Download_TCGA_Data Script

The "Download_TCGA_Data" script is designed to automate the download and preprocessing of RNA and Methylation data from The Cancer Genome Atlas (TCGA) project for multiple cancer types. It ensures that only common samples between RNA and Methylation datasets are retained and renames columns for consistency.

Libraries Used

- **TCGAbiolinks**: For querying, downloading, and preprocessing TCGA data.
- **tidyverse**: For data manipulation and cleaning.
- **maftools**: For analysis of mutation data, not directly used for downloading but useful for TCGA data analysis.
- **SummarizedExperiment**: For handling complex experimental data in an organized form.
- **sesameData** and **sesame**: For handling and analyzing DNA methylation data, particularly from Illumina's platforms.

Main Functions

- **clean_and_rename_columns(df, common_samples)**: Cleans and renames columns of the given data frame based on common samples, ensuring consistency in sample naming across RNA and Methylation datasets.

Workflow

1. **Set the Working Directory**: The script sets the working directory to where the script is located, ensuring all downloaded files are saved in a predictable location.
2. **Download and Preprocess Data for Multiple Cancer Types**: It loops through a predefined list of cancer types, downloading the corresponding RNA and Methylation data for common samples. The data is then cleaned and saved as CSV files.

Usage

1. **Preparation**: Ensure all listed libraries are installed in R.
2. **Execution**: Run the script in an R environment. The script will automatically handle the downloading, preprocessing, and saving of the data.
3. **Output**: For each cancer type in the **cancer_types** vector, two files will be saved in the specified directory:
    - **<Cancer_Type>_RNA.csv**: Contains the RNA data.
    - **<Cancer_Type>_Methylation.csv**: Contains the Methylation data.

Customization

- **Cancer Types**: Modify the **cancer_types** vector to include or exclude specific cancer types based on research needs.
- **Data Categories and Types**: Adjust the **GDCquery** function calls to download different types of data or data from different platforms.
- **Output Location**: Change the path in the **write.csv** function to specify a different saving location for the output files.

Note

- This script makes use of the **matchedMetExp** function from the **TCGAbiolinks** package to identify common samples between RNA and Methylation datasets for each cancer type, ensuring the analysis is performed on matched datasets.
- It is important to verify the availability of data for the specified cancer types and data categories in TCGA before running the script to avoid errors during the download process.