

GSCA User Interface Manual

Zhicheng Ji, Hongkai Ji

January 18, 2015

1 GSCA GUI Overview

The GSCA GUI consists of two main parts: sidebar panel on the left side of the GUI where users can give inputs and specify options, and main panel where plots and GSCA results will appear.

To perform a typical GSCA analysis in the GUI, users should go through four major analysis steps: input gene set, select gene set and compendium, run GSCA analysis and save results. Users can switch between these steps using the main menu on the top-left corner of the UI.

2 Input Gene Set

The first step is to input gene sets into GSCA. Users should first input gene set name for each individual gene set in "Gene Set Name" text input area. Note that the name of the gene set should be easy to remember and does not need to follow any other specific rule. After that, users can choose to directly type in the Entrez Gene ID using keyboard or upload prepared gene set file.

Choose "Specify Gene ID" option to directly type in the Entrez Gene ID. The positive genes (e.g. genes activated by a transcription factor) should go into the "Specify Entrez GeneID for Positive Genes" text box and the negative genes (e.g. genes repressed by a transcription factor) should go into the "Specify Entrez GeneID for Negative Genes" text box. Users should separate different genes with ;. For example, if geneID 10 and 100 are positive genes and geneID 1000 is a negative gene in the gene set, 10;100 should be given in "Specify Entrez GeneID for Positive Genes" text input field and 1000 should be given in "Specify Entrez GeneID for Negative Genes" field. For this input method the weights for all positive genes are 1 and weights for all negative genes are -1.

Choose "Upload Gene set File" option to upload prepared gene set file. The prepared geneset file can contained two or three columns. The first column should be ENTREZ geneID, the second column should be weights (numeric values) and the optional third column should be the gene set names. A short instruction of how to prepare gene set file can be found on the main panel. Click "Choose File" button to select the file path. Users can decide whether to include header, change the separator and change the quote. If nothing appears under "Current gene set" section on the main panel, users should consider changing the options and make sure that the gene set file is in the correct format.

After the input process using either one of the two input methods is finished, users can click "Add Gene Set" button to add the gene set into GSCA. The

main panel showing 1. the current gene set being inputted 2. names of all gene sets already entered in GSCA. Users can also find other two tabs "Gene Set Summary" and "Gene Set Details" above the main panel. "Gene Set Summary" tab provides an overview of all gene sets in GSCA including the gene set name and number of positive and negative genes in each gene set. In "Gene Set Details" tab users can view all genes included in each gene set. Please make sure that all the input gene sets are correctly specified before further analysis steps.

Users can save the current gene sets in GSCA by clicking the "Save" button under "Save current gene sets as csv file". By this way users can conveniently restore all the gene sets when they start a new GSCA analysis next time. If user want to delete one of multiple gene sets, users can go to "Delete existing gene set" panel, choose the gene sets to be deleted and click "Delete Selected Gene Set" button to delete these genes. To start a brand new GSCA analysis, users can click "Reset All Gene Sets" button to delete all existing gene sets.

3 Select Gene Set and Compendium

Select "Select Gene Set and Compendium" in the main menu. On the sidebar panel users should first select the gene sets they want to include in GSCA analysis. By default all the input gene sets will be included in the GSCA analysis. To switch the order of gene sets so they will appear in correct order in GSCA analysis, users can check the "Switch gene set order" option and switch the order of gene sets by choosing the names of the two gene sets to be switched and clicking "Switch" button.

Then users should select with which compendium they want to perform GSCA analysis. Users can either select one of the compendia provided by GSCA or upload their own compendium. Currently there are four preprocessed compendia available: Affymetrix Human Genome U133A Array (GPL96); Affymetrix Mouse Genome 430 2.0 Array (GPL1261); Affymetrix Human Genome U133 Plus 2.0 Array (GPL570); Affymetrix Human Genome U133A 2.0 Array (GPL571). Users can go to the NCBI GEO description site directly by clicking "NCBI GEO description" link. Users can also upload their own compendium by choosing "Upload user's own compendium" under "Select Compendium". The compendium should consist of one gene expression file and one annotation file. Please read the instructions displayed on the main panel carefully before preparing your own compendium.

The main panel will show a summary of how many genes are included in the current compendium. Users should consider changing gene sets or compendium if none of the gene sets has any gene in the compendium. Users can also break down current gene sets into smaller gene sets in "Gene Set Breakdown" tab. For a gene set having more than one gene, users can cluster all genes into user-defined number of sub gene sets (Using slider "Choose number of clusters") using hierarchical clustering. Note that the clustering depends on the compendium that users have selected. Click "Add sub gene sets" button to add all the subgroups into GSCA.

Users can choose to scale expression values across all samples so they have zero mean and unit variance, zero mean or unit variance by checking "Scaling and centering expression values across samples". Users can also select the method of

computing the gene set activities for gene sets with more than one genes. Either weighted average or median of all gene expression values can be selected.

4 GSCA Analysis

Select "GSCA Analysis" in the main menu. The GSCA analysis results will appear on the main panel. Users can choose and tune numeric or interactive gene expression pattern of interest (POI) as well as other options on the sidebar panel.

The settings of numeric POI are similar to one, two or more than two gene sets. There are two ways to specify numeric POI: using slider bar and typing in exact numbers using keyboard. The default method is using the slider bar. For each gene set GSCA will provide a slider bar and users can define POI by changing the slider bar values. Choose "Exact Number" under "Numeric POI" to switch to typing mode. Users can define the lower and upper bound for each gene set by typing in exact numbers. A quantile of the current numeric POI will show up below the input area. Users can also specify POI cutoff in a more rigorous way by checking "More POI cutoff options". Users can specify the upper or lower bound of a gene set to be a specific value of standard deviation from mean, quantile or quantile after fitting a normal distribution to the expressions. Click "Apply New Cutoff" button to apply the new cutoff and either the slider or the text input area will be updated accordingly.

For one gene set GSCA will show a set of histograms as output. The top histogram shows the expression values in all samples and the dashed lines show the current numeric POI. The following histograms show the expression values of top significant biological contexts. For two gene sets GSCA will show a scatterplot as output. The x and y axis corresponds to the activities of the two gene sets and each dot on the plot represent a gene expression sample. The numeric POI is represented as a blue rectangle with dashed lines. The top significant biological contexts will be highlighted using different colors. GSCA will also show the correlation between the activities of the two gene sets as well as other test information below the scatterplot. Users can check whether the correlation is nontrivial. For more than two gene sets GSCA will show two heatmaps as output. The heatmap on the top shows the gene set activities across all samples in the compendium. Each row represent a gene set and each column represent a gene expression sample. The colors stand for gene expression activities. With the blue-red palette the red color represents high gene set activity and the blue color represents low gene set activity. A color bar above the heat map shows the selected samples. Dark blue color shows the samples are selected and cyan color shows the samples are not selected.

Switch to "Ranking Table" tab on the main panel for more details of GSCA results. The results will list all enriched biological contexts with the full names, p-values, fold changes and other information. To change the cutoff of enrichment tests, users can change the p-value cutoff and fold change cutoff under "Specify biological contexts" on the sidebar panel. Users can also select the biological contexts to be displayed on the plots. Users can either select the number of top enriched biological contexts (Display top ranked contexts) or choose specific biological contexts (Display specified contexts).

Choose "Interactive POI" on the top of the sidebar panel to switch to interactive

POI. The settings of interactive POI are different for one, two and more than two gene sets. For one gene set and more than two gene sets, Users can use the slider bar above the line plot (one gene set) or heatmap (more than two gene sets) shown on the main panel to select intervals of samples of interest. Note that sometimes the slider is not aligned to the plot which makes the selection hard. Users should resize the browser window (Windows: Hold "Control" and press "-"; Mac: Hold "Command" and press "-") to align the slider to the plot. Users can also define POI using multiple intervals. Click "Add Slider" button to add a slider and "Delete Slider" to delete the slider. The POI will be defined as the union of all the intervals. After the POI is defined, click "Update Sample Selection" button on the sidebar panel and GSCA will recalculate the results. For more than two gene sets, sometimes users will find it hard to see the details of some small intervals. In this case users can check "Heatmap zoom in" function and the zoom-in view of samples within each interval will be shown below the heatmap.

For two gene sets, users will be able to specify POI by drawing polygons on a scatterplot. Simply click on the scatterplot to define the nodes of the polygons. Click "Finish Drawing Polygon" when you are finished and run GSCA analysis. Note that the GSCA result will appear below the scatterplot where you draw the polygons. Click "Add New Polygon" button to add new polygons, "Undo last Operation" to undo last operation or "Reset" button to reset all polygons. An important function of GSCA is saving and loading the POI so users will be able to reproduce their results with a new GSCA analysis. Users can save the current POI by clicking "Save Current POI" button under "Save Current POI". In a new GSCA session users can load the saved POI file under "Load POI" to recover the old POI. Notice that the gene sets, compendium, scaling options and numeric/interactive POI should be exactly the same as the settings in the previous GSCA analysis or there could be unpredictable errors.

5 Save Results

After GSCA analysis is performed, users can save GSCA output plots and ranking tables. Select "Save Results" in the main menu. First choose numeric POI or interactive POI and on the main panel the GSCA output plots will show up. Click "Ranking Table" tab to check the ranking tables. Save ranking table under "Download ranking table" and save output plots under "Download plots" on the sidebar panel. GSCA provides a variety of options for users to customize GSCA output plots, for example changing main title, range of x and y axis and color palettes. Users can feel free to explore these options until they get satisfactory output plots.

6 Utilities

GSCA provides a small tool for users to convert ENTREZ gene ID between human and mouse to support cross-species analysis. Select "Utilities" in the main menu. First users should upload the file under "Input File". Then users choose "Convert and Download" to convert the ENTREZ gene ID or gene name. Users should specify the column to be converted, the original species and gene

ID/gene name as well as the target species and gene ID/gene name. Click "Convert" to perform the conversion and "Reset" to reset all conversions. Click "Download" to download the converted gene set file. This file can then be directly uploaded to GSCA and users can perform the GSCA analysis in a new species.

7 Contact

To report bugs and provide suggestions for the GSCA GUI as well as the GSCA package, please contact the maintainer Zhicheng Ji (zji4@jhu.edu).