Prediction Script Documentation

The "Prediction" script uses the trained regression models from the "Model" script to predict gene expression levels based on methylation data for a given set of samples. It demonstrates how to use the gene-to-CpG site mapping and the trained Lasso regression models to generate predicted expression levels from methylation data.

Libraries Used

- **glmnet**: Used for applying the Lasso regression models to new methylation data.
- **dplyr**: Provides functions for data manipulation.
- **parallel**: Enhances performance by enabling parallel computing (if applicable in other parts of the script).

Input Files

- Gene-to-CpG site mapping file (**gene_to_cpg_site_mapping.rda**): A mapping between genes and their associated CpG sites.
- Methylation data file for prediction (**Methylation.csv**): Contains methylation levels at specific CpG sites for the samples of interest.
- Regression model file (**regression.output.rda**): Contains the trained Lasso regression models for predicting gene expression levels.

Output File

- Predicted expression data file (**Predicted.rda**): A serialized R data frame saved as an **.rda** file, containing the predicted expression levels for each gene across the samples of interest.

Main Function

- **predict_expression(methylation, gene_to_cpg_site_mapping, regression.output)**: Applies the trained regression models to new methylation data to predict gene expression levels. It iterates over each gene, retrieves the associated CpG sites and the trained model, and uses the model to predict expression levels based on methylation patterns.

Usage

1. Set the working directory to the location of the script.
2. Ensure that the required input files (**gene_to_cpg_site_mapping.rda**, **Methylation.csv**, and **regression.output.rda**) are present in the working directory.
3. Run the script. It will load the necessary libraries, read the input files, and apply the trained models to predict gene expression levels from the new methylation data.

4. The script transposes the matrix of predicted expression levels and converts it into a data frame for easier interpretation and further analysis.
5. The predicted expression data frame is then saved as **Predicted.rda** for future use.

Note

The script assumes that the methylation data for prediction (**Methylation.csv**) is formatted similarly to the methylation data used for training the models, with CpG sites as rows and samples as columns. It is crucial to ensure that the CpG sites and samples are consistent and correctly matched between the training and prediction datasets.