**User Guide**

2019-02-18

**Contact:**

Baoshan Ma ([mabaoshan@dlmu.edu.cn](mailto:mabaoshan@dlmu.edu.cn))

Liming Liang ([lliang@hsph.harvard.edu](mailto:lliang@hsph.harvard.edu))

**Citation:**

[1] Baoshan Ma, Catherine Allard, Luigi Bouchard, Patrice Perron, Murray Mittleman, Marie-France  Hivert, Liming Liang. Locus-specific Methylation Prediction in Cord Blood and Placenta (Epigenetics).

**1. Introduction**

The R functions provided in this package can be used to build prediction model for methylation value within a pair of surrogate and target tissues. The prediction model can then be applied to a new dataset where only surrogate tissue methylation value is available and output predicted methylation value of the target tissue. If you find the program useful, please cite the above reference.

**2. Software requirement**

R including the e1071(svm) package

**3. Input parameters**

The R function “mctp” is included in the R script “methylation\_cross\_tissue\_prediction.R” and is acronym of this file name. Five input parameters are required, they are methylation data file names for (1) training dataset of surrogate tissue, (2) training dataset of target tissue, (3) testing dataset of surrogate tissue (i.e. the samples being predicted in surrogate tissue), (4) predicted target tissue data by single-CpG-based SVM model and (5) predicted target tissue data by multiple-correlating-CpGs-based SVM model. In each methylation data file, rows correspond to samples and columns correspond to methylation CpGs, and the row and column names are required. In addition, corresponding row represents the same sample in the surrogate training dataset and target training dataset, corresponding column represents the same CpG in the surrogate training dataset, target training dataset and surrogate testing dataset.

The detailed parameters for the mctp() function are:

1. x is the file name of surrogate training data
2. y is the file name of target training data
3. z is the file name of surrogate testing data
4. u is the file name of predicted target tissue data by single-CpG-based SVM model
5. v is the file name of predicted target tissue data by multiple-correlating-CpG-based SVM model

**4. How to run the mctp() function and obtain predicted methylation value in target tissue**

(1) Organize three data files corresponding to x, y and z as described above: surrogate training data, target training data and surrogate testing data. You can change the surrogate training data and target training data in our example to your training data, the surrogate testing data to your surrogate tissue of the samples being predicted.

(2) Define output file names for predicted target tissue data, the first one is for single-CpG-based SVM and the second one is for multiple-correlating-CpG-based SVM. For example, “Predicted\_target\_single\_svm.txt” is the predicted target tissue methylation value using single-CpG-based SVM model, “Predicted\_target\_correlating\_svm.txt” is the predicted target tissue methylation value using multiple-correlating-CpG-based SVM. Rows and columns in the output files are corresponding to the rows and columns in the surrogate testing data file “Surrogate\_testing\_data.txt” respectively.

(3) Call mctp() function in your R program, for example,

library(e1071) #load SVM R package

source("methylation\_cross\_tissue\_prediction.R") #load R program

mctp("Surrogate\_training\_data.txt","Target\_training\_data.txt","Surrogate\_testing\_data.txt","Predicted\_target\_single\_svm.txt","Predicted\_target\_correlating\_svm.txt")

The first three file names are corresponding to the surrogate training data, target training data and surrogate testing data. After running this function, we will get two predicted target tissue data files: “Predicted\_target\_single\_svm.txt” and “Predicted\_target\_correlating\_svm.txt”.

**5. How to use our SVM models**

For the 1,012 well-predicted CpG sites, we built the single-CpG-based SVM models using our 169 cord blood and placenta samples. The user could apply the methylation values in cord blood to calculate the methylation values in placenta based on our SVM models. The steps are listed as:

1. Library(e1071)
2. Load the R data file including the established SVM models
3. Obtain the index of the SVM model corresponding to the CpG of being predicted. The 1,012 SVM models are corresponding to the 1,012 CpG sites in the data sheet 1 of “SupplementaryTable3-well\_predicted\_CpGs\_20190218.csv”, and the index of the CpG name is the index of the SVM model.
4. Calculate the methylation values in placenta using the methylation values in cord blood.

For example, the user would like to calculate the methylation values at the CpG site “cg02233190”, the index of this CpG in the data sheet 1 is 2 (i.e., number of row-1), thus the user should choose the second SVM model,

predicted methylation levels at cg02233190 in placenta= predict(cordblood\_placenta\_SVM\_models [[2]], the methylation levels at cg02233190 in cord blood)