Documentation:

Using TRIAGE:

Overview:

The Throughput Ranking by Iterative Analysis of Genomic Enrichment (TRIAGE) platform is designed to facilitate prediction, analysis, and hypothesis generation from genome-wide perturbation studies like those designed with RNAi and CRISPR technologies.

TRIAGE is best utilized as tool for selecting candidates from high-throughput studies to be further validated by rigorous low throughput follow up studies. TRIAGE integrates analysis from gene to gene predicted interactions and statistical enrichments of known gene sets to correct for the false positive and false negative error rates that are associated with normalization methods that require arbitrary cutoffs. By utilizing data sorted into low, medium and high confidence tiers the TRIAGE platform integrates the experimental readout of the analyzed study with *a priori* knowledge about biological networks and interactions.

Getting Started

Uploading Your Data:

To upload your data ensure that your document is in .csv format and includes at least one of these two columns:

1. A column titled “GeneSymbol” with the HGNC gene names of your gene targets.
2. A column titled “EntrezID” with the NCBI entrez ID of your gene targets.

Your document should also include a column with the numeric values by which medium and high confidence values will be selected. Each gene target should only have one value associated with it and listed in the document only once. The name of this column is up to the user. (Additional columns can be included in the document and they will be ignored. Once a file has been uploaded a drop down bar will appear with a list of all the column names from the uploaded document. The user then selects which numeric column should be use to assign one of the three confidence scores based on the cutoffs provided by the user.

Running Your Analysis:

To run your analysis select the parameters that describe your data and the databases you want to use for your analysis.

Select your organism: A drop down menu provides the option to select “Human” or “Mouse”

Select a Database for Enrichment Analysis: The enrichment analysis uses the pathway designations curated by the Kyoto Encyclopedia of Genes and Genomes (KEGG). The user can select the option *KEGG: Biological Processes* which uses only the pathways in the KEGG database that describe biological processes, *KEGG: Disease Pathways* which exclusively uses only the pathways associated with disease description on the KEGG database, or *KEGG: All Pathways* which includes all of the pathways curated in the KEGG database. The default setting is to *KEGG: Biological Processes* and is the recommended option.

Select Interactions for Network Analysis: For the network analysis component TRIAGE 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 with evidence 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 the possible criteria see http://string-db.org/cgi/help.pl?sessionId=Z6t6X5gizxo0)

Interaction Confidence for Network Analysis: The user can select what confidence score (as defined on the STRING database) to consider for the network interactions used in the analysis. The default option is set to Medium (>0.4) corresponding to a score of 400 or greater on the STRING database, with the option to select a lower (>0.15) or more stringent (>0.7) cutoff.

Choose an input file to upload: Upload your .csv file. A progress bar will inform you when your upload is complete, the data from your file will appear in a table under the “Input” tab.

Cutoff Type: Once your file is uploaded a drop down menu will appear with a list of the column names in your document. Select the column that contains the numeric values to be used for the high confidence/medium confidence cutoffs of your targets.

High-conf Cutoff Value & **Med-conf Cutoff Value:** Enter 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 and Med-confidence cutoff value TRIAGE will infer whether the values should be taken as “greater than or equal to” or “less than or equal to”

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 all the “no confidence hits” not included in the results. The added background will not appear as suggested hits by the analysis only to be used as a means to have more robust statistics about the enrichment of pathways. The background genomes use only the known protein coding genes of the selected organisms (Human: HGNC, Mouse: MGI).

Analyze my data: Once you have selected and entered the parameters for your analysis click the "Analyze my data" icon and your analysis will begin. A progress bar at the bottom right corner of your screen will show the progress of your 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.

Reading Your Results:

Enriched Pathways

When the analysis is complete a list of enriched pathways will appear in a table under the "Enriched Pathways" tab. The list includes all pathways that have a statistical score of 0.05 or less in a two-tailed fisher's exact test.

The table provides the following information:
Pathway: The name of the enriched pathway. 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.

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.

TotalGenes: The total number of genes in the pathway.

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

HitGeneNames: The HGNC Gene Symbols of the hit genes in each pathway, 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.

Gene Hits

TRIAGE Gene Hits: This table provides a list of prioritized hits selected by the TRIAGE analysis with supporting information on interacting genes (based on the user selected network criteria) and membership in enriched pathways

EntrezID: NCBI EntrezID identifier.
GeneSymbol: HGNC official Gene Symbol.
ConfidenceCategory: "HighConf" or "MedConf" based on the original category it was assigned to at the start of the analysis by the user provided cutoffs.
TRIAGEhit: "Yes" indicates a hit selected by the TRIAGE analysis.
Pathway: Names of enriched pathways from the analysis which the select gene is a member of.
InteractingGenes: List of genes that are also hits by the TRIAGE analysis that have predicted protein interactions with the specific gene based on the network criteria set by the user at the start of the analysis.
NetworkGenePathways: List of enriched pathways that 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: A table with the input document and the genes added or dropped at each iteration listed in new columns. Genes counted as high confidence hits in an iteration are indicated as "HighConf", genes counted as medium confidence are indicated as "MedConf". The "TRIAGEhit" column indicates the gene hits that are counted as final hits in TRIAGE. 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 TRIAGE Gene Hits table

Graph : Gene Hits By Iteration: A graph showing the number of medium confidence hits and high confidence hits that are selected as TRIAGE hits at each iteration of the TRIAGE analysis. (starting with 0 which is the input file whereby only the original high confidence hits are included).

Pathway Enrichments: This table lists the enriched pathways from the analysis with statistical cutoffs and the gene candidates that drive the enrichment

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 TRIAGE analysis that are in the pathway.
HighScoreGenes: The number of hit genes as selected by TRIAGE 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.
EnirchScore: A calculation representing the robustness of the pathways enrichments by the number of genes represented in the TRIAGE dataset and how many of them are high scoring. The EnrichScore is calculated as $\left(\frac{HitGenes}{GenesInPathway} + \frac{HighScoreGenes}{HitGenes}\right) / 2$

Network

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

To explore this feature the user can select up to three pathways by clicking on the box near the pathway name, once the pathways of interest are selected click on the "Create Network Graph!" icon. A progress bar at the bottom right of the screen will appear

PathNet

Once the graph has been generated the interface will switch to the PathNet tab with two viewing options and an information table.

1st Degreee 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. 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 TRIAGE 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.

PathNetTable: The PathNet table provides the tabulated information for the pathways based network query selected by the user in the "Network" panel. The PathNet table lists all the

TRIAGE hit genes that have primary or secondary predicted interactions with TRIAGE hits in the pathways selected by the user with the following additional information:

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 TRIAGE analysis the gene is a member of.
Allnet.count: the number of other TRIAGE 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 TRIAGE 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 TRIAGE 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 TRIAGE 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 TRIAGE 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 TRIAGE 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 TRIAGE that have predicted interactions with the specific gene.
Total_Path_Hits.net.count: The total number of hit by TRIAGE that are members of all of the selected pathways that have predicted interactions with the specific gene.

Both graphs can be downloaded as interactive HTML files from the "Download" tab as well as a csv file of the PathNet Table.

Download

The "Download all files" icon downloads a zipped folder of the analysis generated by the TRIAGE platform.

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

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

[name of input file]_TRIAGEoutput_ALL.csv: a table listing all the genes from the analysis with their ranking in each iteration, the final iteration showing the list of genes that are selected by the analysis.

PathNet_[name of input file]_[Name(s) of selected pathways]1Degree.html: An HTML file of the graph generated in the 1st Degree Network analysis.

PathNet_[name of input file]_[Name(s) of selected pathways]2Degree.html: An HTML file of the graph generated in the 2nd Degree Network analysis.

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

Help

Contact us: You can contact the TRIAGE team with any issues you encounter by filling out the form under this tab.

Documentation: User guide and associated information. (forthcoming)

Updates: (forthcoming)