

Systems-Level Interactive Data Exploration

(SLIDE v1.1)

USER MANUAL

SOUMITA GHOSH

ABHIK DATTA

2 MAY, 2018

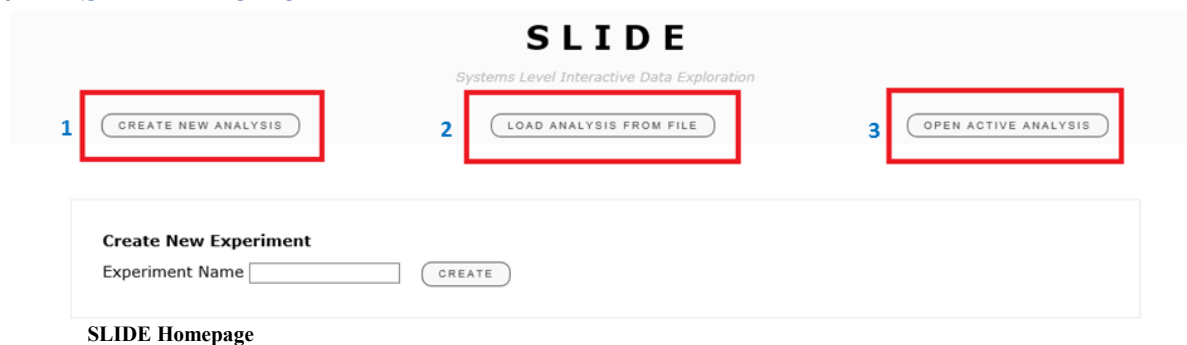
Table of Contents

I.	SLIDE Home	2
II.	Inputs to SLIDE	3
1.	Input Data.....	3
2.	Sample Information File.....	3
III.	Create New Analysis	6
IV.	Visualization Home	9
A.	Control Panel.....	10
B.	Heatmap Views.....	11
C.	Search.....	11
D.	Feature List.....	12
E.	Sub-analysis	14
F.	Annotation of Biological Functions (Enrichment Analysis).....	15
G.	Save Visualization	16
H.	Save Analysis	17
V.	Issue Tracking	18

Systems-Level Interactive Data Exploration (SLIDE) is a user-driven interactive visualization tool for large-scale -omics data. This document describes the step-by-step guideline 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>.

The web version of SLIDE is available at <http://137.132.97.109/VTBox/>

I. SLIDE Home



SLIDE's homepage contains three buttons:

1. **CREATE NEW ANALYSIS** – To enter a name for the workspace and start a new analysis
2. **LOAD ANALYSIS FROM FILE** – To upload a saved workspace (.slide file) for continued analysis
3. **OPEN ACTIVE ANALYSIS** – The button lists all currently running analyses (workspaces including sub-analysis and enrichment analysis that have not been explicitly closed). Clicking a running analysis re-opens it in a new window/tab.

II. Inputs to SLIDE

1. Input Data

SLIDE takes an input data file containing the matrix of expression values in a text delimited format (comma, tab, space, semi-colon, pipe). The data file should not contain single or double quotes. For querying and tagging of features (genes), the input data file must have at least one column containing the following identifiers (referred to as meta data column in SLIDE): Entrez ID, official gene symbols, Ensemble IDs (gene or transcript), RefSeq Accession ID or Uniprot ID. It is still possible to visualize the data without any standard identifiers, although search and tagging of genes will not be available in this case.

A snapshot of the input data file is shown below. The full example data file can be downloaded from https://github.com/soumitag/SLIDE/blob/master/data/Brandes_et_al_GSE42638_Quantiled_Log_Transformed_Data.txt.

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

Caveat: SLIDE is primarily a visualization tool for whole or subsets of genes of interest, and therefore it offers only limited data pre-processing functionalities, including: replicate handling, basic data imputation, and column and row scaling. For advanced pre-processing, users can also first process their data (e.g. normalization of expression values by taking ratios to reference/baseline samples) using R, Matlab, and use SLIDE for visualization purposes afterwards. See **Section IV A** for further details on these functionalities.

2. Sample Information File

Sample information file should be prepared in a **text file with a specific delimiter** (comma, tab, space, semi-colon, pipe), where each line contains the meta data for a sample, whose quantitative molecular data is in a column in the input data. The file is used to infer replicate and sample group information, which determines column ordering during visualization.

The structure of the sample information files varies slightly for different analysis types. Suppose the data contains two experimental conditions. If there are samples in a treatment group and a control group, then the input data can be arranged as follows:

Single Factor Experiment

Treatment	Treatment	Treatment	Treatment	Treatment	Control	Control	Control	Control	Control
Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10

If the same sample group name, i.e. Treatment, is provided (in this case for Patient_1 to Patient_5), SLIDE recognizes that there are five replicates for the Treatment group. We refer to this type of experiment as having a single factor in the experimental design: treatment vs control. SLIDE also allows for grouping samples based on two factors:

Two Factor Experiment

Treatment	Treatment	Treatment	Treatment	Treatment	Control	Control	Control	Control	Control
Dose 1	Dose 1	Dose 2	Dose 2	Dose 3	Dose 1	Dose 1	Dose 2	Dose 2	Dose 3
Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10

We refer to the above description of experiment as having two factors: 1. treatment vs control and 2. Dose_1 vs Dose_2 vs Dose_3. The structure of the sample information file for various cases is described below:

No.	Analysis Type	Sample Information File Structure
1	Analysis without any sample grouping information and with no replicates	No sample information file is required in this case.
2	Group comparisons with a single factor	<p>The first entry in each line is a column header name (sample name) in the input data file and the second entry is the sample group information (e.g. Treatment / Control).</p> <p>If replicates are present, they must share the same sample group name.</p>
3	Group comparisons with two factors	<p>The first entry in each line is a column header name (sample name) in the input data file, the second entry is the first sample group information (for factor one, e.g. Treatment / Control), and the third entry is the second sample group information (for factor two, e.g. Dose_1, Dose_2, Dose_3).</p> <p>Replicates must share the same names for the two sample groups. For instance, in the two factor experiment depicted above Patient_1 and Patient_2 are inferred as replicates for the Treatment, Dose_1 group.</p>
4	Analysis with replicates but no other sample grouping factors	<p>In this case, the replicates should be considered as a single sample grouping factor.</p> <p>The first entry in each line is a column header name in the input data file and the second entry is the sample name. Replicates must share the same sample name.</p>

Note: In all cases, **the sample names in the sample information file must be identical to the column headers in the input data file**. Any extra lines in the sample information file should start with '#'. Likewise, any headings in the sample information file must begin with a # (as shown in the above example). Both files can contain empty lines for ease of formatting. Lines beginning with # will be regarded as comments and will not be processed. The metadata columns (i.e. columns containing row identifier information like gene symbols or entrezs) should not be included in the sample information file.

The sample names in the sample information file are case-sensitive and must not contain white spaces. These names should be identical to the column headers in the input data file.

Example Datasets and Sample Information Files

Two example input data files and their corresponding sample information files are available at: <https://github.com/soumitag/SLIDE/tree/master/data>.

The first example (input data file: *Tan_et_al_Post_Pre_Mouse_Infection_Data.txt*, sample information file: *Tan_et_al_Post_Pre_Mouse_Infection_Sample_Information.txt*) is a dataset with one grouping factor comprising of two experimental conditions (pre-infection and post-infection). The input data file has a total of 10 columns, of which the first two columns (geneID, symbol) are the row identifiers – we refer to these as meta data columns in SLIDE. The rest of the columns containing expression data are referred to as non-metadata columns. The sample information file contains the group information about the eight non-metadata columns – see Sample Information File A below.

Note: The sample information file contains just the column headers and the sample group information only (no other meta data). Also note that the header starts with a ‘#’.

Sample Information File A

#Column Names	Sample Group
Pre-infection_1	Pre_infection
Pre-infection_2	Pre_infection
Pre-infection_3	Pre_infection
Pre-infection_4	Pre_infection
Post-infection_1	Post_infection
Post-infection_2	Post_infection
Post-infection_3	Post_infection
Post-infection_4	Post_infection

The second example (input data file: *Brandes_et_al_GSE42638_Quantiled_Log_Transformed_Data.txt*, sample information file: *Brandes_et_al_GSE42638_Sample_Information.txt*) is of a data with two grouping where factor one specifies the virus strain/dosage of infection (SHAM, PR80.2, PR80.6, PR810, PR8100 and Tx91) and factor two specifies the different time-points post infection (Day_1, Day_2, Day_3, Day_10) that the data was collected.

Note: Combinations of the same sample group and timepoint information are recognized as replicates in SLIDE. In this data, there are seven replicates per condition.

Sample Information File B

#Column	Sample Group	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

III. Create New Analysis

Create New Analysis

Analysis Data

Select Input Data File
Select a delimited text file. The first row of the input data file must contain UNIQUE column headers.

Select Delimiter used in the Text File

Analysis Parameters

Select Species

How Many Rows Should be Read from the Data File?
The first row (containing column names) is row 0 and is always read. The first row of data is row 1.

Select a Data Imputation Strategy
If your data has missing values please select an imputation strategy, otherwise select "None".

Identify All Metadata Columns
Metadata columns contain non-expression data that should not be visualized in the heatmap. To identify a metadata column select it from the drop-down list and click ADD.

Mark Metadata Columns as Row Identifiers (if any)
If any of the metadata columns is a row identifier (such as Entrez, Genesymbol), please select them from the drop-down list, identify their type and click ADD. You can add multiple identifiers if available.

Does the Data Contain Sample Groups?
Samples can be grouped by up to two factors. For instance, the two factors can be experimental conditions (such as Treatment vs Control) and time points.

Sample Attributes

How many Sample Grouping Factors are there?

Select a Sample Information File
Click the "Download Sample Information Template" button to download a dummy sample information file for this dataset. Click "Yes!" for further information.

Select Delimiter used in Sample Information File


CREATE ANALYSIS


Selecting 'Yes' to 'Does the Data Contain Sample Groups' opens 'Sample Attributes' panel

Create New Analysis

The images above show various input arguments to the SLIDE's interface. The numbers marked in the image corresponds to those in the table below, where each argument is described in detail.

No.	Argument/Options/Buttons	Description
1	Select input data file	The input data file must be a tab, comma, space, semicolon or pipe delimited text file. The first row of the input data file must contain UNIQUE column headers. The data file should not contain single or double quotes.
2	Select delimiter used in the text file	User must specify the delimiter in the input data file. For instance, if the input data file is tab-delimited file, please select 'Tab' from the list. If the file is a comma separated file (.csv), select 'Comma' from the list.
3	UPLOAD	Click 'UPLOAD' to upload the input data file from your local disk to SLIDE's server. The upload status will be displayed next to the input field. On successful upload, the 'Analysis Parameter' tab will open up.
4	PREVIEW	Clicking 'PREVIEW' opens a snapshot of the uploaded data file. (This allows the user to verify that the data file has been correctly uploaded and to help enter some of the Analysis Parameters below).

5	Select species	Currently SLIDE supports querying and tagging functions based on genes, transcripts, protein identifiers, GO Terms and biological pathways pertaining to human and mouse. If the input data is neither human or mouse, select 'Other'. In this case SLIDE can be used for visualization, but the querying and tagging functionalities will not be available in the first release.
6	How many rows should be read from the data file?	In case the user wants to use only part of your data file in SLIDE, specify the row numbers of the file that SLIDE should read and visualize. The first row of data (i.e. the row after the column headers) is row 1. For instance, if you specify rows 1 to 100, SLIDE will read in 100 rows, i.e. row 1 through row 100 of the data. The default is 'All', in which case the whole data file will be read.
7	Select data imputation strategy	If the input data has missing values, select an appropriate imputation strategy. If the selected imputation strategy is 'None' and SLIDE encounters missing values in your data, then it will report an error and stop processing the file. If the data does not contain missing values, please select 'None'. * Users who prefer other imputation strategy such as K-nearest neighbor-based imputation should impute data in appropriate packages outside SLIDE.
8	Identify all meta data columns	Meta data columns contain non-expression values, i.e. attributes of features (genes) that can be used as labels in the heatmaps. The dropdown list displays all the column headers from the uploaded data file. Select each metadata column and click 'Add' to identify them. The remaining columns that have not been identified as a part of the meta data will be processed as count/expression data and therefore they must contain only numeric data. SLIDE will report an error if there are non-numeric values in columns not identified as meta data. Adding a column header as metadata creates a tag like this:  To delete a tag, click the 'X' inside it.
9	Mark meta data columns as row identifiers (if any)	At least one of the metadata columns identified in Section 8 contains gene identifiers: 1. Entrez 2. Gene symbol

		<p>3. Ensembl 4. Refseq 5. Uniprot</p> <p>then create a meta data column -> identifier mapping. To create a mapping select the meta data column and the associated identifier type, and click 'Add'.</p> <p>These mapping(s) allows querying and tagging of biological functions in the downstream analysis.</p> <p>Adding a meta data column -> identifier mapping creates a tag like this:</p>  <p>To delete the mapping, click the 'X' inside the tag.</p>
10	Does the data contain sample groups?	If the input data contains sample groups (based on factors such as Treatment vs Control), selecting 'Yes' will allow SLIDE to visualize samples in different groups / orders. See Section 2. Sample Information File for details. The default is 'No' in which case no sample information file is required.
11	How many sample grouping factors are there?	SLIDE allows grouping samples based on up to two factors. For instance, the two factors can be: viral strain and time point. For each factor the samples can be grouped in different ways. Viral strain for instance can have three groups and for each strain there can be multiple time points. See Section 2. Sample Information File for details.
12	Select a sample information file	A sample information file is mandatory when the data contains sample groups or if the data has replicates. The sample information file can be tab, comma, space, semicolon or pipe delimited text file. The sample information file can have grouping information up to two factors.
13	Select delimiter used in sample information file	User must specify the delimiter used in the sample information file. For instance, if the sample information file is tab-delimited file, please select 'Tab' from the list. If the file is a comma separated file (.csv), select 'Comma' from the list.
14	UPLOAD	Clicking UPLOAD will upload the sample information file to the server.
15	Download sample information template	A sample template sample information file for the uploaded data can be generated using the 'Download Sample Information Template' button. The downloaded file can be edited outside SLIDE and then used in SLIDE. Please refer to Section II for details on editing/creating a sample

		information file.
16	CREATE ANALYSIS	Click to start visualizing your data.
17	HELP	<p>The help menu (top-right of the page) provides links to:</p> <ol style="list-style-type: none"> 1. this User Manual, 2. a GitHub link for reporting persistent issues (see section V below for details), 3. an example data file and 4. an example sample information file

IV. Visualization Home



A: Control Panel

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 (which may be suboptimal at first). For any data set, a first step is 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 appropriate binning range and appropriate data transformation, the changes in expression level may not be visualized properly (all the cells in the heatmap may appear in similar colors).

If the data has already been pre-processed (for instance, log transformed and mean centered), the user needs to change only the visualization parameters and may ignore the data transformation parameters. If the relative patterns of expression level change between genes are 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.

IMPORTANT: The parameters for data transformation are applied in the order they appear in the control panel (top to bottom). 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 parameter and applies 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. ‘Show All Replicates’ displays all replicates, ‘Mean’ displays replicate average value and ‘Median’ displays replicate median for each feature. The *default* is ‘Show All Replicates’.
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 Scaling:** Column scaling independently standardizes the values in each sample to similar ranges. The *default* is ‘None’. The ‘Modified Pareto Scaling’ transforms the data matrix \mathbf{X} as follows:

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

here, μ_j and σ_j are the mean and standard deviation for the j^{th} column (sample), x_{ij} is the ij^{th} element of \mathbf{X} , and N is the total number of elements in \mathbf{X} .

5. **Row Centering:** The range of values in each row (feature) of the input data file may vary widely (e.g. abundance levels are wildly variable). Row centering removes the bias of visualization due to variable abundance levels. The ‘Mean Center Rows’ option independently shifts each row so that their means are at 0. The *default* is ‘None’.
6. **Group Columns By:** This option allows grouping the samples based on the sample group factor one or two.
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 option* is performing no hierarchical clustering.
Note: 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 fixed set 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. **Row Label:** If at least one **meta-column->identifier mapping** is provided, the standard identifiers (such as Entrez, Gene Symbol, RefSeq, Ensembl and Uniprot identifiers) will be available for use as row labels in the heatmaps. If no **meta-column->identifier mapping** is provided but unmapped meta data columns are provided, these columns will be available for use as row labels in the heatmaps. If no mappings or meta data columns are present in the data, default labels ('-') will appear as row labels.
11. **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.
12. **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).
Note: The histogram (right side of the screen) can be used to check the data range and adjust the binning range accordingly.
13. **Leaf Ordering:** A leaf ordering scheme determines the placement of genes/samples in the dendrogram tree. The output of hierarchical clustering, i.e. 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'.
14. **Heatmap Color Scheme:** SLIDE has four commonly used color schemes available for heatmaps. The default is the blue-white-red color scheme.

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/unclustered 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, or biological functions/pathways. The associated genes are tagged in vertical bars next to all heatmaps. Wildcard search, using approximate keywords, can be performed 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*.

1 2 3 4

Entrez ID = [] SEARCH

Search Bar

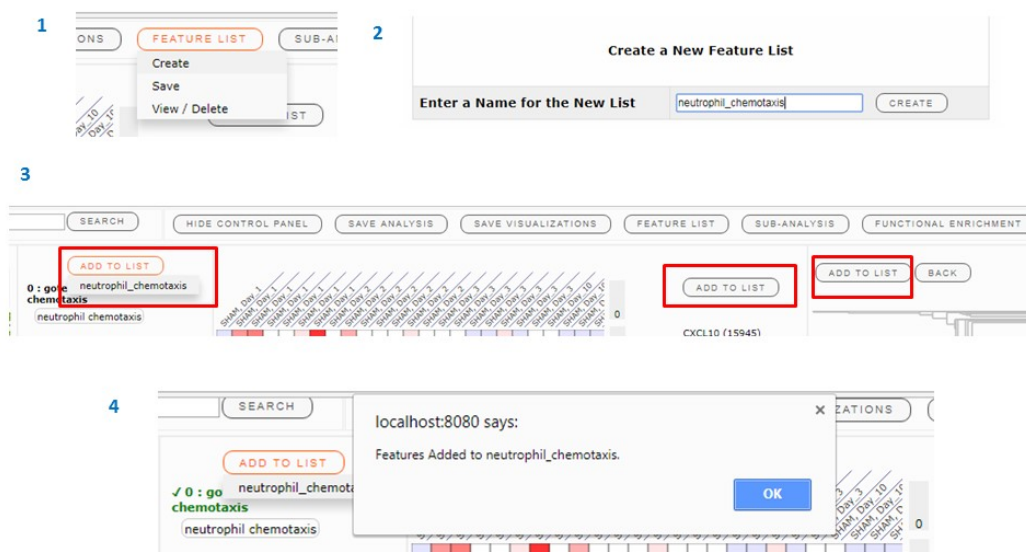
To use the ‘search’ functionality, do the following:

1. Select one of the available query types from the drop-down list:
Entrez ID, Gene symbol, RefSeq ID, Ensembl gene ID, Ensembl Transcript ID, Ensembl Protein ID, Uniprot ID, Gene Ontology ID, GO Term, Pathway ID or Pathway Name
2. **Select the type of search: exact (=) or wildcard (≅).** The default is set to exact search.
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 (genes) 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

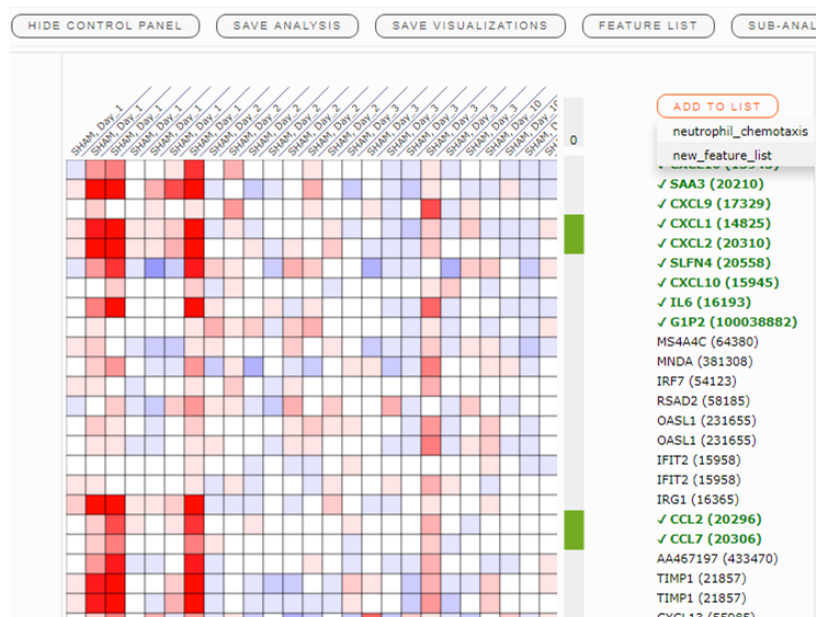
To add features to a **list**, do the following to create a feature list:

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:

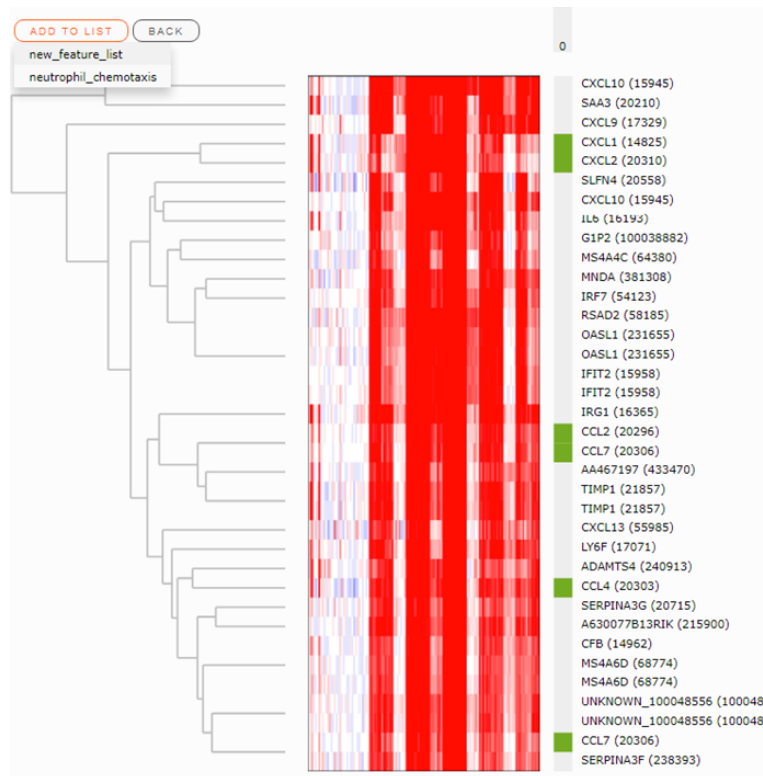
3. The newly created *feature list* will appear on mouse over of the **ADD TO LIST** button in all three heatmap views. In the figure above, for instance, on mouse over of the **ADD TO LIST** button in *global view*, the *feature list* named 'neutrophil_chemotaxis' appears.
4. To add features from the *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 wished by the user.

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.

-
-
- To save the sub-analysis open
-

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.

IMPORTANT: 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. 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.

1. **FUNCTIONAL ENRICHMENT**

2. **Create Enrichment Analysis**

Enter a Name for the Enrichment Analysis: enrichment_analysis_study

Enter the Enrichment Type: Pathway

Select Feature Lists to be included in the Enrichment Analysis and Click Add

Select Feature Lists: up-regulated_features_list, down-regulated_features_list

Include Functional Groups with p-value Lower Than: (Significance Level): 0.05

Include Functional Groups That Contain At Least These Many Genes: 5

Include Functional Groups That Contain At Least These Many Genes: 5

Include Ontologies (Only used for Gene Ontology Enrichment): ☐ Biological Processes ☐ Molecular Function ☐ Cellular Components

3. **Enrichment Parameters**

Significance Level: 0.05

Minimum Functional Group Feature List Intersection: 5

Minimum Functional Group Size: 5

Clustering Parameters

☒ Perform Hierarchical Clustering

Linkage Function: ☒ Average ☐ Complete ☐ Median ☐ Centroid ☐ Ward ☐ Single

Distance Function: ☒ Euclidean ☐ Manhattan ☐ Cosine ☐ Correlation ☐ Chebyshev

Visualization Controls

Number of Color Bins: 25

Scaling Ranges

☒ Use Min/Max of Data ☐ Use Z-Score ☐ Use Symmetric Bins (about 0)

☐ Use Range

Start: and End:

Leaf Ordering

☒ Largest Child First ☐ Smallest Child First ☐ Heat Shrink Child First ☐ Largest Shrink Child First

Heatmap Color Scheme:

4. **Enrichment analysis enrichment_analysis_study created. Click here to open.**

To save the enrichment analysis open it and click the "Save Analysis" button.

5. **Enrichment Analysis Results**

6. **Enrichment Analysis Results**

7. **Enrichment Analysis Results**

8. **Enrichment Analysis Results**

9. **Enrichment Analysis Results**

10. **Enrichment Analysis Results**

11. **Enrichment Analysis Results**

12. **Enrichment Analysis Results**

13. **Enrichment Analysis Results**

14. **Enrichment Analysis Results**

15. **Enrichment Analysis Results**

16. **Enrichment Analysis Results**

17. **Enrichment Analysis Results**

18. **Enrichment Analysis Results**

19. **Enrichment Analysis Results**

20. **Enrichment Analysis Results**

21. **Enrichment Analysis Results**

22. **Enrichment Analysis Results**

23. **Enrichment Analysis Results**

24. **Enrichment Analysis Results**

25. **Enrichment Analysis Results**

26. **Enrichment Analysis Results**

27. **Enrichment Analysis Results**

28. **Enrichment Analysis Results**

29. **Enrichment Analysis Results**

30. **Enrichment Analysis Results**

31. **Enrichment Analysis Results**

32. **Enrichment Analysis Results**

33. **Enrichment Analysis Results**

34. **Enrichment Analysis Results**

35. **Enrichment Analysis Results**

36. **Enrichment Analysis Results**

37. **Enrichment Analysis Results**

38. **Enrichment Analysis Results**

39. **Enrichment Analysis Results**

40. **Enrichment Analysis Results**

41. **Enrichment Analysis Results**

42. **Enrichment Analysis Results**

43. **Enrichment Analysis Results**

44. **Enrichment Analysis Results**

45. **Enrichment Analysis Results**

46. **Enrichment Analysis Results**

47. **Enrichment Analysis Results**

48. **Enrichment Analysis Results**

49. **Enrichment Analysis Results**

50. **Enrichment Analysis Results**

51. **Enrichment Analysis Results**

52. **Enrichment Analysis Results**

53. **Enrichment Analysis Results**

54. **Enrichment Analysis Results**

55. **Enrichment Analysis Results**

56. **Enrichment Analysis Results**

57. **Enrichment Analysis Results**

58. **Enrichment Analysis Results**

59. **Enrichment Analysis Results**

60. **Enrichment Analysis Results**

61. **Enrichment Analysis Results**

62. **Enrichment Analysis Results**

63. **Enrichment Analysis Results**

64. **Enrichment Analysis Results**

65. **Enrichment Analysis Results**

66. **Enrichment Analysis Results**

67. **Enrichment Analysis Results**

68. **Enrichment Analysis Results**

69. **Enrichment Analysis Results**

70. **Enrichment Analysis Results**

71. **Enrichment Analysis Results**

72. **Enrichment Analysis Results**

73. **Enrichment Analysis Results**

74. **Enrichment Analysis Results**

75. **Enrichment Analysis Results**

76. **Enrichment Analysis Results**

77. **Enrichment Analysis Results**

78. **Enrichment Analysis Results**

79. **Enrichment Analysis Results**

80. **Enrichment Analysis Results**

81. **Enrichment Analysis Results**

82. **Enrichment Analysis Results**

83. **Enrichment Analysis Results**

84. **Enrichment Analysis Results**

85. **Enrichment Analysis Results**

86. **Enrichment Analysis Results**

87. **Enrichment Analysis Results**

88. **Enrichment Analysis Results**

89. **Enrichment Analysis Results**

90. **Enrichment Analysis Results**

91. **Enrichment Analysis Results**

92. **Enrichment Analysis Results**

93. **Enrichment Analysis Results**

94. **Enrichment Analysis Results**

95. **Enrichment Analysis Results**

96. **Enrichment Analysis Results**

97. **Enrichment Analysis Results**

98. **Enrichment Analysis Results**

99. **Enrichment Analysis Results**

100. **Enrichment Analysis Results**

Create Enrichment Analysis

To initiate Enrichment Analysis, do the following:

1. Click on the '**FUNCTIONAL ENRICHMENT**' button to open the enrichment analysis input form.

Enter a suitable name for the analysis and do the following:

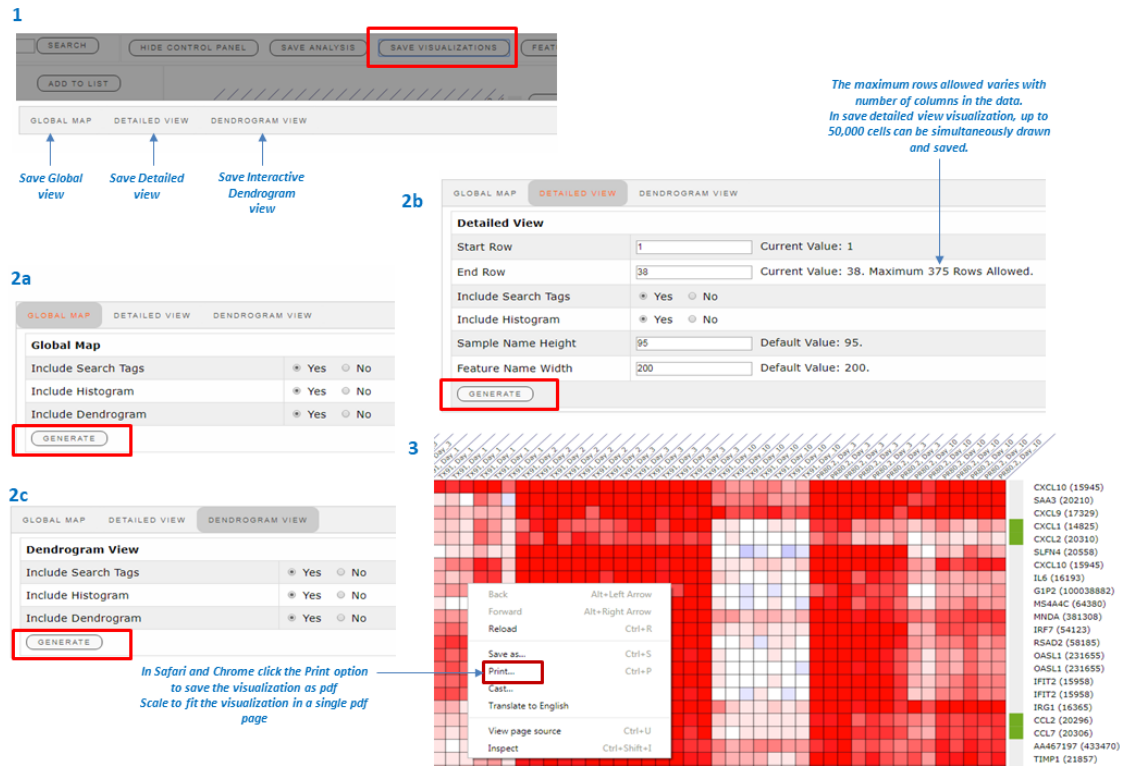
- a. Select the Enrichment Type: Pathway Enrichment or Gene Ontology Enrichment
 - b. Select a *feature list* to be used in the analysis and click the '**Add**' button. Continue adding as many *feature lists* as desired in a similar manner. Two feature lists were added in the example above.
 - c. Set 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 the GO term categories 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 will be 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

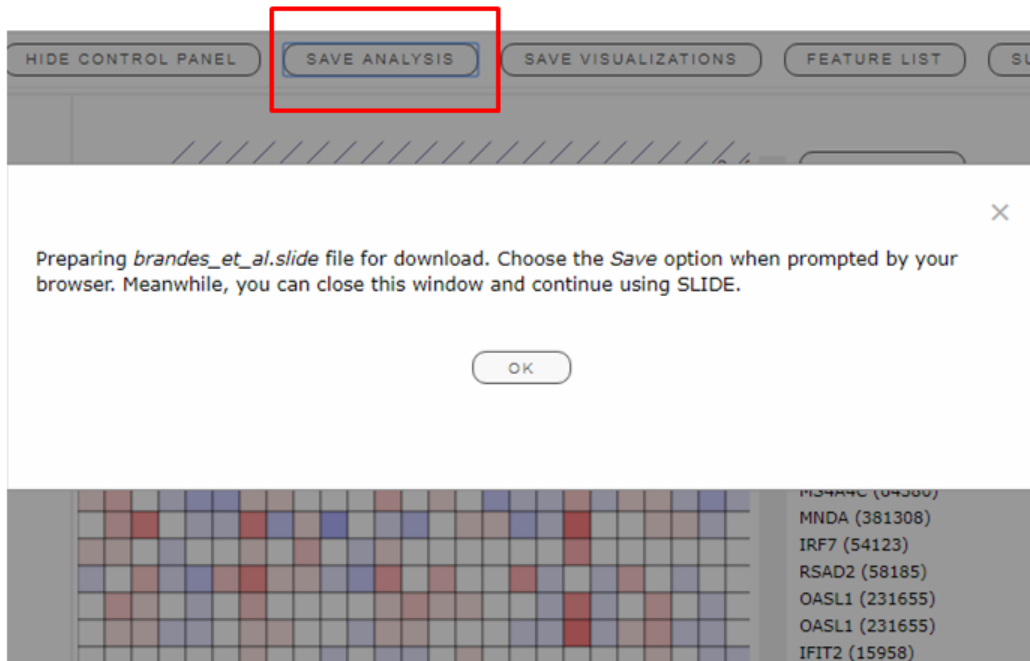
To save visualizations, do the following:

1. Click on 'SAVE VISUALIZATIONS' 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.



V. Issue Tracking

In case of any persistent issues, please report it using GitHub's issue tracker:
<https://github.com/soumitag/SLIDE/issues>.