# Documentation for Project 'IamComparison':

## main.R

Call to all R files necessary to run the whole analysis (except for MALA)

- Set path to directory where the project (raw, bin, src, results, …) is located on the computer
- Specify name of biological input data object
- Specify name for current results run
- Set relative path to sub directories
- Subsequently run analysis on biologic and synthetic data

## synthetParameter.R

File holding all parameter settings for the data simulation and the application of methods to the synthetic datasets. Change the parameters here to create the data and/or run the analysis with different settings.

## biologParameter.R

File holding all parameter settings for the application of methods to the biological datasets. Change the parameters here to run the analysis with different settings.

## syntheticDataES.R

Generate synthetic datasets with different simulation parameter sets.

- Simulate gene expression and DNA methylation dataset
- Create synthetic datasets for 9 sets (combinations) of simulation parameters
- For each set of simulation parameters create 100 datasets for statistics purpose

## preprocTCGAData.R

General preprocessing of biological input dataset.

- select data types to integrate as specified in the parameter file from raw data
- split dataset into subsets of tumor and normal samples

## biologicalData.R

Subset the biological dataset to obtain a data matrix suitable for the application of all methods, i. e. multiple omics levels measured on the same sample under different conditions.

- reduce data subsets to participants (patients) common to all data and tissue types
- remove features containing more than 10% NAs (only occurring in Methylation datasets)
- remove features containing more than 10% zeroes (applies to RNASeq data only)
- impute remaining NAs/zeroes with half of the lowest value in each feature
- do data transformation: log2 transform RNA-seq, 1-x transform methylation data
- plot histograms before and after transformation
- calculate fold change/difference in mean between conditions and set threshold to determine the approximate number of genes of interest → percentage is used as parameter to predefine the number of features to be selected by each method

- prepare data subsets for cross validation

### sCCA.R

Apply sCCA to synthetic/biological datasets and save results

- parameter tuning to determine the appropriate constants for the restriction of the L1 norm in order to obtain a predefined number of features selected by sCCA
- apply sCCA implemented in R package PMA to synthetic data runs
- apply sCCA implemented in R package PMA to biological datasets and do cross validation (apply sCCA to training and test datasets)
- save flat list of features selected by sCCA
- For synthetic data runs draw ROC curves and calculate AUCs and other performance measures
- For biological data evaluate performance (calculate TPR, FPR and accuracy) on training and test datasets

### preprocForNMF.R

Prepare synthetic/biological datasets specifically for the application of NMF

- standardization of columns (features) in the datasets
- scaling of datasets to achieve equal Frobenius norms
- make dataset fit the non-negativity constraint

### NMF.R

Apply NMF to synthetic/biological datasets and save raw results

- apply NMF to synthetic data runs
- apply NMF to biological datasets and do cross validation (apply NMF to training and test datasets)

### postprocOfNMF.R

Post process raw results of NMF to obtain a flat list of the predefined number of features selected by NMF.

- Z-transform weights corresponding to features and extract predefined number of top-weighted features
- For synthetic data runs draw ROC curves and calculate AUCs and other performance measures
- For biological data evaluate performance (calculate TPR, FPR and accuracy) on training and test datasets

### preprocForMALA.R

Prepare synthetic/biological datasets specifically for the application of MALA

- standardization of rows (features) in the datasets

### MALA.R

Apply MALA to synthetic datasets. Run MALA parallel on multiple synthetic data runs on an Ubuntu/debian Linux machine with large RAM resources and save raw results

- calls runMALA.R which actually runs MALA
- add up the features resulting from the application of MALA to a datasets until a predefined number of features or a maximum number of splits is reached

MALA_linux.R

Apply MALA to biological datasets. This file is not embedded in the project workflow → MALA and the datasets to be analyzed must be copied to an Ubuntu/debian Linux machine, MALA is executed there, and the results must be transferred to the project folder manually.

- MALA is applied to each preprocessed dataset located in the same folder as MALA and MALA_linux.R
- results for each dataset are saved to a separate folder named after the input dataset

postprocOfMALA.R

Post process raw results of MALA application to synthetic/biological datasets to obtain a flat list of the features selected by MALA.

- Extract features from csv files containing the logic formulas for classification of samples resulting from the application of MALA to synthetic/biological datasets
- For synthetic data runs draw ROC curves and calculate AUCs and other performance measures

syntheticComparison.R

Assess performance of methods to detect DE features and perturbed pathways in synthetic datasets. Load AUCs or other performance measures for synthetic datasets created with different simulation parameters. Visualize and compare performance of methods under different parameter settings as box plots of AUCs and other performance measure of multiple statistics runs.

- Load arrays containing AUCs or other performance measures and create box plots for different simulation parameter settings and multiple statistics runs
- Combine boxplots in one figure for direct comparison of method performance

methodComparison.R

Assess congruency of feature lists, GO terms and pathways resulting from the application of methods to the biological datasets. Assess overlap of results with curated cancer gene signatures.

- Plot Venn diagrams on feature/gene level and on the level of over-represented GO terms and perturbed Reatome pathways.
- Create tables (pdf, txt, tex) listing features/genes, GO terms and pathways resulting from individual methods for the supplement.
- Assess significance of overlap with curated cancer gene signatures and plot Venn diagrams
- Calculate average classification accuracy on training and test sets

**ADDITIONAL R files (called within files above):**

assessPerformance.R

Assess classification power of methods on biological datasets.

crossValidation.R

Assess classification power of methods on biological datasets.

drawROC.R

Calculation of TPR and FPR, construction of ROC curves, calculate AUC and other performance measures for results on synthetic datasets.

extractMALAFeatures.R

Extract feature names from logic formula files resulting from MALA.

helpParameters.R

Extract the parameters used to create synthetic datasets from their names.

multiNMF_mm.R

Re-implementation and extension to multiple datasets of TriNMF_mm.m function implementation by Shihua Zhang.

multiNMF_residue.R

Re-implementation and extension to multiple datasets of TriNMF_residue.m function implementation by Shihua Zhang.

multiNMF_residue_comodule.R

Re-implementation and extension to multiple datasets of TriNMF_residue_comodule.m function implementation by Shihua Zhang.

setDirPath.R

Set all paths to the project sub directories.

setSubDirPath.R

Set all paths to the project result run sub directories.