DToxS: SVD-based pipeline for identification of condition-selective differentially expressed genes and outlier responses

Hansen et al. 2024

Isolated use of the R-scripts 'SVD1-17' for own datasets

For the identification of condition-selective differentially expressed genes in own datasets using the R-scripts 'SVD1-17' in isolation, follow these instructions:

1) Open the file "SVD_global_parameter.R".
   a) Specify the "overall_lincs_directory" on your hard drive that contains all downloaded subdirectories, e.g. "D:/LINCS_DToxS_SVD/".
   b) Select the fraction of available cores ("fraction_of_used_cores") for parallel processing.
   c) Add the names of your dataset(s) that shall be subjected to the SVD pipeline to the array "datasets".
   d) If you want to project one dataset into the subspaces identified by another dataset, add both datasets as a key-value pair to the inputDataset_inputDatasetForSvd_list. The key is the dataset that will be projected into the value's dataset subspaces. In this case the treatment conditions of interest (see below) have to overlap in both datasets. Ensure that the treated entities (e.g., cell lines, see below) have different names in both datasets. The key datasets should not be mentioned in the datasets array (1c). If you do not have datasets that should be projected into another dataset's SVD space, leave the list empty.
   e) If you want remove any eigenarrays from the data before identification of drug-selective subspaces, add the dataset and the numbers of the eigenarrays to be removed as a key-value pair to the list "substract_svds". Here, you can remove eigenarrays whose contribution to a gene expression profile correlates with the number of its significantly expressed genes. Such eigenarrays would represent a general response that increases with the magnitude of the perturbation and is independent of the treatment condition. The SVD pipeline will analyze how much this is the case for each eigenarray. Leave the list empty for no removals.
      Each key defined in "substract_svds" has to be added as a key to the list "inputDataset_inputSVDDataset_list". The value is the datasets whose eigenarrays will define the removed components. In the regular case the keys and values should be the same.
      The datasets after removal of the selected eigenarrays will have the same name as the initial dataset plus the ending "_no1stSVD". This is just a name that is independent of the removed eigenarrays. If the generated dataset shall be subjected to the SVD pipeline as well, they have to be added to the dataset array (1c).
2) Prepare your data. The following instructions have to be followed for each dataset that you specify in the SVD_global_parameter.R within the dataset array (1c) and the inputDataset_inputDatasetForSvd_list (1d), except for those datasets that will be generated by removal of selected eigenarrays (1e). For those datasets the following steps can be ignored, since the SVD pipeline will automatically generate all directories and files. The files and directories in the folder "Results_to_run_R_SVD1to17_in_isoloation" give an example for the following instructions.

a) Generate a subdirectory in the “Results/” folder whose name is identical with the dataset name. Within the subdirectory create the subdirectory “1_DEGs/”.
b) Generate a file containing your data that has the name of your dataset plus “_topall.txt” (e.g., “SVD_DEGenes_iPSCdCMs_P0_Signed_minus_log10pvalue_topall.txt”). Your data should be part of a tab-delimited spreadsheet. The first column of the spreadsheet should contain the features, e.g. genes, and be called “Symbol” in the headline. All other columns have to contain expression values, e.g., $\log_2$(fold changes) or signed minus $\log_{10}$(p-values). The larger the absolute value, the more significant the expression of a given feature in a given condition. Positive and negative values indicate up- and downregulation, respectively. The file should contain all identified features for each condition without prior application of any significance filters. The names for the conditions should contain six entries that are separated by hyphens. The first two entries should be “Diffrna” and “H48”, the last “Plate.0”. The third entry can specify a class, e.g. a drug class, the fourth the treated entity, e.g., cell line, subject or animal model and the fourth the treatment condition of interest, e.g. a drug or disease. Here is an example condition from our LINDS DToxS data: “Diffrna-H48-Anthracycline-Cell_line.MSN01_03R-DAU-Plate.0”. Our pipeline will search for expression profiles that are shared among all conditions sharing the same name at the fourth entry, e.g. the treatment condition of interest. The generated file should be copied into the “1_DEGs/” folder.
c) The information in the “DEG_summary.txt” file will be used to quantify how strongly each eigenarray represents a general response that increases with the magnitude of perturbation as quantified by the number of significant genes. It should contain a tab-delimited spreadsheet with the column names “Drug_type”, “Plate”, “Cell_line”, “Treatment” and “Significant_degs_based_on_FDR_count”. Each condition in the data file has to be listed in this file. Entries within the first four columns have to agree with the related entries within the condition names in the data file, e.g. the third, sixth, fourth and fifth entries, respectively (2b). The last column should contain the number of significant DEGs in each condition, e.g. using an FDR cutoff of 10%. This file should be copied into the “1_DEGs/” folder as well.

3) Open the file “SVD_0000000_main_Run_pipeline.R”. Specify the working directory that contains the downloaded R-code (e.g., "D:/LINCS_DToxS_SVD/AA_R_code/"). Set “memory_larger_than_16GB” to “FALSE” or comment the line “source('SVD_0c_Correlation_with_GTEx.R')”.
4) Copy the file “Report_finished_1st_part_by_Csharp.txt” from the subdirectory “Results_to_run_R_SVD1to17_in_isolation/” into the “Results/” subdirectory or comment “source('SVD_0___wait_for_C#_script_to_finish.R')” in the file “SVD_0000000_main_Run_pipeline.R”.
5) Run the code. This analysis is finished, once the message "Waiting for C# script to finish 3rd part (already waited for 0 min)" appears.
6) For quick enrichment analysis of identified treatment condition-selective DEGs, followed by pathway network-based integration and visualization, download our desktop application MBCO PathNet (mbc-ontology.org) and follow the instructions in the folder (“14_DEGs_MBCO_PathNet”) within the dataset folder you generated in the “Results/” directory.

7) Optional - define colors and full names for results visualization. The user can define colors for treatment conditions, treated entity conditions and classes within the file “SVD_colors.R”. The defined colors for treatment conditions will also be imported into the MBCO PathNet application.
   a) Colors for treatment conditions can be defined in the array pair “drugs_for_highlight_colors” and “drug_highlight_colors”. Treatment conditions will be assigned to the color at the same index.
   b) Colors for treated entity conditions can be defined in the array pair: “cellLines_for_highlight_colors” and “cellLine_highlight_colors”. Treated entities will be assigned to the color at the same index.
   c) Color for classes can be defined in the list “drugClass_colors”. Keys indicate classes, values colors.
   d) Full names for abbreviations of treatment conditions can be defined in the list “full_drugNames”. Keys indicate abbreviations that have to match the fourth entry within the condition column names in the data matrix (2b), values the full names that will be visualized in the dendrograms.

   The function “Get_all_color_and_label_vectors” in the file “SVD_13_generate_all_heatmaps.R” contains the described mappings.