GSCAR User Interface User Manual

Zhicheng Ji, Hongkai Ji
Department of Biostatistics, Johns Hopkins University

January 16, 2014

Introduction

Although techniques of chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-seq) or tiling array hybridization (ChIP-chip) have came into being for a long time, it still remains difficult to generate a quality ChIPx (i.e., ChIP-seq or ChIP-chip) data set due to the tremendous amount of effort required to develop effective antibodies and efficient protocols. Especially with recent cuts in research fundings, most labs are unable to easily obtain ChIPx data in more than a handful of biological contexts. Thus, standard ChIPx analyses primarily focus on analyzing data from one experiment, and the discoveries are restricted to a specific biological context. We propose to enrich this existing data analysis paradigm by developing a novel approach, GSCAR, which superimposes ChIPx data on large amounts of Publicly available human and mouse gene Expression Data containing a diverse collection of cell types, tissues, and disease conditions to discover new biological contexts with potential geneset activity patterns. GSCAR could also serve as an informative guide for biologists to prescreen interested biological contexts when designing their experiments.

Overview

The purpose of GSCAR is to predict the biological contexts, defined as the cell or disease type and associated treatment or condition, in which a certain geneset activity pattern exhibits. GSCAR accomplishes this by first requiring the users to specify a number of genesets with activated (positive) and repressed (negative) genes defined from experimental data from one or more cell types. Users are also required to specify a particular geneset activity pattern they want to study. Given the genesets, GSCAR will then search for biological contexts that are significantly enriched with the specific geneset activity pattern by examining the activity value of each given geneset across all of the biological contexts in the gene expression compendium. Finally, a complete report including result tables and plots will be generated.
As an enhancement to the standard GSCAR R package, GSCAR user interface aims to provide users who do not have prior knowledge in computer program-

ming or statistics a user-friendly tool to conduct standard GSCAR functions as well as more powerful and insightful analyses. Thanks to the R shiny server, users do not even need to have R or any necessary R package installed on their computers to run GSCAR user interface. The easiest way is just to open the web browser, type in the URL and the user interface will be launched.

How to use

1. Initiating

A standard way to initiate GSCAR user interface is to first download and install GSCAR and GSCARdata packages from githut. Simply run the following code in R console, and you may need to install package devtools first. Notice that GSCARdata package is about 840 MB large so downloading could take some time. Also make sure that your computer should have at least 1GB free memory when doing the installation of GSCARdata.

```
require(devtools)
install_github("GSCAR", "zji90")
install_github("GSCARdata", "zji90")
```

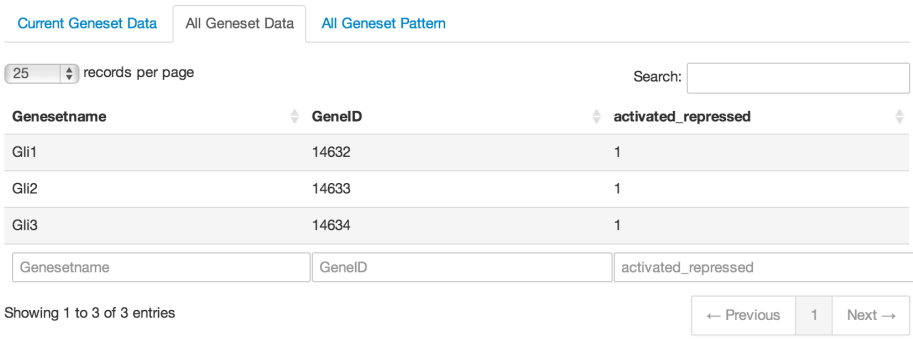
After R finishes downloading and installing the two packages, you can run GSCARui function to launch the GSCAR user interface in your web browser.

```
GSCARui()
```

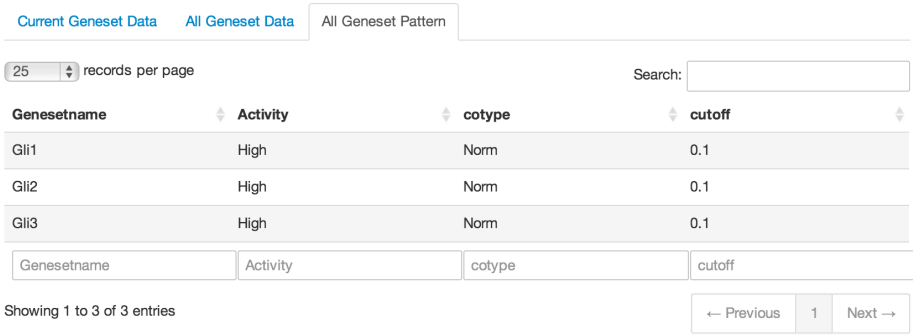
Currently there could be some parsing, warning and error messages bumping in R console. Just ignore these messages.
Another easier way to launch GSCAR user interface is to directly go to URL: `http://spark.rstudio.com/jzc19900805/GSCAR/`
GSCAR user interface will run on a server provided by rstudio. This does not require users to have R or any dependent R packages installed on their computers. However, running the UI on the server could be slower than running the UI on users' own computers. If users wish to explore GSCAR functions with depth, the standard way of installation is recommended.
Users are recommended to check the help pages of GSCAR package (especially GSCAR function) before they run any analysis on GSCAR user interface. Details about inputs and methods used by GSCAR will not be included in this user manual.
After GSCAR user interface has been displayed on the web browser, a set of radiobuttons (Main menu) will appear on the top left corner of the browser. These buttons list the main steps of using GSCAR and users can change among different steps freely.

2. Input Geneset Data

The primary example used in the following parts of the user manual is GSCAR analysis on mouse gene Gli1,Gli2 and Gli3, all members of GLI-Kruppel family. Their corresponding Entrez GeneID are 14632,14633 and 14634.
To input geneset data, select "Input Geneset Data" in the main panel (will automatically be in this mode when startup).
The first step is to input genedata and pattern into the program. On the left side below the main menu panel there are two panels where users can specify their genedata and pattern inputs one by one. On the middle is the main panel showing the current genedata being input and all genedata which have already entered in the program. Note that every action taken in the sidepanel could change the display in "Current Geneset Data" so users can know what their input is.
To input genedata, first type in the name of the genedata 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 existing genedata file. Choose "Specify Gene ID and +/-" option to directly type in the Entrez GeneID and specify whether genes are activated or repressed. Users should seperate different genes with ';'. For example, 10;100;1000 should be given in "Specify Gene ID" text input field and 1;1;-1 should be given in "Specify activation or repression" field if geneID 10 and 100 are activated genes and geneID 1000 is repressed gene in the genedata. Choose "Upload Geneset File" option to upload existing genedata file. 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 depending on whether the program has correctly read in the data.
To input corresponding pattern, choose geneset activity pattern, cutoff type from the pull down menu and specify the cutoff value in the text input area. Please check the help page of GSCAR function for details.
After the genedata and pattern are inputted, click "Add Geneset Information" button on the bottom of the page to add the current genedata and corresponding pattern to the program. Users can check all genedata and pattern which have been inputted in "All Geneset Data" and "All Geneset Pattern" tabs.
In our example, the final input of genedata and pattern should look like this:
genedata:

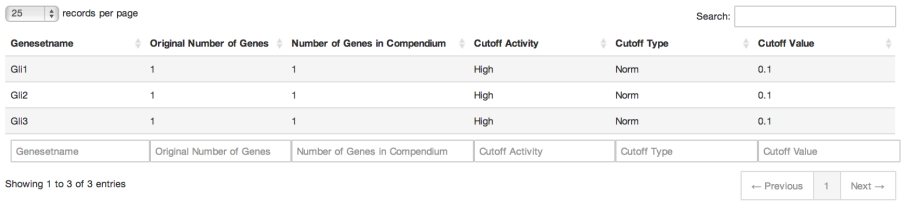


pattern:



3. Select Geneset and Compendium

Select "Select Geneset and Compendium" in the main menu. On the left side under the main menu users can select the genedata they want to include in the analysis and the compendium. Currectly two compendiums, human and mouse, are available, see help pages of GSCARdata package for details of the two compendiums. A summary about the genedata being selected will apear in the middle of the page. Only genedata which contain at least one gene included in the compendium will appear in the summary table. If there is no genedata which contains at least one gene included in the compendium, then no summary page will appear. Note that users can always go back to this page to reselect genedata and compendium during the analysis.
In our example, the summary table will look like this if all three genedata are selected:

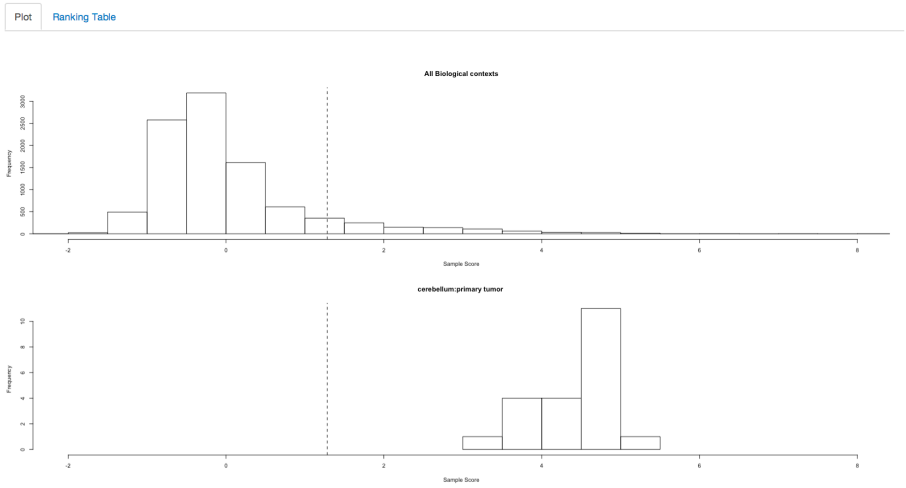


4. GSCAR

Select "GSCAR" in the main menu. On the left side below the main menu is the sidebar panel with which users can specify geneset activity pattern and alter plotting options. There are basically two ways to specify geneset activity pattern in GSCAR: default way and interactive way. Common options users can specify in both ways are p-value and foldchange cutoffs and biological contexts users want to display in the plots. Users can select the enrichment p-value and foldchange cutoffs in "Enrichment P-value cutoff" and "Enrichment Foldchange cutoff" text input area and only biological contexts which meets the both cutoffs will be selected. Note that smaller p-value and bigger foldchange cutoffs impose more stringent requirement of context selection. Users can also specify biological contexts displayed in the plot in "Choose Biological Context displaying Method" area. Select "Display top ranked Contexts" radio button option and change the value in the slider to display different numbers of top ranked biological contexts. Select "Specify Contexts" radio button option to specify any biological context to be displayed. Note that only biological contexts that have at least one samples in the geneset activity region can be selected. To select multiple contexts, hold Control key in Windows or Command key in Mac and click on different contexts. GSCAR plot and ranking table will be displayed in the middle of the page, and users can select between displaying plot and displaying ranking table using tabs on the top. GSCAR will automatically generate different sidebar panels and plots according to the number of genedata users want to include in their analyses. We will introduce GSCAR usage that have not been included in previous introductions seperately.

Case of one genedata

Select "Gli1" only in the previous step. GSCAR will display a set of histograms of geneset activity values of all samples and of samples in top ranked enriched biological contexts.
In the default enrichment region selection method, the enrichment region is defined based on users' previous input of pattern. Check help pages of GSCAR package for detail. The GSCAR plot will look like this:



The ranking table will look like this: (click on "Foldchange" to reorder the table according to foldchange)

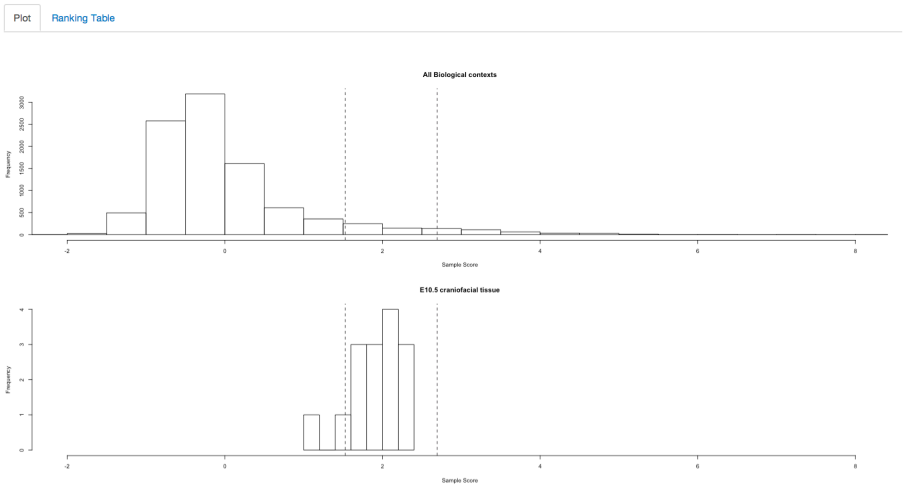
Plot | Ranking Table

25 records per page

Search:

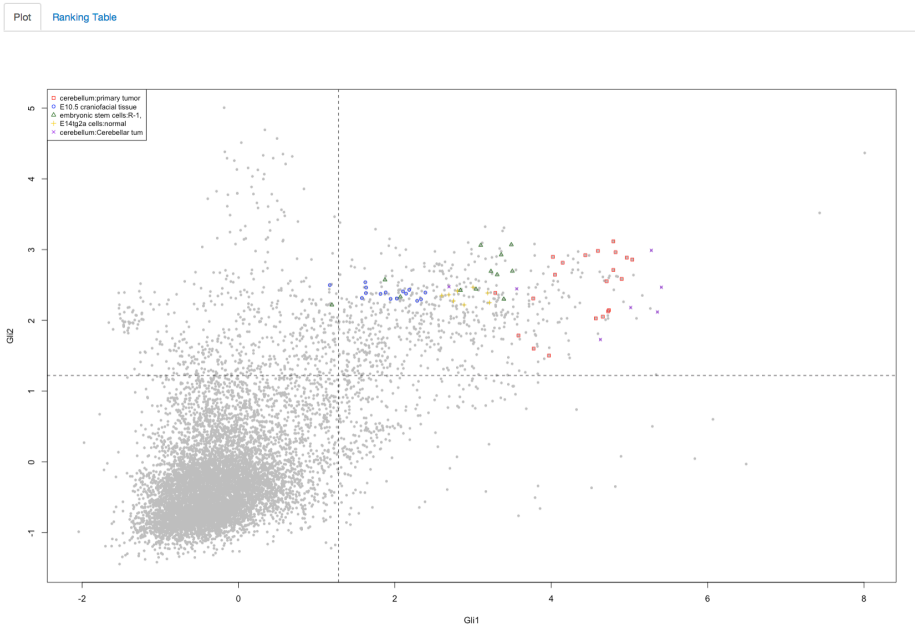
Rank	Active	Total	FoldChange	Adj.Pvalue	SampleType	ExperimentID
27	8	13	5.980	9.32e-03	embryonic stem cells:normal	GSE10476;GSE10573;GSE10553;GSE10610;GSE10806;GSE9954
28	8	13	5.980	9.32e-03	retinal progenitor cell P0:normal	GSE9811
21	8	10	7.611	4.25e-04	neonatal small intestine:normal	GSE6065
4	10	11	8.700	1.04e-06	retinal progenitor cell E16.5:normal	GSE9811

Choose "Interactive Enrichment Region Selection" radio button option to select enrichment region interactively. A slider called "Choose your interested region" will then appear. Users can change the two values of the slider and the enrichment region will be the region between the two values. The GSCAR plot will look like this:



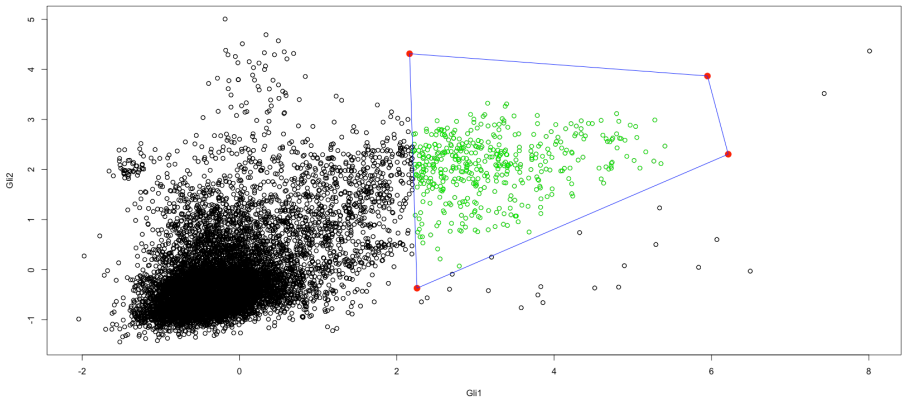
Case of two genedata

Select both "Gli1" and "Gli2" in the previous step. In the case of two genedata, users can specify whether to display the biological contexts outside the geneset activity region via "Show Enriched Context only in interested region" checkbox. GSCAR plot of default enrichment region selection looks like this:

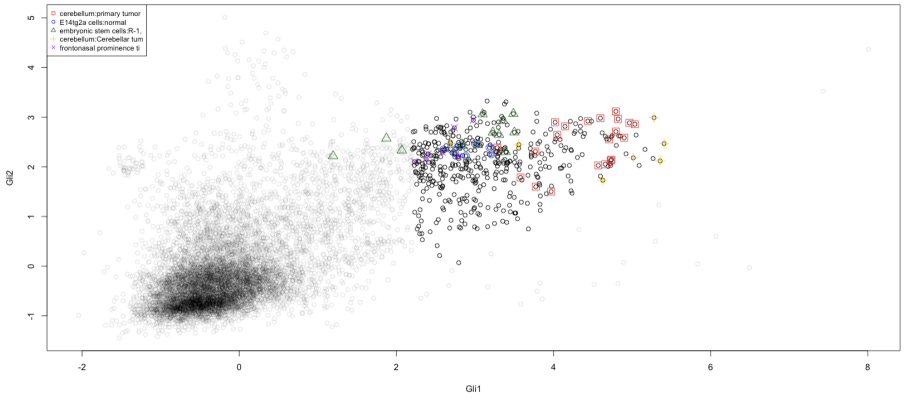


If interactive enrichment region selection is chosen, a scatter plot will appear in the middle of the page. In the first plot, users are supposed to draw a polygon to define the geneset activity region and GSCAR will calculate all

enriched biological contexts within this region. The drawing of the polygon can be finished using the three buttons in "Choose your interested region" area. To start, click on the plot where the first node of the polygon should be, and there appears a red node in the position. Continue clicking on the plot to add new nodes of the polygon and all previous nodes will be connected with blue lines. After the last click, hit "Close Polygon" button to close the polygon and the program will automatically calculate and show all samples staying within the polygon with green color. If users misclicked some nodes, simply hit "Remove last point" button to undo the last click. Hit "Reset" button to clear out the current polygon and start again from scratch. The polygon drawing process will look like this:

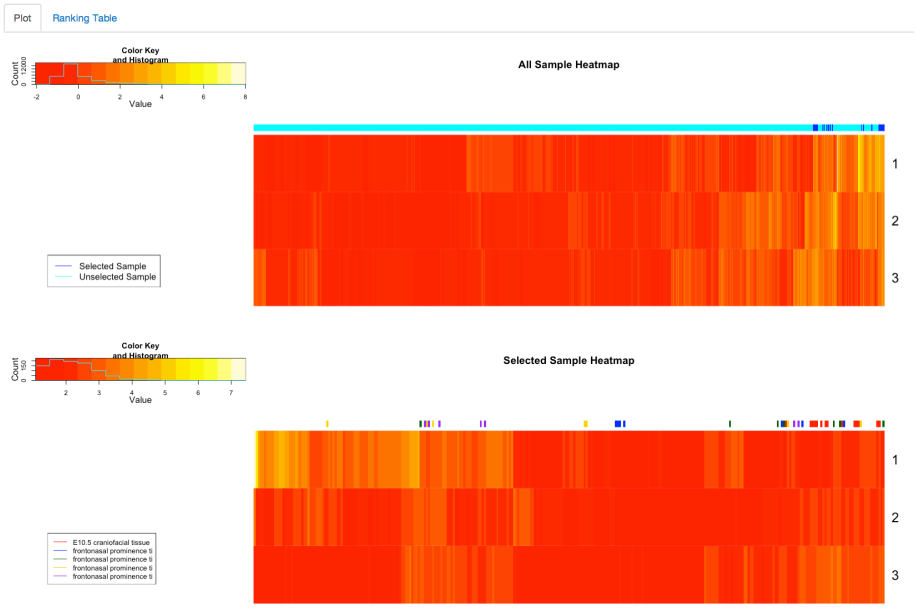


Each time the button "Close Polygon" is hit, GSCAR will automatically display GSCAR plot underneath the polygon drawing plot. Note that samples staying within the polygon are shown black and samples staying outside the polygon are shown gray. GSCAR plot will look like this:

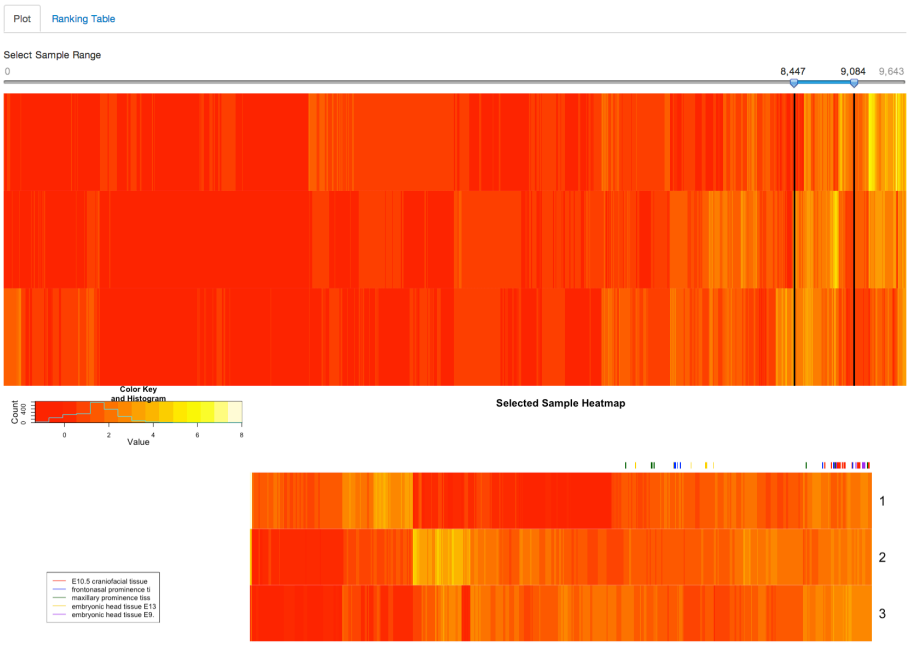


Case of more than two genedata

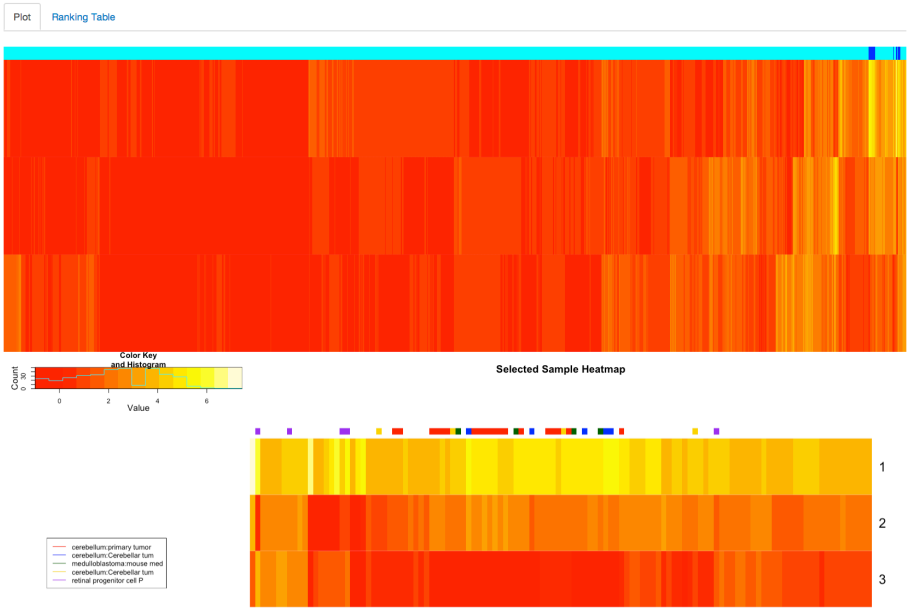
In the case where more than two genedata are included in the analysis, two heatmaps will always be displayed. The first heatmap shows the geneset activity values of all samples and samples staying within or not within the genedata activity region will be indicated by a column colorbar. The second heatmap shows the geneset activity values of samples staying within the genedata activity region and all top ranked or specified biologcial contexts are indicated in the column colorbar.
Note that because there are tens of thousand samples in either of the two compendiums, drawing the first heatmap could take some time. Please be patient waiting for the heatmaps to appear.
GSCAR plot of default enrichment region selection looks like this, note that legend of the second heatmap does not reveal the full name of the biological context. In this case users can always check the ranking table:



There are two ways to specify geneset activity region interactively, by sample or by value. To specify by sample, choose "Sample" radio button selection. The first heatmap will appear in the middle and a slider also appears above the heatmap. Change the left and right values of slider and the geneset activity region will be all samples between the two sliders. Click "Update Sample Selection" button to let GSCAR calculate and display the second heatmap. The selection process as well as the GSCAR plot looks like this:



To specify by value, choose "Value" radio button selection. The first heatmap will appear in the middle and several sliders corresponding to each genedata will appear in the left sidebar panel. Change the values of the sliders to alter the geneset acitivity region. After finishing changing the sliders, click "Update Sample Selection" button to let GSCAR calculate and display the second heatmap. The selection process as well as the GSCAR plot looks like this:



5. Download

GSCAR provides multiple options for users to customize their GSCAR plots so that they can easily fit the plots in scientific reports or papers. Select "GSCARdefault" or "GSCARinteractive" in "Select Genedata Acitivity Region" radio button set to choose using default geneset activity region or using interactive geneset activity region that has already be defined in "GSCAR" step. If users want to change geneset activity region, they should go back to previous steps to make the changes.

Note that previews of GSCAR plot and ranking table is available in the middle of the page.

GSCAR enables users to download ranking table or GSCAR plot. For the ranking table, after selecting file type (txt or csv), specifing file name and choosing whether to include column names, users can click "Save Ranking Table" button and the ranking table will be saved in users' default download path. For GSCAR plot, after selecting file type (pdf,png,ps,jpeg,bmp,tiff) and specifing file name users can click "Save Plot" or "Save Heatmap" button and the corresponding plot will be saved in users' default download path. To further customize plots, check "Change Plotting Details" area. In the case where less than or equal to two genedata are included, users can make changes of main title, x or y axis titles and x or y axis ranges. In the case where more than two genedata are included, users can specify the palette used to draw the heatmap as well as whether to include column dendrograms in the heatmaps. Note that including dedrogram in the first heatmap could take some time.

More downloading options are coming in the future.