Systems-Level Interactive Data Exploration (SLIDE v1.0)

USER'S MANUAL

SOUMITA GHOSH

Table of Contents

Systems-Level Interactive Data Exploration (SLIDE) is a user-driven interactive visualization tool for large-scale –omics data. This document describes the steps for using SLIDE's functionalities. SLIDE is distributed under the BSD license. See the LICENSE.txt in the source distribution or http://opensource.org/licenses/BSD-2-Clause.

I. SLIDE Home

SLIDE Homepage

SLIDE's homepage contains three buttons. Their functionalities are as follows:

1. Create New Analysis – Enter a name for the workspace
2. Load Analysis From File – Upload a saved workspace (.slide) for continued analysis
3. Open Active Analysis – The currently running analyses (workspaces including sub-analysis and enrichment analysis) are logged here.

II. Inputs to SLIDE

1. Input Data

SLIDE takes an input data file containing the matrix of expression values in text delimited format. The input data file must have at least one column of feature meta data such as entrez and gene symbols. SLIDE can automatically map missing meta data information if either entrez or gene symbols are available.

Input Data File

ID_REF	Gene	Entrez	SH_d1_r1	SH_d1_r2	Tx91_d1_r1	Tx91_d1_r2	SH_d2_r1	SH_d2_r2	Tx91_d2_r1	Tx91_d2_r2
ILMN_1	ABC		-0.07709	0.03832	-0.02144	0.01178	0.1740	0.1307	-0.0519	0.2546
ILMN_2		11223	-0.04110	-0.02241	-0.17497	-0.01753	0.0819	0.0525	0.1215	-0.0612
ILMN_3	DEF	44556	0.00271	0.0389	0.1590	0.0166	-0.0313	0.1794	0.0400	0.0322

2. Sample Mapping File

A sample mapping file is a tab delimited text file where each row contains meta data for a column of the input data. The first entry in each row of the sample mapping file is the column header name in the input data file. The second and third columns contain the sample grouping and time point information. A sample mapping file corresponding to the input data file above is shown below. Here, a time-course experiment with two time-points, two experimental conditions (SHAM and Tx91) and two replicates per time-point and condition combination is shown.

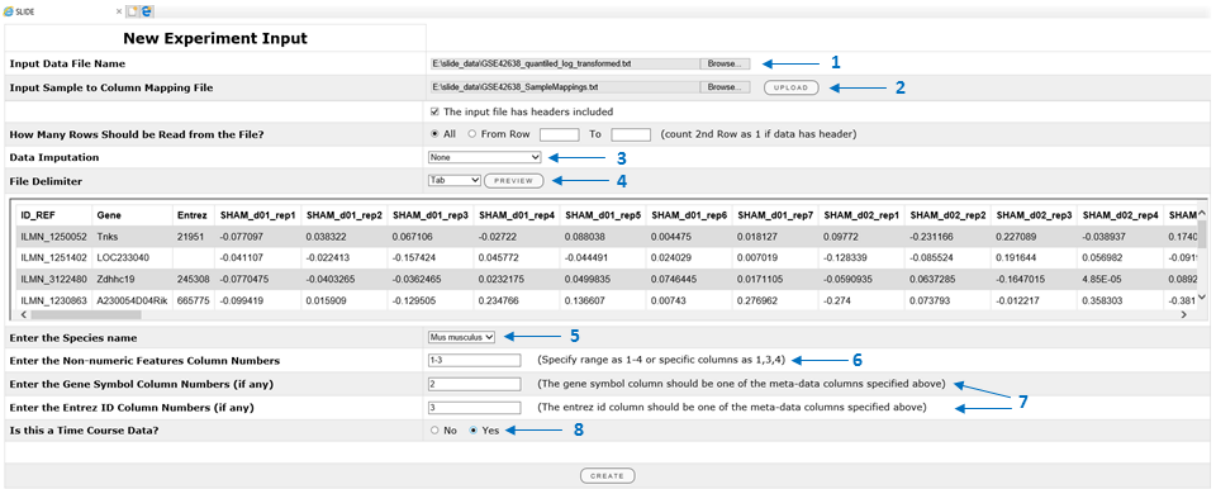
Sample Mapping File

#Column	SampleGroup	Timepoint
#Day 1		
SH_d1_r1	SHAM	Day_1
SH_d1_r2	SHAM	Day_1
Tx91_d1_r1	Tx91	Day_1
Tx91_d1_r2	Tx91	Day_1
#Day 2		
SH_d2_r1	SHAM	Day_2
SH_d2_r2	SHAM	Day_2
Tx91_d2_r1	Tx91	Day_2
Tx91_d2_r2	Tx91	Day_2

Lines beginning with # are comments and are not processed. The file can have line breaks for ease of formatting. For a replicate, non-time-series experiment, only the first two columns are used. For a replicate, time-series experiment, the first three columns are used. In case of independent experiment design (no replicates, each sample is independent), only the first column is needed.

An example of the sample information file for the mouse data set is provided along with the software package.

III. Create New Experiment



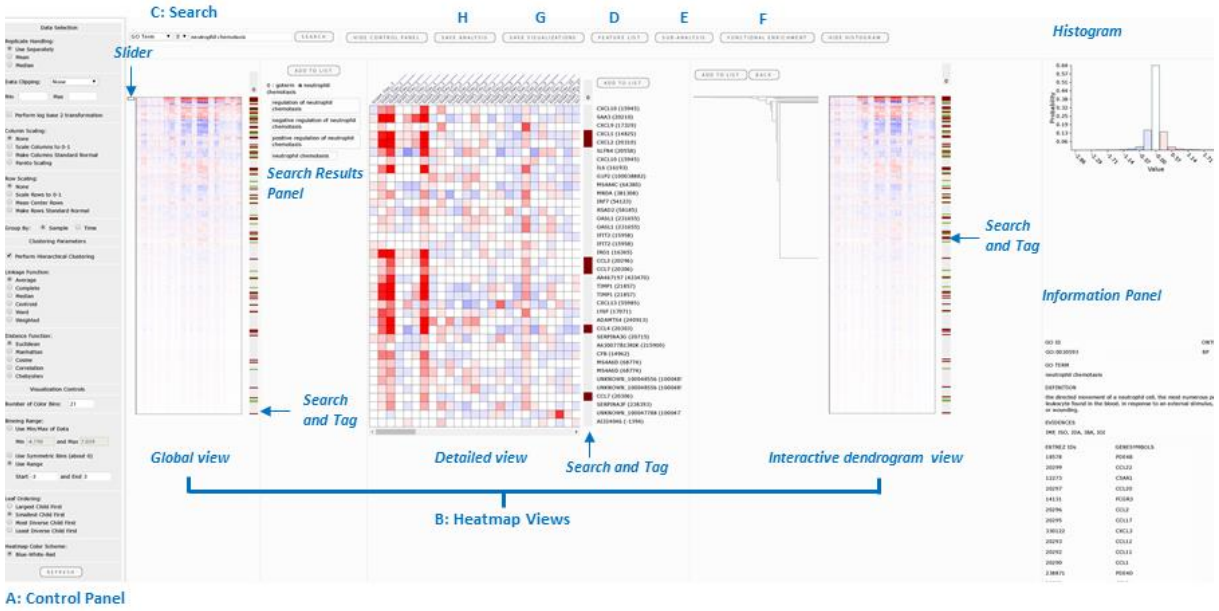
Create New Experiment Page

The New Experiment Form requires the following inputs:

1. The path to the input data file.
2. The path to the sample mapping file.
3. An appropriate data imputation method. If you do not want to apply any data imputation, please select ‘None’.
4. The input file’s delimiter. Click the ‘Preview’ button to verify if the file was read in correctly.

5. The appropriate species name.
6. The non-numeric columns that contain feature information in the input data file. In the data file preview shown in the figure above, columns 1, 2 and 3 contain feature metadata information corresponding to ID_REF, Gene Symbol, and Entrez ID, respectively. These non-numeric, meta data columns are identified in the textbox (marked by 6) as 1-3 or as comma-separated 1, 2, 3.
   Note: Column indexing in SLIDE starts from 1, not 0.
7. The columns of the input data file that contain Gene Symbols and Entrez ID.
   Note: While it is possible to use SLIDE without both gene symbols and entrez information, providing at least one of the two is necessary to fully utilize the meta data based search and tagging functionality.
8. The experiment type: time-series or non-time-series experiment.

IV. Visualization Home



SLIDE Visualization Page

Once the data is successfully loaded into SLIDE, the *default binning range* (see A.11 for details), the maximum and minimum of the data, is applied. The binning range is the data range that is mapped to the color bins. A first step would be to change the *binning range* depending on the range of the data. This step can be useful in reducing the effect of outliers and allows highlighting the meaningful differences in expression levels. Without an appropriate binning range and appropriate data transformation the changes in expression level may not be immediately perceivable (all the cells in the heatmap may appear as similar colors).

If your data has already been pre-processed (for instance, log transformed and mean centered), you will need to change only the visualization parameters and may ignore the transformation parameters.

If the relative patterns of expression level change between features is of interest, use the *mean center rows* (see A.5 for details) option.

Users can further explore the effects of various data transformation, visualization and clustering parameters using the *control panel* described below.

A. Control Panel

The *control panel* lists the parameters for data transformation, visualization and clustering.

Note: The parameters for data transformation are applied in the order they appear in the control panel. After applying each option, users have to click on the Refresh button in the bottom of the panel to apply the changes. Each time the user chooses a set of data transformation parameters and applies them using the Refresh button, the selected set of transformations are applied on the raw data.

Control panel parameters:

1. *Replicate Handling*: Select the appropriate option to visualize the quantitative expression of each replicate of a sample group. 'Use separately' displays all replicates, 'Mean' displays replicate average value and 'Median' displays replicate median for each feature. The *default* is 'Use separately'.

2. *Data Clipping*: Optionally remove outliers in the data that can potentially skew the visualization. The *default* is 'None'. If other, options are selected the 'Min' and 'Max' range has to be specified.

3. *Perform log base 2 transformation (optional)*: If checked, the data will be $\log_2$ transformed.

4. *Column Centering and Scaling*: The range of values in each column (sample) of the input data file may vary widely. A column with a wide range of values will influence the visualization and clustering results. Column scaling independently standardizes the values in each column to similar range. The *default* is 'None'. The 'Modified Pareto Scaling' transforms the data matrix **X** as follows:

$$\frac{\mathrm{x}_j - \mu_j}{\sigma_j} + \frac{\sum_{x_{ij} \in \mathbf{X}} x_{ij}}{N}$$

here, x_j is the j[th] column of the data, μ_j and σ_j are the mean and standard deviation for the j[th] column, x_{ij} is the ij[th] element of **X** and *N* is the total number of elements in **X**.

5. *Row Centering and Scaling*: The range of values in each row (feature) of the input data file may vary widely. A row with a wide range of values will influence the visualization and clustering results. Row scaling standardizes the range of each row independently to a common range. The 'Mean Center Rows' option independently shifts each row so that their means are at 0. The option is useful to discern patterns across features. The *default* is 'None'.

6. *Group By*: This option allows grouping the samples based on the sample group names or based on the time-point.

7. *Perform Hierarchical Clustering (optional)*: If 'checked', SLIDE performs agglomerative hierarchical clustering. To perform hierarchical clustering select the linkage and distance functions as mentioned in 8 and 9. The *default* is no hierarchical clustering.

Hint: SLIDE caches the result of hierarchical clustering. For each combination of data transformation and hierarchical clustering parameters, the clustering is performed only once and the results are cached. For a combination of parameters, for which clustering has been performed before, the clustering can be re-applied in real time.

8. *Linkage Function:* The linkage functions are used to compute the distance between two clusters. The *default* linkage function is 'Average'.

9. *Distance Function:* The distance functions are used to compute the similarities between data points. The *default* distance function is 'Euclidean'

10. *Number of Color Bins*: Specify the number of bins used to discretize the color range. The *default* is 21 for feature-level visualization and 51 for group-level visualization.

11. *Binning Range:* The data range to be mapped to colors. The *default* range is 'Use Min/Max of Data' where the minimum and maximum of the data are mapped to the purest blue and purest red respectively. Custom ranges can be specified using 'Use Range' which can be used to reduce the effects of outliers. Specifying a custom range of say -2 to 2 for a data that contains -100 will give all cells that have -2 or less value the same color (purest blue).
Hint: The histogram (right side of the screen) can be used to check the data range and adjust the binning range accordingly.

12. *Leaf Ordering*: A leaf ordering scheme is necessary to visualize the dendrogram tree. The output of hierarchical clustering, the dendrogram, is a series of binary splits. The leaf ordering determines which sub-tree will be visualized on top at each split of the binary tree. The appropriate leaf ordering ensures that similar clusters are grouped together. The *default* is 'Largest Child First'.

B. Heatmap Views

SLIDE offers three heatmap views, at multiple resolutions, to navigate through the data: *global view*, *detailed view* and *interactive dendrogram view*.

The *global view* displays a heatmap of the clustered data in its entirety.

The *detailed view* displays a selected portion of the data in a zoomed-in view. A slider attached to the heatmap in the *global view* as shown in the figure above allows the user to scroll through the entire data and select the portion of the data to be visualized in the *detailed view*.

The *interactive dendrogram view* is displayed only after hierarchical clustering is performed. The branches of the dendrogram can be clicked to visualize a subset of features. In each view, the dendrogram sub-tree starting at the selected root node and expanding down till twenty-five leaf nodes are found is displayed. When the tree reaches a certain depth, the feature labels are displayed alongside the heatmap.

C. Search

Individual genes and functional terms can be queried in SLIDE. The *search results panel* (see figure above) displays the outcome of user queries. Users can search genes, biological functions/pathways. The associated genes are tagged in vertical bars next to all heatmaps. The search can be a wildcard search and can include multiple comma separated search terms.

Clicking a keyword in the *search results panel* highlights the associated search tags along the three heatmaps. The *information panel* displays the details of features, pathways and gene ontologies selected (clicked) by the user in the heatmap views and the *search result panel*.

Search Bar

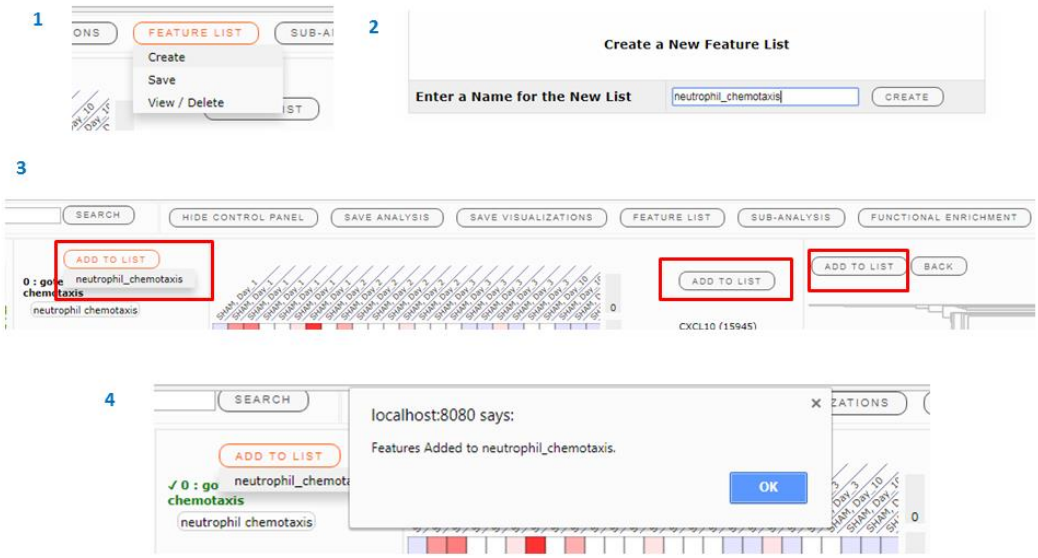
To use the 'search' functionality, do the following:

1. Select one of the available query types:
   Entrez ID, Gene symbol, GO ID, GO Term, Pathway ID or Pathway Name
2. **Select the type of search: exact (=) or wildcard (≅)**.
   Note: Exact search returns results matching the whole word(s)/term(s)/string in the database, whereas wildcard search returns all matches that contain the string in it.
3. Enter the term to search.
   Note: To search multiple terms, enter them as a comma-separated string.
4. Click 'Search' button to see the results of the search.

D. Feature List

Multiple lists of user-selected genes can be maintained in SLIDE, for further *sub-analysis* or enrichment analysis. The *feature lists* created within SLIDE can also be saved in text file format.

SLIDE provides multiple ways of adding features to these user-created lists. For instance, individual genes can be added to the *feature lists* from the *detailed view* panel. Likewise, clusters of genes can be added to the *feature lists* from the *interactive dendrogram view* (by selecting a branch), or functionally related genes can be added to the *feature lists* from the *search results panel*.



Feature List Creation

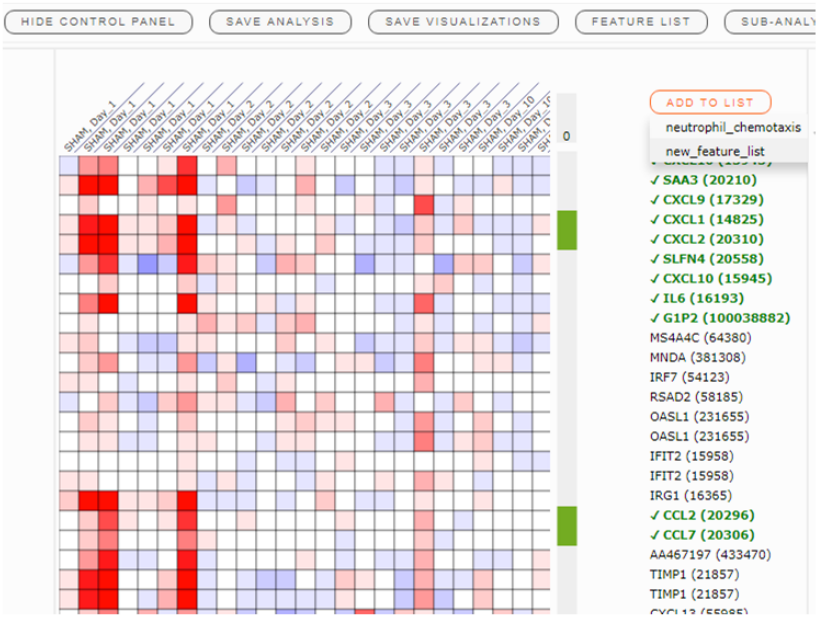
To add features to a list, do the following to first create a feature list:

1. Mouse over the ‘Feature List’ button and click ‘Create’.
2. Enter a new name for the ‘Feature List’. Click ‘Create’ and close the modal window. This creates an empty *feature list*.

In the figure above, an empty *feature list* named ‘neutrophil_chemotaxis’ is created. To add genes to this *feature list*, do the following:

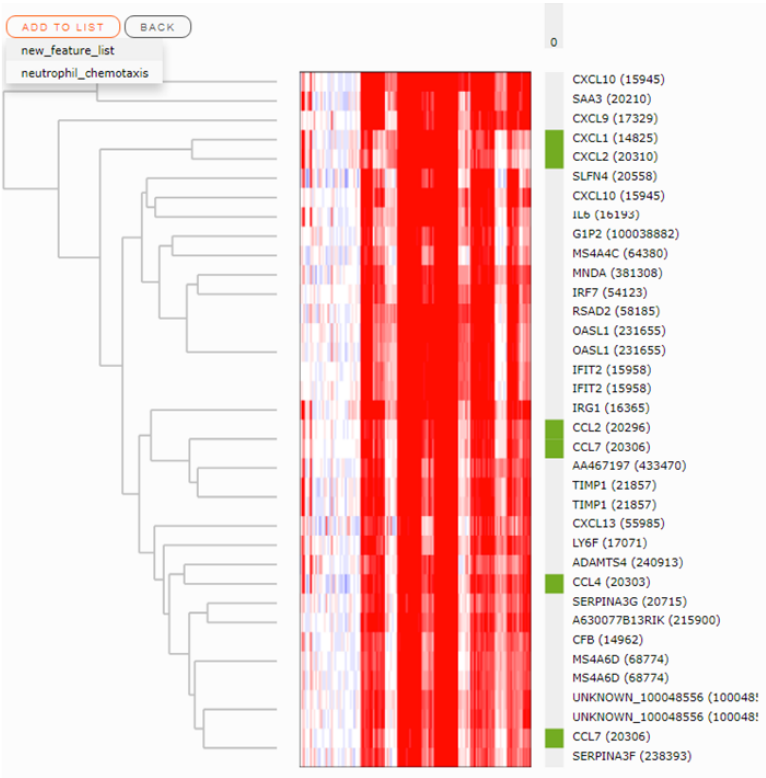
3. The newly created *feature list* will appear on mouse over of the ‘Add To List’ button in all three heatmap views. In the figure above, for instance, on mouse over of the ‘Add To List’ button in *global view*, the *feature list* named ‘neutrophil_chemotaxis’ appears.
4. To add features from the *search results panel*, click on the search terms (shown in bold). A green tick will appear, indicating that the features are ready to be added to the lists. Click on the desired *feature list* name in the ‘Add To List’ button to add the selected genes to the list. To deselect, click on the highlighted item again. In the figure above, genes belonging to the GO term ‘neutrophil chemotaxis’ are added to the *feature list* ‘neutrophil_chemotaxis’.

   Likewise in the *detailed view*, the feature labels can be individually selected and added using the ‘Add To List’ button, as shown below.



Add Features to User-created Lists from Detailed View

In the *interactive dendrogram view*, all features in the current view can be added to a list directly using the 'Add To List' button, as shown below.

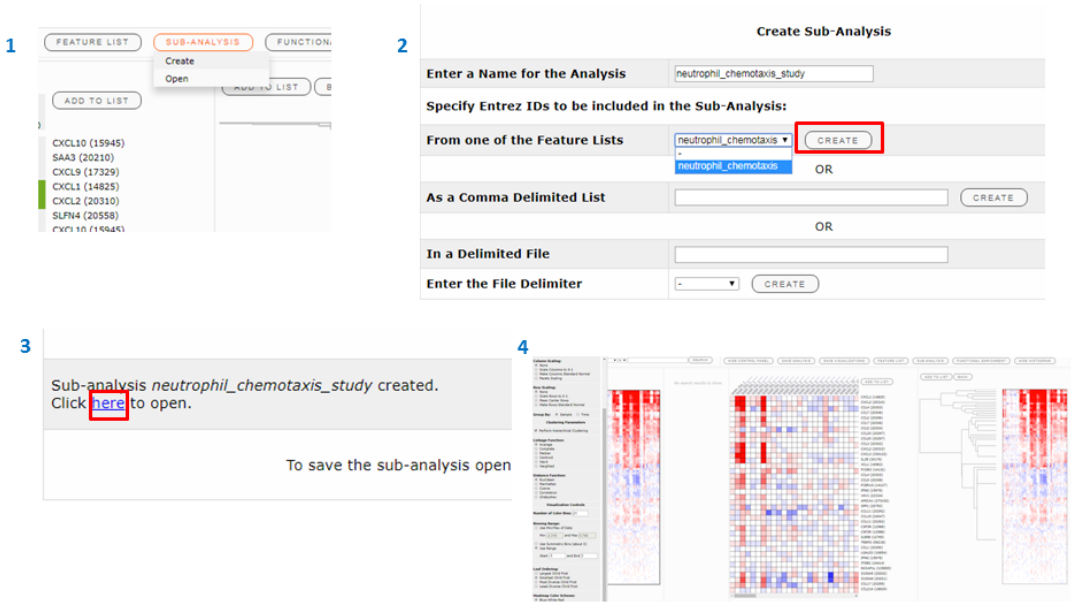


Add Features to User-created Lists from Interactive Dendrogram View

Using the feature-list menu the created feature lists can be viewed, saved to external delimited text files or deleted as required.

E. Sub-analysis

A *sub-analysis* can be created from the user-created *feature lists* and visualized in a separate window. Each new *sub-analysis* creates a new set of visualizations, where further querying and clustering can be performed, independent of the original analysis. Multiple sub-analyses can be recursively created from the existing ones.



Create Sub-analysis

To perform a sub-analysis, do the following:

1. Mouse over the 'Sub-Analysis' button and click 'Create'.
2. Enter a new name for the *sub-analysis*, select a relevant *feature list* from the dropdown list, and click 'Create'. Alternatively, specify a comma-separated string of features or a delimited text file with feature names in it, and click the corresponding 'Create' button.
3. Upon successful creation of *sub-analysis*, a link is displayed in the modal window. Click the link ('here') to open the sub-analysis in a new tab in the browser.
4. Apply the *control panel* parameters to customize the visualizations.

Note: Each new *sub-analysis*, inherits only the transformed data (not the original data) from the analysis where it is created. The transformed data becomes the raw data for the newly spawned sub-analysis. Therefore, any additional transformations applied in the sub-analysis is effectively applied on the transformed data. For instance, if a sub-analysis is created after log transforming the data, it inherits the log-transformed data as its raw data. Re-applying the log transformation in the sub-analysis will therefore log transform the already log transformed data.

Also, since a sub-analysis contains a subset of features the hierarchical clustering cannot be propagated to it. Therefore, hierarchical clustering must be re-applied in the sub-analysis.

F. Annotation of Biological Functions (Enrichment Analysis)

In SLIDE, users can perform statistical test of enrichment using the popular hypergeometric test on selected *feature lists* (gene sets). SLIDE visualizes the enriched biological pathways or GO terms using a similar heatmap-driven interface.

Create Enrichment Analysis

To initiate Enrichment Analysis, do the following:

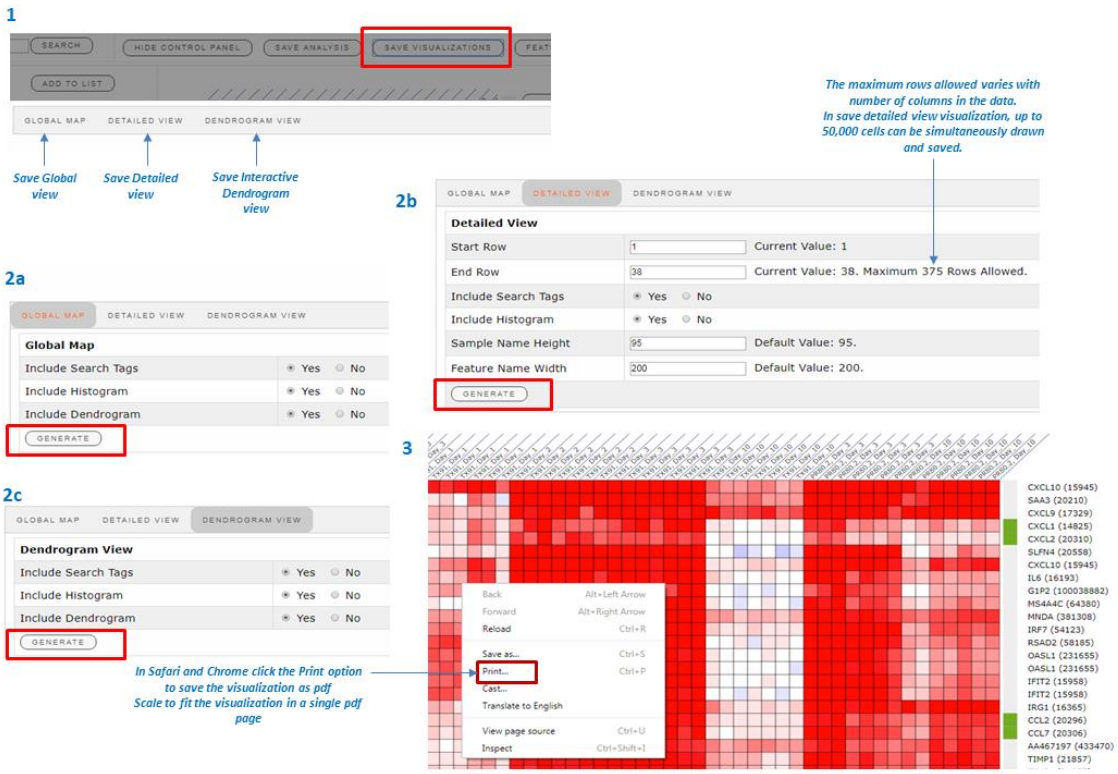
1. Click on ‘Enrichment Analysis’ button to open the enrichment analysis input form. Enter a suitable name for the analysis and do the following:
    a. Select the Enrichment Type: Pathway Enrichment or Gene Ontology Enrichment
    b. Select a *feature list* to be used in the analysis and click the ‘Add’ button. Continue adding as many *feature lists* as desired in a similar manner. Two feature lists were added in the example above.
    c. Set the desired significance level used to determine significantly enriched functional terms. *Default* is 0.05. Only functional terms with significance level less than this value are visualized.
    Note: This parameter can be changed in the control panel later.
    d. Set the minimum overlap (no. of common genes) between each functional group and *feature list*. The *default* is 0. See note below for details.
    e. Set the minimum size (no. of genes) of the functional group. The *default* is 0.
    Note: In hypergeometric test, functional groups that have very few genes can become enriched if they have even one gene common with the feature list. The above two parameters (d and e) are used to filter out such functional terms. These parameters can be changed in the control panel later.
    f. Select GO terms to be included in the analysis: biological process (BP), molecular function (MF) or cellular components (CC). The *default* is to include all categories. This parameter is only used for GO enrichment, but not for pathway enrichment analysis.
2. Upon successful creation of enrichment analysis a link is displayed in the modal window. Click the link (‘here’) to open the enrichment analysis in a new tab in the browser.
3. Apply the *control panel* parameters to customize the visualization.

The colors in the heatmaps represent the magnitude of statistical significance ($-\log_{10}(p_{value})$). Darker red color indicates greater statistical significance of enrichment. Each column in the heatmap represents a user-created feature list and each row represents a functional term. Unlike the *control panel* for feature-level visualization, the data scaling options are not available in group-level visualization. In group-level visualization, the search is limited only to functional terms.

G. Save Visualization

SLIDE provides several customization options while saving visualizations. Users can choose to save the *global view*, *detailed view* and *interactive dendrogram view*. One can choose to include the search tags as well as the histogram in the image. When saving the *detailed view*, user can also specify the start and end rows.

The steps for saving is web-browser specific. In Internet Explorer, right-click and select 'Save as Picture' to save the image in SVG or PNG formats, or use 'Print' from Tools menu to save in PDF format. In Chrome, right-click and select 'Print' option to save the visualization in PDF format. In Safari, use 'Export As Pdf' option in the file context menu to save the visualization in PDF format.



Save Visualizations

To save visualizations, do the following:

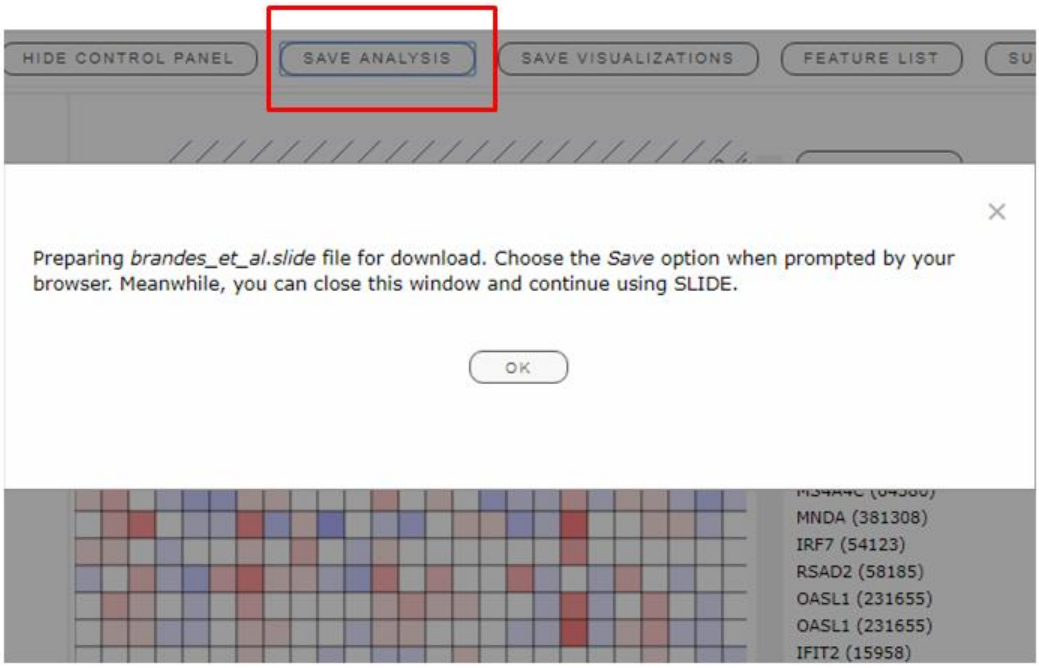
1. Click on 'Save Visualization' button. This will open a modal window with three tabs, 'Global Map', 'Detailed View' and 'Dendrogram View'.
2. Select a preferred view to save

a. Selecting ‘Global Map’ gives the option to customize the *global view* heatmap, e.g., including or excluding search tags, histogram and dendrogram (only if clustering was performed).
b. In ‘Detailed View’, in addition to the search tags, histogram and dendrogram customization, the number of rows as well as the range of rows can be specified by the user. Users can also customize the feature label and sample label widths to ensure sample and feature names are not truncated.
c. The ‘Dendrogram View’ has similar customization options to those in ‘Global Map’, described in 2a.

3. Click ‘Generate’ button to generate the visualization. This opens a new browser tab with the visualization. As mentioned above, saving the visualization in SLIDE is web-browser specific.
Note: The scale may have to be adjusted to fit the visualization in a single page when saving as pdf. Using Internet Explorer the visualization can be saved in SVG format, a format that is resolution independent and can be used to generate very high-quality images.

H. Save Analysis

Clicking the ‘Save Analysis’ button saves the entire workspace as a *.slide* file. This file can be later uploaded back into SLIDE for continued analysis.



Save Analysis