GSCA User Interface Manual

Zhicheng Ji, Hongkai Ji

March 23, 2014

1 Overview

Gene Set Context Analysis (GSCA) is a new approach and user-friendly software tool to help biomedical investigators to effectively utilize Publicly available gene Expression Data (PED) in their daily research. GSCA is built upon a collection of 30,000+ expert-annotated and consistently normalized gene expression samples in human and mouse representing 1000+ biological contexts. The primary goal of GSCA is to allow users to conveniently explore gene and gene set activities across these samples and link interesting expression patterns to biological contexts. Given multiple genes or gene sets, GSCA displays their activities across all samples in our data compendium. Users can interactively specify which activity pattern is of interest, and GSCA will then identify biological contexts in the compendium that are significantly associated with the pattern. Conceptually, one can use GSCA to answer questions such as "which diseases are associated with high activity of pathway A, low activity of pathway B, and medium activity of pathway C".
In addition to this user manual, users can also refer to the Youtube short video demo which is available at:
`https://www.youtube.com/watch?v=1OeZ1PAUMhw`

2 Launching

The easiest way of launching GSCA user interface is to go to URL:
`http://http://spark.rstudio.com/jzc19900805/GSCA/`
Notice that this approach does not require R or any R packages to be installed on users' computers. However, running GSCA UI on the web server could be slow when multiple users are visiting the website simultaneously.
The GSCA R package and four data packages are avaiblable on Bioconductor. If you have not installed Bioconductor previously, run the following commands in R console:

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

To install the GSCA R package, run the following commands in R console:

```
source("http://bioconductor.org/biocLite.R")
biocLite("GSCA")
```

At least one of the four data packages should be installed to run GSCA. The data packages are called Affyhgu133aExpr, Affymoe4302Expr, Affyhgu133A2Expr and Affyhgu133Plus2Expr. To install these data packages, run the following commands in R console:

```
source("http://bioconductor.org/biocLite.R")
biocLite("Affyhgu133aExpr")
biocLite("Affymoe4302Expr")
biocLite("Affyhgu133A2Expr")
biocLite("Affyhgu133Plus2Expr")
```

After at least one of the four data packages has been installed, run the following command in R console to launch GSCA UI in locally:

```
GSCAui()
```

3 Analysis Workflow

3.1 UI layout

The UI consists of two main parts: sidebar panel on the left side of the UI 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 UI, users should go through four major analysis steps: input geneset, select geneset and compendium, run GSCA analysis and save results. Users can switch between these steps via the main menu on the top-left side of the UI.
There is a small information panel on the top showing status of GSCA. If the panel shows GSCA is computing, please wait GSCA to finish computing before doing further analysis or change other options.

3.2 Input geneset

The first step is to input genesets into GSCA. Users should first input geneset name for each individual geneset in "Input Geneset Name" text input area. Note that the name of genedata should be easy to memorize and does not need to follow any other specific rule. After that, users can choose to directly type in the Entrez GeneID or upload prepared geneset file. Choose "Specify Gene ID" option to directly type in the Entrez GeneID and specify whether genes are activated or repressed. Users should seperate different genes with ;. For example, if geneID 10 and 100 are activated genes and geneID 1000 is repressed gene in the geneset, 10;100 should be given in "Specify Activated Entrez GeneID" text input field and 1000 should be given in "Specify Repressed Entrez GeneID" field. For this input method the weights for all activated genes are 1 and weights for all repressed genes are -1. Choose "Upload Geneset File"

option to upload prepared geneset 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 geneset names. Click "Choose File" button to select the designated file path. Then users can decide whether to include header, change the separator and change the quote. If nothing appears in the main panel on the right side, there should be something wrong when reading in the file and users should consider changing the options. Click "Add geneset" button to add the geneset being inputted into GSCA.
In the middle is the main panel showing 1. the current geneset being input 2. a summary of all genesets already entered in GSCA and 3. details of individual genesets already entered in GSCA. Users should focus on tab 1 when inputting genesets. Tab 2 and 3 are helpful to check the inputs before going to the following steps.
If there are input mistakes with one or multiple genesets, users can go to "Delete Existing Geneset" panel and delete the chosen genesets. Then users can reinput these genesets. Users can also click "Reset all genesets" button to delete all existing genesets when starting a brand new GSCA analysis.
Finally, to store all inputted genesets for further analysis, users can click "Save" button under "Save current genesets as csv file". Users can simply upload this csv file to reload all genesets in a new GSCA session.

3.3 Select geneset and compendium

Select "Select geneset and compendium" in the main menu. In the sidebar panel users should first select the genesets they want to include in GSCA analysis. Then users should select with which compendium they want to perform GSCA analysis. Users can either select preprocessed compendiums or upload their own compendiums. By default users should choose one of the preprocessed compendiums. Currently there are four preprocessed compendiums available: Affymetrix Human hgu133a Array (GPL96); Affymetrix Mouse 430 2.0 Array (GPL1261); Affymetrix Human Genome U133 Plus 2.0 Array (GPL570); Affymetrix Human Genome U133A 2.0 Array (GPL571). When any of the compendium is selected, information of how many samples and genes are included in this compendium will show up below. Users can also go to the NCBI GEO description site directly by clicking "NCBI GEO description" link. Users can also upload their own compendiums which consists of preprocessed gene expression data and annotation file. Please read the instructions displayed in the UI carefully when preparing your own compendiums.
The main panel will show a summary of how many genes are included in the current compendium. Users should consider changing genesets or compendium if none of the genesets has any gene in the compendium. Users can also break down current genesets into smaller genesets in "Geneset breakdown" tab. For genesets having more than one gene, users can cluster all genes in any of such genesets into user-defined number of sub genesets (Using slider "Choose number of groups") using hierarchichal clustering. Note that the clustering depends on the compendium users have selected. Click "Add sub genesets" to add all the subgroups into GSCA.
There are also other options provided in the bottom left area. Users can choose to scale expression values across all samples so they have zero mean and unit

variance. Users can also select how to compute the geneset activity value for genesets having more than one genes. Either weighted average or median of all gene expression values can be selected.

3.4 GSCA analysis

Select "GSCA analysis" in the main menu. The GSCA analysis results will appear in the main panel and users can change gene expression pattern of interest (POI) in the sidebar panel. For one geneset GSCA will give a set of histograms as output. For two genesets GSCA will give a scatterplot as output. For more than two genesets GSCA will give two heatmaps as output. Switch to "Ranking Table" tab in the main panel for more details of GSCA results.
GSCA UI provides two ways for users to specify POI, numeric POI and interactive POI. By default, GSCA will first run analysis on numeric POI. Users can specify numeric POI by changing slider values in "Numeric POI" sidebar panel. More specific cutoffs are also supported when "More options" radiobutton is checked. These cutoffs includes standard deviation from the mean, quantile of a normal fit to the expression values or the quantile. After selecting geneset, upper/lower bound, cutoff pattern and giving in cutoff value, click "Apply new cutoff" to apply the changes.
To switch to interactive POI, change the radiobutton from "Numeric POI" to "Interactive POI". For one geneset case and more than two genesets case, users can specify POI using a set of sliders. Simply change the slider value to include an interval of samples to be selected. Click "Add Slider" and "Delete Slider" buttons to add/delete number of sliders. Note that the final POI will be the union of selected samples from all sliders. Click "Update Sample Selection" button in the sidebar panel to update the POI and run GSCA analysis. For two genesets case, users can 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.
GSCA provides additional options for users to further explore GSCA results or customize outputs. These options are available under "Options" wellpanel in the sidebar panel. Users can change enrichment p-value and foldchange cutoffs to include more/less significantly enriched biological contexts. Generally speaking, smaller p-value cutoffs and larger foldchange cutoffs leads to more strict criterion of selecting biological contexts. Users can also choose what biological contexts to be displayed on the output plots. Select "Display top ranked contexts" and change the slider value to display arbitrary number of top biological contexts ranked by enrichment p-values. Select "Display specified cutoffs" to choose specific biological contexts to be displayed. Other options are also provided under "Options" wellpanel. For two genesets case, users can choose whether designated enriched biological contexts are plotted within POI or on the whole scatterplot by checking "Show enriched context only in POI". For more than two genesets case, users can choose to suppress the color range of heatmap by changing the slider values in "Suppress heatmap color range". This function can come in handy when there exist outliers in expression values. Users can also choose to cluster on genesets (rows) by checking "Cluster on rows".

Users can save current POI in the "Save/Load POI" wellpanel in the sidebar panel. In a new GSCA session simply load the saved POI file to recover the old POI. Notice that the genesets should be exactly the same as those in the previous GSCA analysis or there could be unpredictable errors.

3.5 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 which POI to use, numeric POI or interactive POI. In the main panel the exact GSCA output plots will show up. Click "Ranking Table" tab to check the ranking tables. Save ranking table under "Download ranking table" wellpanel in the sidebar panel. Save output plots under "Download plots" wellpanel in 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.