# README: Integrated Machine-learning and Multi-omics Profiling for Luminal Breast Cancer Biomarker Identification Pipeline

- **Execution Order & Description**

1. **Download_TCGA_Data Script**

   - Purpose: Automates the download of RNA and methylation data from TCGA for multiple cancer types.
   - Main Functions: Queries TCGA, downloads data, retains common samples, and renames columns for consistency.
   - Output: <Cancer_Type>_RNA.csv, <Cancer_Type>_Methylation.csv

2. **Preprocessing RNA Data Script**

   - Purpose: Processes RNA-seq count matrices, including normalization, gene mapping, and filtering.
   - Main Functions: Loads data, maps Ensembl IDs, filters low-expression genes, normalizes data, and adjusts outliers.
   - Output: Processed RNA count matrix (.rda format)

3. **Preprocessing Methylation Data Script**

   - Purpose: Processes DNA methylation beta value matrices, including imputation, normalization, and outlier handling.
   - Main Functions: Loads beta values, imputes missing data, applies quantile normalization, and adjusts outliers.
   - Output: Processed Methylation Matrix (.rda format)

4. **BRCA Samples Subtyping Script**

- Purpose: Classifies BRCA samples into molecular subtypes (LumA, LumB, Basal, Her2, Normal) using the PAM50 classifier.
- Main Functions: Maps gene names to probe IDs, prepares expression data, and performs PAM50 subtyping.
- Output: Subtype sample lists (.rda format)

5. **Mapping Script**

- Purpose: Generates a mapping between genes and CpG sites based on genomic coordinates.
- Main Functions: Imports reference data, processes genomic locations, and creates gene-to-CpG site mapping.
- Output: Gene-to-CpG site mapping (.rda format)

6. **Model Script**

- Purpose: Trains predictive models using Lasso regression to predict gene expression from methylation data.
- Main Functions: Loads data, applies Lasso regression, stores model outputs.
- Output: Regression model file (.rda format)

7. **Prediction Script**

- Purpose: Uses trained Lasso models to predict gene expression levels from new methylation data.
- Main Functions: Loads trained models, applies them to new methylation data and generates predicted expression values.
- Output: Predicted expression data (.rda format)

8. **DEGs Loop (t-test & Wilcoxon) Script**

- Purpose: Identifies differentially expressed genes (DEGs) between normal and cancer samples using statistical tests.
- Main Functions: Performs normality testing, t-tests or Wilcoxon tests, multiple testing correction, generates volcano plots.
- Output: DEG results (.csv), significant genes (.csv), volcano plots (.png)

## • Usage Instructions

- Ensure all required R packages are installed before running the scripts.
- Follow the execution order to maintain data integrity and consistency
- Modify cancer types, parameters, and paths as needed.
- Use the generated .rda and .csv files for downstream analysis and visualization.