Gene Set Characterization Pipeline

QUICKSTART GUIDE

Updated: 08/30/2017

In this pipeline, the user submits a set of genes to determine if they are enriched for a pathway, a Gene Ontology term, or other types of annotations. This pipeline tests your gene set for enrichment against a large compendium of annotations. Gene Set Characterization can be done using a standard statistical test or in a Knowledge Network-guided mode (using DRaWR, PMID: 27153592)

Pipeline Selection

1. Click on “Start A New Pipeline” on the homepage or on “Analysis Pipelines” at the top.
2. Hover the cursor over “Gene Set Characterization” and click on “Start Pipeline” next to it.

Data Files

3. Click on the **“**Select species**”** dropdown**.** This quickstart demo run uses a human gene set, so keep the default setting of “Human”.
4. Upload your gene set(s):
    a. Click on the “Use Demo Data” button. The file “demo_GSC.list.txt” will immediately begin loading into the data table.
    b. Once the demo file appears in the data table, make sure the checkbox to the right of the filename is selected (it should be checked automatically when you use demo data).
    c. Click “Next” at the bottom right corner.

Parameters

5. Choose public gene sets: This page allows you to choose collections of public gene sets that represent gene properties and annotations that will be tested for enrichment in your gene set. Select the following six collections:
    a. *Ontologies: Gene Ontology*
    b. *Pathways: Enrichr Pathway Membership*
    c. *Tissue Expression: GEO Expression Set*
    d. *Disease/Drug:* KEA Kinase Signatures
    e. *Disease/Drug: MSigDB Chemical and Genetic Perturbations*
    f. *Protein Domains: PANTHER Classification*
6. Click on “Next”

Knowledge Network

7. On this page you are indicating if you want to use the Knowledge

Demo File & Acceptable File Formats

Demo files are available within the data selection page for each pipeline setup.

Currently the Platform accepts tsv formatted files only.

About the Demo File:

The sample file for this demo lists a set of 125 genes associated with breast cancer downloaded from the Enrichr database.

Uploading Spreadsheets with Multiple Gene Sets

You may also select or upload a tab-separated spreadsheet file. This allows you to specify multiple gene sets (one per column) at the same time, for parallel analysis.

The first row (header) should should contain the names of the gene sets in the corresponding columns. The first column of the spreadsheet should be the gene identifiers corresponding to each row.

For each entry in the spreadsheet table, a “1” indicates that the corresponding row gene is part of the corresponding column gene set, a “0” means it is not.

Network in your analysis. For this demo, we'll select "Yes" then select the network "'STRING Co-expression", keep the Network Smoothing at its default settings, and click "Next".

Review & Submission

8. This page is a summary of various choices you've made until now, and you can go ahead and click "Submit Job". The page should immediately tell you the job is "Running !" From here you have the option to "Start a new Pipeline" or "Go to Results Page". Click on "Go to Results Page".

Results and Visualization

10. You should now see a listing of all jobs you've run, with the latest one at the top. (It should have the name "gene_set_characterization_<date>"). Wait for a few minutes until the status column for this entry turns to a green check. Click on the row in the table and then on the "View Results" button in the right panel. (You can also double-click on the job name in the Results table to open the visualization view.)
11. The main panel shows a grid view of gene sets (rows) and properties (columns), with colored cells indicating an association detected between that gene set and property. Since this sample analysis was done with one gene set, the grid has only one row.
12. You may use the slider called "Filter matches by score"' to increase or decrease the number of associations shown.
13. Using the left panel, you can select which groups of annotations (e.g., "Pathways", "Ontologies", etc.) you want to include in the grid view.

Results Download

13. Click on "Results" at the top of the page. You should see a listing of the various jobs you've run, with the current job at the top of the list. Click on it and then click the Download Icon in the right panel. Wait a few moments until a .zip archive is downloaded.
14. In the downloaded archive, there is a directory for each of the six public gene set collections we chose for our run. Each directory contains a "*.df" file that contains the results of the analysis with that public collection. The first two columns show the name of the user submitted gene set and the KnowEnG internal id of the public gene set. The third column is the score of the association. With DRaWR, this is a normalized difference between the query and baseline probabilities with the best score observed as one. With Fisher Exact test, this is the negative log10 p-value of the one-sided statistical test. There is also an "*.edge" file which describes the gene sets by their Ensembl gene ids. The file README-GSC has more information about the downloaded results.

Using the Knowledge Network

If you don't wish to use the Knowledge Network, select "No" and go to the next step.

In a nutshell, using the Knowledge Network widens your ability to infer connections via network neighbors for sparsely annotated genes.

Advanced Information

Additional information about using the Knowledge Network is available in the Info Panel on the right side of each page.

If it's not visible click on the "?" to the right of each page.