

# Download TCGA Data Script

The "**Download\_TCGA\_Data.R**" script automates 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 mutation data analysis (not directly used in downloading but essential for downstream analysis).
  - **SummarizedExperiment**: For handling complex experimental data in an organized format.
  - **sesameData** and **sesame**: For managing and analyzing DNA methylation data, particularly from Illumina platforms.
- 

## Main Functions

- **clean\_and\_rename\_columns(df, common\_samples)**:  
Cleans and renames the columns of a 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 the script's location, ensuring all downloaded files are saved in a predictable location.

2. **Download and Preprocess Data for Multiple Cancer Types:**

Loops through a predefined list of cancer types to:

- Download corresponding RNA and Methylation data for common samples.
  - Clean and preprocess the data.
  - Save the results as CSV files.
- 

# Usage

1. **Preparation:**

Ensure that all listed libraries are installed in R.

2. **Execution:**

Run the script in an R environment. The script will handle the downloading, preprocessing, and saving of the data.

3. **Output:**

For each cancer type in the `cancer_types` vector, the script generates two output files:

- `<Cancer_Type>_RNA.csv`: Contains RNA data.
  - `<Cancer_Type>_Methylation.csv`: Contains Methylation data.
- 

# Customization

- **Cancer Types:**

Modify the `cancer_types` vector to include or exclude specific cancer types based on your research needs.

- **Data Categories and Types:**

Adjust the `GDCquery` function calls to download different data types or data from various platforms.

- **Output Location:**

Change the path in the `write.csv` function to specify a different save location for the output files.

---

## Notes

- This script uses the `matchedMetExp` function from the **TCGAbiolinks** package to identify common samples between RNA and Methylation datasets for each cancer type. This ensures the analysis is performed on matched datasets.
- 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.