# Sample Clustering Pipeline

Cancer Genomics Cloud (CGC) version
QUICKSTART GUIDE

knoweng

Updated: 01/29/2018

This Quickstart guide describes how to run the Sample Clustering (SC) pipeline/workflow on the Cancer Genomics Cloud (CGC).

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.

## Initial Steps

1. Login to the CGC at https://cgc.sbgenomics.com/.

   If you don't have a CGC account, create one using the "Create a free account" link on this page just below the login box.
2. Click on the project's link to go to the project dashboard.
   If you don't have a project to use, create one using the "Create a project" button.
3. Click on the "Apps" tab on the project dashboard.
   If you don't have the KnowEnG Samples Clustering Workflow app in your project, add it using the "Add app" button:
   a. *Click on the green "Add app" button.*
   b. *Search for the app, e.g., enter "samples clustering" in the search box.*
   c. *Click on the "Copy" button in the app box.*
   d. *Click on the green "Copy" button.*
   e. *Click on the "X" in the top right of the window to close it.*
4. Click on the KnowEnG Samples Clustering Workflow link to go to the app page and view information about the app.

### Running Apps

There are actually several ways to run apps on the CGC; for example, on the Apps tab, each app has some Actions, one of which is a green arrow run button; in addition, if you have run the app before, you can go to the Tasks tab, and each task includes an Action to re-run the task; and if you are on a task page, there is a green "Edit and rerun" button.

## Run the App

5. Click on the Run button to run the app.

## Data Files

6. These instructions use sample input files already available in this project (demo_SC.genomic.txt and demo_SC.phenotypic.txt). For information on uploading your own data to the CGC, see the section "Uploading Files to the CGC" further down in this document.
   a. ***Genomic Spreadsheet File**:*
      i. *Click on the "Select file(s)" button.*
      ii. *Enter "demo" in the search box.*
      iii. *Find the input file in the list of files returned (the file's name is "demo_SC.genomic.txt").*
      iv. *Click on the checkbox to the left of the file name.*
      v. *Click on the "Save selction" button.*
   b. ***Phenotypic Spreadsheet File**:*
      i. *Click on the "Select file(s)" button.*
      ii. *Enter "demo" in the search box.*
      iii. *Find the input file in the list of files returned (the file's name is "demo_SC.phenotypic.txt").*
      iv. *Click on the checkbox to the left of the file name.*
      v. *Click on the "Save selction" button.*

### **About the Demo File**: Genomic spreadsheet

The genomic spreadsheet preprocessed for this demo, demo_SC.genomic.txt, 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 sample phenotypic file for this demo, demo_SC.phenotypic.txt, 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

7. Click on the "Next: Define App Settings" button (or the "Define App Settings" tab).
   The parameters are listed below in their suggested order of entry, although they may be listed in a different order on this page.
   For each parameter, the "?" link contains a description of what the parameter is used for and what type of value is expected.
   a. ***Number of Clusters**: enter an integer value. This value is required.*
   b. ***Number of Top Genes**: enter an integer value.*
      *This value is optional; if not specified, the default value is 100.*
   c. ***Knowledge Network Edge Type**: sample values are listed at*

*KN Contents by Gene-Gene Edge Type (the values are listed within parentheses); e.g., STRING_experimental" is a sample value. Leave this value blank if you do not wish to use the Knowledge Network.*

d. ***Species Taxon ID**: sample values are listed at KN Contents by Species (the values are listed within parentheses); e.g., "9606" (for human) is a sample value.*
*(Species Taxon ID is listed twice on this page, but the input boxes are linked together so that when you enter the value in one text box it appears in the other.)*
*(This value is only relevant if the Knowledge Network is used.)*

e. ***Amount of Network Influence**: enter an integer value between 0 and 100 (inclusive).*
*This value is optional; if you do not enter a value, 50% will be used. A greater value means greater contribution from the network interactions. (This value is only relevant if the Knowledge Network is used.)*

f. ***Number of Bootstraps**: enter an integer value; if you want to use bootstrapping, enter a value greater than 1.*
*This value is optional; if not specified, the default value is 0 (i.e., no bootstrapping).*

g. ***Bootstrap Sample Percent**: enter an integer value between 0 and 100 (inclusive).*
*This value is optional; if you do not enter a value, 80% will be used. (This value is only relevant if bootstrapping is done.)*

h. ***Processing Method**: choose a value from the pull-down menu; the possible values are **serial** and **parallel**.*
*This value is optional; if not specified, the default value is **serial**. (This value is only relevant if bootstrapping is done.)*

### Launch the App

8. When the parameters are all entered, click on the green Run button to start the workflow running.

   The app generally takes a few minutes to run (e.g., 4 to 9 minutes).

### Results and Download

9. When the task is finished, the task page will show a green "COMPLETED" image to the left of the task name. The page shows the input files, the parameter settings, and the output files.

   The output files include the **Sample Labels** (‘sample_labels_by_cluster.txt’), the **Consensus Matrix** (‘consensus_matrix.txt’), and the **Gene Map File**, a file mapping the input gene ids to their internal KnowEnG identifiers (‘gene_map.txt’). There is also the **Top Genes File** (`top_genes_by_cluster.txt`), 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. There are also several output files that give information about the run. The **README** file has more information about the output files.

   To view an output file, click on the file name, and scroll down to the bottom of the page to view the file's contents (or just a portion of the contents, for large files). (You may need to click on the “Display raw data” button to view the contents.) From that page, you can also Download the file (the "Download" link is available under the "..." "More actions" button).

### Using the Knowledge Network

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

If you do not wish to use the Knowledge Network, simply leave the Knowledge Network Edge Type parameter blank.

### Links

KnowEnG Main Website
KnowEnG Analytics
Knowledge Network Overview
Knowledge-Guided Pipelines
SC Pipeline
Pipeline Quickstart Guides
SC Pipeline Quickstart
KnowEnG YouTube Channel

You can also download a file, or multiple files, one at a time, from the task page by clicking the folder image "Browse files" button next to a particular file, or the "Outputs" header.

### Uploading Files to the CGC

There are several ways to upload files to the CGC, as described on this page:

https://docs.cancergenomicscloud.org/docs/upload-to-the-cgc

The best way will depend on where the files are, how you access them.

If they are on your personal computer, you can use the CGC Uploader GUI tool.

If they are on a server, you can use the Command Line Uploader.

If they are accessible via FTP or the web, you can use the FTP/HTTP(S) import tool.

And you can also upload from a cloud volume, such as Amazon Web Services (AWS) or Google Cloud Storage (GCS).