Script Documentation: `comparative_statistical_analysis.R`

This script is designed to perform comparative statistical analysis between two datasets, specifically focusing on the comparison of functional abundances. The script filters functions based on user-defined abundance and prevalence thresholds, checks for common functions between datasets, and performs statistical tests (Wilcoxon rank-sum) to identify significantly different functions between the groups. Additionally, it generates a variety of outputs, including Venn diagrams, barplot, heatmap and files with common and exclusive functions.

1. **Usage**

To execute the script, use the following command:

Rscript comparative_statistical_analysis.R <input_data1> <dataset_name1> <input_data2> <dataset_name2> <sample_type> <abundance_threshold> <prevalence_threshold> <padj> <log_threshold> <color_data1> <color_data2> <heatmap_color>

2. **Arguments**

<input_data1>: The input data file for the first dataset.

<dataset_name1>: The name of the first organism or condition being analyzed (e.g., "Human", "Ruminants", "decrease_RME", "increase_RME").

<input_data2>: The input data file for the second dataset.

<dataset_name2>: The name of the second organism or condition being analyzed.

<sample_type>: The type of sample (e.g., "Mags", "Genes", "miRNA").

<abundance_threshold>: The minimum abundance level used to filter functions (must be a non-negative numeric value).

<prevalence_threshold>: The minimum prevalence level for filtering functions (must be a numeric value between 0 and 1).

<padj>: The adjusted p-value threshold for significance testing (must be a numeric value between 0 and 1).

<log_threshold>: The log2 fold change threshold for filtering results (any numeric value is accepted).

<color_data1>: The color for visualizing the first dataset (e.g., "blue" or "#FF5733").

<color_data2>: The color for visualizing the second dataset (e.g., "red" or "#33FF57").

<heatmap_color>: The color palette for heatmap visualization (e.g., "purple", "green").

3. **Help**

You can access the help information by running the script with the `--help` argument:

Rscript comparative_statistical_analysis.R --help

4. **Requirements**

R Packages: The script requires the following R libraries:

- **stats:** for statistical functions and modeling.

  install.packages("stats")

- **ggplot2:** for plotting.

  install.packages("ggplot2")

- **reshape2:** for reshaping data.

  install.packages("reshape2")

- **dplyr:** for data manipulation and transformation.

  install.packages("dplyr")

- **VennDiagram:** for creating Venn diagrams.

  install.packages("VennDiagram")

- **grid:** for grid graphics.

  install.packages("grid")

- **futile.logger:** for logging messages.

  install.packages("futile.logger")

5. **Script Functionality**

a. Input Validation

The script checks if the correct number of arguments is provided. If not, it stops execution and displays usage information.

It ensures that both datasets have the same classification level and sample type, otherwise, it throws an error.

The script validates that the `prevalence_threshold` and `padj` values are within the required range (0-1).

b. Data Processing

The input files are read, and functions are filtered based on the `abundance_threshold` and `prevalence_threshold` values provided by the user.

Common and unique functions between the two datasets are identified.

c. Venn Diagram Creation

A Venn diagram is created to visualize shared and exclusive functions between the two datasets. The diagram is saved as both PDF and PNG with the following filenames:

Venn_diagram_<sample_type>_level_<dataset_name1>_vs_<dataset_name2>.pdf
Venn_diagram_<sample_type>_level_<dataset_name1>_vs_<dataset_name2>.png

d. Saving Common and Exclusive Functions

The script identifies common and exclusive functions between the datasets and saves them to text files:

Common_functions_<sample_type>_level_<dataset_name1>.txt: Abundance values of common functions for the first dataset
Common_functions_<sample_type>_level_<dataset_name2>.txt: Abundance values of common functions for the second dataset
exclusive_functions_<sample_type>_level_<dataset_name1>.txt: Exclusive functions for the first dataset
exclusive_functions_<sample_type>_level_<dataset_name2>.txt: Exclusive functions for the second dataset

e. Statistical Testing

A Wilcoxon rank-sum test is applied to compare the abundance levels of the common functions between the two datasets.

Adjusted p-values (Benjamini-Hochberg method) are calculated for the functions, and those with p-values less than the user-defined `padj` threshold are considered significant.

If no significant functions are found, the script exits with a message.

f. Result Generation

The script generates a volcano plot, barplot and heatmap to illustrate the differences in function expression between the two datasets.

- **Volcano Plot**

The volcano plot is a type of scatter plot that displays statistical significance (adjusted p-value) versus the magnitude of change (log2 fold change) between two conditions or organisms. It helps visualize both the magnitude and the reliability of differences between two datasets.

x-axis: log2 fold change, indicating the magnitude of the difference in gene expression or function between the two groups.

y-axis: -log10(p-value), which shows the statistical significance of the differences (the higher the point, the more significant it is).

Generated Output

A PNG image showing the volcano plot for the comparative analysis:

Volcano_Plot_<sample_type>_level_<dataset_name1>_vs_<dataset_name2>.png

A PDF image showing the volcano plot for the comparative analysis:

Volcano_Plot_<sample_type>_level_<dataset_name1>_vs_<dataset_name2>.pdf

- **Bar Plot**

The bar Plot visualizes the common differentially abundant functions between the first and second datasets. It provides an easy visual summary to understand which functions are more abundant in each group.

x-axis: log2 fold change, indicating the magnitude of the difference in gene expression or function between the two groups.

y-axis: Functional categories from SuperFocus.

Generated Output:

A PNG image illustrating the bar plot that highlights the functions with significant differences identified through the comparative analysis between the two groups.

Functions_significant_<sample_type>_level_<dataset_name1>_vs_<dataset_name2>.png

A PDF image illustrating the bar plot that highlights the functions with significant differences identified through the comparative analysis between the two groups.

Functions_significant_<sample_type>_level_<dataset_name1>_vs_<dataset_name2>.pdf

- **Heatmap**

The heatmap is a graphical representation of data where individual values are represented by colors.

Rows: Functional categories from SuperFocus.

Columns: Organism or experimental conditions.

Color Scale: Indicates the level of expression (e.g., red for high abundance, white for low abundance).

Generated Output: Heatmap_abundance_

A PNG image showing the heatmap, with the level of expression or abundance encoded in the color scale between the two groups.

Heatmap_abundance_<sample_type>_level_<dataset_name1>_vs_<dataset_name2>.png

A PNG image showing the heatmap, with the level of expression or abundance encoded in the color scale between the two groups.

Heatmap_abundance_<sample_type>_level_<dataset_name1>_vs_<dataset_name2>.pdf

6. **Example:**

Rscript comparative_statistical_analysis.R data1.txt "Human" data2.txt "Ruminants" "Mags" 0.5 0.2 0.05 1 "#FF5733" "#33FF57" "purple"

This command compares two datasets, filters functions based on an abundance threshold of 0.5 and a prevalence threshold of 0.2, and performs statistical analysis with an adjusted p-value cutoff of 0.05. The results will be visualized using the specified colors.