Sample Clustering Pipeline

QUICKSTART GUIDE

knoweng

Updated: 08/30/2017

This tutorial describes running the Sample Clustering Pipeline on the KnowEnG Platform. With this pipeline, users can find clusters of samples that have similar genomic signatures, such as cancer patient subtypes. If you also have phenotypic descriptions for each sample, e.g treatment outcomes, this pipeline can identify phenotypes that are highly correlated with each cluster. Sample clustering can be done in a Knowledge Network-guided mode, and with optional use of bootstrapping to achieve robust cluster assignment.

Pipeline Selection

1. Click on "Start A New Pipeline" on the homepage or on "Analysis Pipelines" in the main navigation.
2. Hover the cursor over "Sample Clustering" and click on "Start Pipeline" next to it.

Data Files

3. "Select species": This demo run uses human cancer data, so keep the default setting of "Human".
4. ***Upload "omics"spreadsheet***
    a. Click on the "Use Demo Data" button. The file "demo_SC.genomic.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.
5. ***Upload phenotype spreadsheet***
    a. Click on the "Use Demo Data" button. The file "demo_SC.phenotypic.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"

About the Demo File: Genomic Spreadsheet

The genomics file for the sample clustering demo contains gene-level non-synonymous mutations of 381 lung adenocarcinoma cancer patients from TCGA. *You can use a spreadsheet software such as Excel to view it locally if you are curious.*

About the Demo File: Phenotypic Spreadsheet

The phenotypic file for this demo contains values of 10 different phenotypes for many of the lung cancer patients in the transcriptomics demo file. The phenotypes here are descriptions of cancer stage, days survival, smoking status, etc.

Parameters

6. On this page keep the default final clustering algorithm set at, "k-means". Change the number of clusters to "6" in the "Enter the

number of clusters you wish the analysis to return" textbox, and click "Next".

Knowledge Network

7. Select "Yes" to use the Knowledge Network, select the network "HumanNet Integrated Network", keep the network smoothing at its default settings, and click "Next".

Bootstrapping

8. Select "Yes" to use bootstrapping, keep the default settings otherwise and click "Next".

Review & Submission

9. 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. Then click on it and then on "View Results" 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 heatmap of the consensus clustering. The rows and columns are samples and the cells indicate the percentage of time two samples were clustered together in the same bootstrap. The cluster sizes are reported in the table on the right.
12. You may use the 'ADD PHENOTYPES TO COMPARE' feature under the heatmap to add sample phenotypes to the visualization. Each phenotype is visualized using a mini-heatmap that is aligned with the main heatmap above and represents the same samples in the same order. A p-value of significance is on the left and the heatmap color legend is on the right of the mini heatmap. Click on the legend to display the values for the phenotype. With these mini-heatmaps, you can visually assess whether patterns in the values correspond to the clusters in the heatmap.

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. 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, you can download the sample to cluster_id mapping ('cluster_labels.tsv'), the values of the consensus matrix ('concensus_matrix.tsv'), and a file mapping the input gene ids to their internal KnowEnG identifiers ('gene_name_mapping.tsv'). There is also a files called top_gene_by_cluster.tsv that is a binary valued spreadsheet with gene names as rows and cluster ids as columns that indicates if a gene was in the top 500 important genes for the cluster. The file README-SC 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.

Advanced Information

Additional information about making Knowledge Network and Bootstrapping selections 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.