Reproducibility in the Application to the Kabuki Syndrome Data

This file contains instructions for reproducing the results and figures in the application to the Kabuki syndrome data. The codes are in "code" folder. Please set the working directory to the source file location ("code" folder).

The following code files can be directly run as their input data are in the "input\_data" and "result\_data" folders: Step1\_Preprocessing\_data.R, Step2\_BIC\_analysis.R, Step3\_BUSbeta.R, Step4\_BUS.R, Step5\_ComBat.R, Step6\_BEclear.R, Step7\_Figure2.R, Step8\_Kmeans.R, Step9\_GMM.R and Step10\_MetaSparseKmeans.R.

**Data**

The Kabuki syndrome data consist of two data sources available in GEO with accession numbers GSE218186 and GSE97362, serving as 2 batches. The data contain 206 samples in total, which are assigned to 2 subgroups: 26 Kabuki syndrome patients and 180 healthy controls. The data file “GSE218186\_series\_matrix.txt” is downloaded from the link <https://ftp.ncbi.nlm.nih.gov/geo/series/GSE218nnn/GSE218186/matrix/GSE218186_series_matrix.txt.gz>. The data file “GSE218186\_RAW.txt” is downloaded from the link <https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE218186&format=file>. And the data file “GSE97362\_series\_matrix.txt” is downloaded from the link <https://ftp.ncbi.nlm.nih.gov/geo/series/GSE97nnn/GSE97362/matrix/GSE97362_series_matrix.txt.gz>. Please first download all the data files to the "input\_data" folder.

In the data preprocessing procedure, we collect the subgroup information for each sample and select 148 informative CpG sites from the file “informative\_cpg\_Kabuki.txt” in the “input\_data” folder, which capture distinct DNA methylation patterns across subgroups, according to Aref-Eshghi et al. (2020).

The preprocessed data includes:

* "Y\_preprocessed.RData": The preprocessed data list.
* "Z\_real.RData": The true subgroups list.

**Code**

**Step 1: " Step1\_Preprocessing\_data.R "**

Read data from the data files “GSE218186\_series\_matrix.txt,” “GSE218186\_RAW.txt” and “GSE97362\_series\_matrix.txt” in "input\_data" folder using the R package vroom. Then preprocess the raw Kabuki syndrome data matrix. The outputs of this step are "Z\_real.RData" and "Y\_preprocessed.RData."

**Step 2: "Step2\_BIC\_analysis.R"**

Calculated the BIC values for subgroups value K ranging from one to ten, and draw Figure 2(b) in Section 6 to determine the number of subgroups for the following BUSbeta analysis. The output of this step is "Figure2(b).png" in "figures" folder.

**Step 3: "Step3\_BUSbeta.R"**

Implement BUSbeta on the data from Step 1, and conduct the MCMC posterior inference. Then compute the ARI values in each single batch and the overall ARI values for BUSbeta. The outputs of this step are "Y\_BUSbeta.RData" which saves the corrected data, "Z\_BUSbeta.RData" which saves the clustering labels and "ARI\_BUSbeta.csv" in "result\_data" folder.

In the "ARI\_BUSbeta.csv":

* "Batch1," "Batch2": samples in each batch.
* "Overall": samples combined from all batches together.
* "BUSbeta ARI": ARI values by BUSbeta.

**Step 4: "Step4\_BUS.R"**

Implement BUS model using the R package BUScorrect. Then compute the ARI values in each single batch and the overall ARI values for BUS. The outputs of this step are "Y\_BUS.RData" which saves the corrected data, "Z\_BUS.RData" which saves the clustering labels and "ARI\_BUS.csv" in "result\_data" folder.

In the "ARI\_BUS.csv":

* "Batch1," "Batch2": samples in each batch.
* "Overall": samples combined from all batches together.
* "BUS ARI": ARI values by BUS..

**Step 5: "Step5\_ComBat.R"**

Implement ComBat method using the R package sva. The output of this step is "Y\_ComBat.RData" which saves the corrected data in "result\_data" folder.

**Step 6: "Step6\_BEclear.R"**

Implement BEclear method using the R package BEclear. The output of this step is "Y\_BEclear.RData" which saves the corrected data in "result\_data" folder.

**Step 7: "Step7\_Figure2.R"**

Draw Figures 2(a) and 2(c)-(f) in Section 6 using the data from Step 1 and the corrected data of all methods.

**Step 8: "Step8\_Kmeans.R"**

Implement Kmeans on the data from Step 1 and compute the ARI values in each single batch. The output of this step is "ARI\_Kmeans.csv" in "result\_data" folder.

In the "ARI\_Kmeans.csv":

* "Batch1," "Batch2": samples in each batch.
* "Overall": samples combined from all batches together.
* "Kmeans ARI": ARI values by Kmeans.

**Step 9: "Step9\_GMM.R"**

Implement GMM on the data from Step 1 using the R package mclust, and compute the ARI values in each single batch. The output of this step is "ARI\_GMM.csv" in "result\_data" folder.

In the "ARI\_GMM.csv":

* "Batch1," "Batch2": samples in each batch.
* "Overall": samples combined from all batches together.
* "GMM ARI": ARI values by GMM.

**Step 10: "Step10\_MetaSparseKmeans.R"**

Implement MetaSparseKmeans on the data from Step 1 using the R package MetaSparseKmeans, and compute the ARI values in each single batch and the overall ARI values. The output of this step is "ARI\_MetaSparseKmeans.csv" in "result\_data" folder.

In the "ARI\_ MetaSparseKmeans.csv":

* "Batch1," "Batch2": samples in each batch.
* "Overall": samples combined from all batches together.
* "MetaSparseKmeans ARI": ARI values by MetaSparseKmeans.