gene perturbation screen.

Once the values have been added upload your file to SIGNAL. Under "Cutoff Type" select the name of the column in which you have added the new values. In the

field “High Confidence Cutoff Value” enter the value 1, in the field below it (“Medium Confidence Cutoff Value”) enter the value 0.5. This will run the iterative analysis using the high, medium, and low confidence tiers that you’ve assigned outside of SIGNAL.

Appendix B: Data security

The SIGNAL application web interface is hosted by the National Institute of Allergy and Infectious Disease (NIAID) Office of Cyber Infrastructure and Computational Biology (OCICB) (https://www.niaid.nih.gov/about/cyber-infrastructure-computational-biology-contacts). The analysis of uploaded data is run behind two internet security firewalls. Access by external users to the SIGNAL interface pass through two secure firewalls. The incoming request first passes through the NIH web hosting firewall after which the analysis and SIGNAL application go through the NIAID firewall where the analysis is hosted. Migrations to external websites outside of the firewall (i.e. the KEGG interface) are accompanied with a warning message for the user.

SIGNAL uses a secure encrypted HTTPS connection, and all requests are handled using encrypted connections. Using these encrypted connections, only the browser from the IP address where the request originated from can access the data generated and uploaded. When a file is uploaded, a new unique directory is created where the input file and all the subsequently generated analysis files are temporarily saved, ensuring that only the user generating the connection can access the directory with their files. The directory is kept on the server only for the duration of the session (i.e. as long as the user is using the site). Once the sessions ends (i.e. close of browser window or move to a new site) the directory and all its files are removed from the SIGNAL server. This decreases the security risk for the user and ensures that the results are only stored locally after the analysis.

Data collected for each session is limited to the country where the request comes from and the time spent using the site (and specific pages). File names, analysis choices, user IDs, and results are neither collected nor stored.

An additional option to increase the security of the data and generated analysis is to download SIGNAL source code and run the application locally on your own device see "Appendix C: Running SIGNAL with alternative or bespoke databases".

For any further questions about data security, please reach out to us via the ‘Contact us’ tab under the “Help” tab.

Appendix C: Running SIGNAL with alternative or bespoke databases

To make SIGNAL an adaptable framework for iterative analysis with different datasets and databases beyond the databases and settings used on this platform, an R script version of a standalone SIGNAL function can be downloaded. The SIGNAL function relies on calling two separate analysis function, a pathway enrichment function and a network analysis function. The master SIGNAL function applies the pathway and network function iteratively, and the results are tested for when the analysis converges on a single set.

The list of input variables that can be selectively assigned in the adaptable SIGNAL function in R and their required formats are:

screen.datafame: A data frame of the screen.
ID.column: A column within the screen.dataframe for the identifiers of the targets (EntrezID, GeneSymbol, etc.).
criteria.column: A column within the screen.dataframe of the criteria for being considered a hit.
highconf.criteria: A criteria each target has to meet to be considered a "high confidence" hit.
midconf.criteria: A criteria each target has to meet to be considered a "mid confidence" hit.
criteria.setting: Whether the function should be using "equal", "greater than or equal", or "less than or equal" when assessing if confidence criteria are met. criteria.setting input should be in the format of "equal", "greater", or "less".
enrichment.dataframe: A data frame to be used for pathway membership in the format of a column of IDs (should be same as ID column in screen.dataframe in ID type and column title) and a column of which group they are part of (each ID~group relationship needs to be in its own separate row).
enrichment.title: Name of the column with the names of the enrichment groups the targets are members of.
stat.test: Name of the statistical test to be used for measuring enrichment confidence. Needs to be in the format of either "pVal", "FDR", or "Bonferroni".
test.cutoff: A numeric value which a less than value in stat.test will be considered a significant enrichment.
network.igraph: an igraph of the network to be used for network analysis (network igraph must use the same ID type as screen.dataframe)

The user provided variables are then used to apply the iterative function as in the previous paragraph. The adaptable version of SIGNAL broadens the scope of its application beyond the use of the specific databases and settings it was designed in.

The SIGNAL function provides an output in the format of a R script list that contains three data frames:

1. The input data frame with an appended 'SIGNAL.hit' column.
2. A data frame of high confidence and medium confidence designation at each iteration of the analysis.

3. A data frame of final SIGNAL enrichments from the provided enrichment data frame.