BRCA Samples Subtyping Script Documentation

The BRCA Samples Subtyping script is designed to classify breast cancer (BRCA) samples into molecular subtypes using the PAM50 classifier. The script processes gene expression data, maps gene names to probe IDs and Entrez Gene IDs, and performs subtyping using the genefu package. The resulting subtypes (LumA, LumB, Basal, Her2, Normal) are saved as .rda files for further analysis.

Libraries Used

- **limma:** For linear models and differential expression analysis.
- **affy:** For handling Affymetrix microarray data.
- **hgu133plus2.db:** Annotation database for Affymetrix Human Genome U133 Plus 2.0 Array.
- **AnnotationDbi:** Provides interface to annotation databases.
- **stringr:** For string manipulation.
- **genefu:** Contains the PAM50 classifier for breast cancer subtyping.
- **xtable:** For generating LaTeX tables.
- **rmeta:** For meta-analysis.
- **Biobase:** Base functions for Bioconductor.
- **caret:** For classification and regression training.
- **biomaRt:** For accessing BioMart databases to map gene IDs.
- **dplyr:** For data manipulation.
- **tibble:** For modern data frames.

Input Files

1. BRCA Gene Expression Data:

File format: .csv.

- Contains gene expression data for BRCA samples.
- The first column should contain gene names and subsequent columns should contain expression values for each sample.

Example: BRCA_RNA.csv.

Output Files

1. Subtype Sample Lists:

File format: .rda.

- Contains lists of samples classified into each PAM50 subtype:
- LumA: LumA_Samples.rda.
- LumB: LumB_Samples.rda.
- Basal: Basal_Samples.rda.
- Her2: Her2_Samples.rda.
- Normal: Normal_Samples.rda.

Main Steps

1. Load and Prepare Data

- The script reads the BRCA gene expression data from a .csv file.
- The first column is renamed to Genes to ensure consistency.

2. Map Gene Names to Probe IDs

- The script uses biomaRt to map gene names to Affymetrix probe IDs.
- Duplicate probe IDs are removed, and the Genes column is replaced with matched probe IDs.

3. Map Probe IDs to Entrez Gene IDs

- The script maps probe IDs to Entrez Gene IDs using biomaRt.
- Rows with missing Entrez Gene IDs are removed.

4. Prepare Data for Subtyping

- The gene expression data is transposed to fit the input format required by the PAM50 classifier.

5. Perform PAM50 Subtyping

- The script uses the **molecular.subtyping** function from the genefu package to classify samples into PAM50 subtypes.
- Subtypes include LumA, LumB, Basal, Her2, and Normal.

6. Save Subtype Results

- The script saves lists of samples for each subtype as .rda files.

Usage

1. Set Up Input File:

- Ensure the BRCA gene expression data file (**BRCA_RNA.csv**) is in the working directory.

2. Run the Script:

- Execute the script to:
  - Load and preprocess the gene expression data.
  - Map gene names to probe IDs and Entrez Gene IDs.
  - Perform PAM50 subtyping.
  - Save the results for each subtype.

3. Review Output:

- Check the generated .rda files for lists of samples classified into each subtype.