DEGs Loop (t-test & Wilcoxon) Script Documentation

The DEGs Loop (t-test & Wilcoxon) script is designed to identify differentially expressed genes (DEGs) between normal and disease samples using statistical tests. It performs normality testing, calculates log fold changes, applies multiple testing corrections, and generates visualizations (volcano plots) to highlight significant genes.

Libraries Used

- **stats:** Provides functions for statistical tests (t-test, Wilcoxon test, Shapiro-Wilk test).
- **ggplot2:** Used for creating visualizations, such as volcano plots.
- **ggrepel:** Enhances ggplot2 by preventing label overlap in plots.

Input Files

1. Normal Sample Data:

File format: .rda or .RDS.

- Contains gene expression data for normal samples.

Example: LumA_Normal.rda, LumB_Normal.rda.

2. Disease Sample Data:

File format: .RDS.

- Contains gene expression data for disease samples.

Example: LumA_Predicted.rds, LumB_Predicted.rds.

Output Files

1. Test Results:

File format: .csv.

Contains the results of statistical tests for each gene, including:

- Shapiro-Wilk test p-values (normal and disease samples).
- Test type (t-test or Wilcoxon test).
- Test statistic and p-value.
- Adjusted p-value (Benjamini-Hochberg correction).
- Log fold change.

Example: LumA_test_results_with_correction.csv.

2. Significant Genes:

File format: .csv.

Contains genes that meet the significance criteria:

- Adjusted p-value < 0.05.
- Absolute log fold change > 1.

Example: LumA_significant_genes.csv.

3. Volcano Plot:

File format: .png.

- Visualizes the relationship between log fold change and adjusted p-value.
- Highlights upregulated (red) and downregulated (blue) genes.

Example: LumA_volcano_plot.png.

Main Functions

process_samples(normal_file, disease_file, output_prefix)

This function processes a pair of normal and disease samples to identify DEGs. It performs the following steps:

1. Load Data:

- Loads normal sample data from .rda or .RDS files.
- Loads disease sample data from .RDS files.

2. Filter Genes:

- Ensures only genes present in both normal and disease datasets are analyzed.

3. Statistical Analysis:

- Performs Shapiro-Wilk tests to check for normality.
- Applies t-test (if data is normally distributed) or Wilcoxon test (if data is not normally distributed).
- Calculates log fold change for each gene.

4. Multiple Testing Correction:

- Adjusts p-values using the Benjamini-Hochberg method.

5. Save Results:

- Saves the full test results and significant genes to .csv files.

6. Visualization:

- Generates a volcano plot to visualize significant DEGs.
- Saves the plot as a .png file.

Usage

1. Set Up Input Files:

Ensure the normal and disease sample files are in the correct format and located in the working directory.

2. Define Samples:

Update the samples list in the script to include the file paths and output prefixes for each sample pair.

Notes

1. Normality Assumption: The script uses the Shapiro-Wilk test to determine whether to apply a t-test (for normally distributed data) or a Wilcoxon test (for non-normally distributed data).
2. Significance Criteria:

– Genes are considered significant if:

- Adjusted p-value < 0.05.
- Absolute log fold change > 1.

3. Volcano Plot:

– The volcano plot highlights upregulated (red) and downregulated (blue) genes.

– Dashed lines indicate thresholds for significance (p-value = 0.05) and fold change (|log2FC| = 1).

4. Multiple Testing Correction: The Benjamini-Hochberg method is used to control the false discovery rate (FDR).