Preprocessing RNA Data Script Documenta

The Preprocessing RNA Data script is designed to process RNA-seq count matrices to ensure high-quality input for downstream analysis. It includes steps for loading raw count data, mapping Ensembl IDs to gene symbols, filtering low-expression genes, normalizing expression values, and handling outliers. The script supports multiple input formats and automates preprocessing steps to streamline data preparation.

Libraries Used

- **edgeR:** For normalization and filtering of RNA-seq count data.
- **biomaRt:** For gene ID mapping from Ensembl.
- **dplyr:** For data manipulation and aggregation.
- **e1071:** For skewness-based threshold adjustments.
- **limma:** For linear modeling and expression analysis.

Input Files

– RNA-seq count matrix (CSV or TSV format): Contains raw gene expression counts.

Output Files

– Processed RNA count matrix (.rda format): A serialized R data file containing the preprocessed RNA-seq data.

Main Functions

preprocess_count_matrix(file_path, dataset = "hsapiens_gene_ensembl")

- Loads the raw RNA-seq count matrix.
- Maps Ensembl gene IDs to gene symbols.
- Filters out lowly expressed genes.
- Normalizes expression values.
- Adjusts outliers to improve data consistency.

load_count_matrix(file_path)

- Reads RNA-seq count matrices in CSV or TSV format.
- Checks file extensions and handles unsupported formats.

map_and_aggregate(count_matrix, dataset)

- Converts Ensembl IDs to gene symbols using biomaRt.
- Aggregates expression counts for duplicated gene symbols.

filter_lowly_expressed_genes_and_handle_zeros(count_matrix)

- Normalizes gene expression using TMM (Trimmed Mean of M-values) from edgeR.
- Filters out genes with low expression variability using median absolute deviation (MAD).
- Identifies missing values and applies dynamic thresholding to handle them.

adjust_outliers(filtered_log_count_matrix)

- Detects and adjusts outliers using the Interquartile Range (IQR) method.
- Dynamically adjusts outlier thresholds based on the dataset's skewness.

Usage

1. Set the working directory to the location of the script.
2. Ensure required input files (RNA-seq count matrices in CSV/TSV format) are available.
3. Run the script, which will:
4. Load the count matrix.
5. Map gene identifiers and normalized expression values.
6. Filter lowly expressed genes and adjust outliers.
7. Save the processed data in .rda format.

Notes

- This script assumes that the count matrix contains gene expression data with Ensembl gene IDs.
- The normalization and filtering thresholds are dynamically adjusted based on dataset characteristics.
- Users working with non-human datasets should specify the appropriate biomaRt dataset name.