Preprocessing Methylation Data Script Documentation

The Preprocessing Methylation Data script is designed to process DNA methylation beta value matrices for downstream analysis. It includes steps for loading raw methylation data, imputing missing values, normalizing beta values, and handling outliers. The script automates preprocessing to ensure high-quality data for analysis.

Libraries Used

- **IlluminaHumanMethylation450kanno.ilmn12.hg19**: Annotation for Illumina 450K methylation arrays.
- **minfi:** Methylation data preprocessing and normalization.
- **impute:** Missing data imputation.
- **dplyr:** Data manipulation and aggregation.
- **e1071:** Skewness-based threshold adjustments.
- **preprocessCore:** Quantile normalization.
- **data.table:** Efficient data handling and reading.
- **BiocParallel:** Parallel processing.
- **lumi:** Package used to change beta values into M values.

Input Files

- Methylation beta value matrix (CSV or TSV format): Contains raw methylation beta values for CpG sites.

Output Files

- Processed Methylation Matrix (.rda format): A serialized R data file containing the preprocessed methylation data.

Main Functions

preprocess_methylation_matrix(file_path)

- Loads the beta value matrix.
- Imputes missing values and filters out low-quality data.
- Normalizes beta values using quantile normalization.
- Adjusts outliers to improve data consistency.

load_beta_value_matrix(file_path)

- Reads methylation beta value matrices in CSV or TSV format.
- Aggregates duplicate CpG site values by averaging.

filter_and_impute(beta_matrix)

- Identifies missing values and applies dynamic thresholding.
- Filters low-quality CpG sites and samples.
- Uses K-nearest neighbor (KNN) imputation for missing values.
- normalize(imputed_filtered_beta_matrix)
- Applies quantile normalization to standardize beta values across samples.

adjust_outliers(normalized_beta_matrix)

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

Usage

1. Set the working directory to the location of the script.
2. Ensure required input files (methylation beta value matrices in CSV/TSV format) are available.
3. Run the script, which will:
    - Load the beta value matrix.
    - Impute missing values and filter low-quality data.
    - Normalize and adjust outliers in the dataset.
    - Save the processed data in .rda format.

Notes

- This script assumes that the input matrix contains beta values for CpG sites.
- The imputation and filtering thresholds are dynamically adjusted based on dataset characteristics.
- Users working with different array platforms may need to adjust processing parameters.