Mapping Script Documentation

The "Mapping" R script is designed to generate a mapping between genes and CpG sites based on genomic coordinates and methylation data. It involves reading gene reference data and a methylation manifest, filtering according to criteria, and then creating a mapping that associates each gene with relevant CpG sites based on their proximity to the gene's location on the genome.

Libraries Used

- **rtracklayer**: For importing gene reference data in GTF format.
- **dplyr**: For data manipulation and filtering.
- **parallel**: For parallel processing to speed up the mapping generation.
- **pracma**: Optional, included for potential mathematical operations not used in the current script.

Input Files

- Gene reference file in GTF format (**hg19.refGene.gtf**): Contains genomic coordinates and other information about genes.
- Methylation manifest file (**HumanMethylation450_15017482_v1-2.csv**): Contains information about CpG sites, including their genomic locations.

Output File

- Gene-to-CpG site mapping file (**gene_to_cpg_site_mapping.rda**): A serialized R data frame saved as an **.rda** file, which maps each gene to its associated CpG sites.

Main Functions

- **get.full.ref.gene.df()**: Imports the GTF file and converts it into a data frame.
- **get.gene.coordinates.df(force.update = FALSE)**: Retrieves or calculates the coordinates for each gene. If **force.update** is **TRUE**, it recalculates the gene coordinates even if a saved file exists.
- **create.gene.to.cg.site.mapping(gene.coordinates.df, upstream.margin.in.bases, downstream.margin.in.bases, force.update = FALSE)**: Creates the mapping between genes and CpG sites. **upstream.margin.in.bases** and **downstream.margin.in.bases** define the genomic range around each gene's coordinates to consider CpG sites as associated with that gene.

Usage

1. Set the working directory to the location of the script.
2. Adjust the input file paths if necessary.

3. Run the script. It will automatically load the required libraries, import the gene reference and methylation manifest files, calculate gene coordinates, and generate the gene-to-CpG site mapping.
4. The resulting mapping will be filtered to exclude genes with fewer than 2 associated CpG sites and those not present in both the expression and methylation datasets.
5. The final mapping is saved as an **.rda** file.

Note

This script assumes that the gene reference data and methylation manifest are based on the HG19 human genome assembly. Users working with data from other genome assemblies or methylation platforms may need to adjust the input files and possibly the script itself to accommodate different formats or content.