Summary

The *triple classifier* allows for classification of patients with clear cell Renal Carcinoma as high or low risk for disease progression. The classification is based on methylation data (beta-values) from tumor samples together with clinical information including age, gender, tumor diameter, Fuhrman grade, TNM-stage, hemoglobin level, albumin level, calcium concentration, gamma-glutamyltransferase, thrombocyte particle count, alkaline phosphatase and creatinine levels. The model is built using the first five principal components of 12 clinical variables, 64 methylation sites and 40 methylation biomarkers obtained using a novel approach called Directed Cluster Analysis (DCA).

Installation

No installations is required, run the code in your R-console.

Data

To run the code, you need methylation data (beta-values) from Infinium Human Methylation 450k BeadChip or Infinium MethylationEPIC arrays and clinical variables for the patients to classify. You also need the provided files `CpG_Panel`, `DCA_Matr`, `MeanSD`, `rotation_matr` and `model_coeff`. `CpG_Panel` contains the names of CpG sites that have been identified in the literature as relevant in ccRRCC prognosis. The information of which CpG sites that are used for construction of DCA variables are stored in `DCA_Matr`. The `MeanSD` file contains values for standardization of the data, and `rotation_matr` is used for construction of the first five principal components. `model_coeff` contains the weights for the principal components in the logistic regression model.

Usage

The main function is `tripleClassif`, which calls functions `ConstructDCA`, `CombD`, `NewPCA` and `Classify`.

The `ConstructDCA` function makes consensus variables by calculating the mean beta value of the defined CpG sites for each DCA variable.

`CombD` combines clinical input data with the methylation panel that is created in `tripleClassif` and the DCA consensus variables created by `ConstructDCA`.

The function `NewPCA` normalizes the combined clinical and methylation data and creates the first five principal components based on the provided rotation matrix.

The function `Classify` provides the posterior probabilities for each patient calculated by the provided logistic regression model.

The function `tripleClassif` requires the following 7 arguments:

- ***Clin*** - A data.frame with 12 columns, where column names are: "Age", "Gender", "Tumor_diam", "Grade", "TNM", "Hb", "Albumin", "Calcium", "ALP", "GGT", "TPC", "Creatinine". Row names are sample id (same as column names in ***MethylationD***).
- ***MethylationD*** –A data.frame with normalized beta values from Illumina EPIC. methylation arrays, where column names are sample id and rownames are CpG id.
- ***model*** - A logistic regression model.
- ***PCAMatr*** – A matrix containing the loadings from the principal component analysis.
- ***DCAMatr*** – A matrix that indicates which CpG sites that are included in the 40 DCA consensus variables.
- ***PICpG*** – A vector containing names of CpG sites that previously have been identified as relevent in ccRCC.
- ***MeanSD*** – A data.frame with two columns: Mean_v and SD_v, containing mean values and standard deviations that are used for standardizing variables before principal component analysis.

NAs are not allowed in Clin or in MethylationD for CpGs included in PICpG. Missing values can be imputed using mean values in MeanSD. The output of `tripleClassif` is a numeric vector with classification label (1- high risk for progression, 0- low risk for progression).