

User Guide

2013-8-20

Contact:

Baoshan Ma (mabaoshan@dlmu.edu.cn)

Liming Liang (lliang@hsph.harvard.edu)

Website:

<http://www.hsph.harvard.edu/liming-liang/cross-tissue-methylation/>

Citation:

Baoshan Ma, Elissa Wilker, Saffron A.G. Willis-Owen, Hyang-Min Byun, Kenny C.C. Wong, Valeria Motta, Andrea A. Baccarelli, Joel Schwartz, William O.C.M. Cookson, Kamal Khabbaz, Murray A. Mittleman, Miriam F. Moffatt, Liming Liang. (2013) Predicting DNA methylation level across human tissues (submitted)

1. Introduction

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

2. Software requirement

R including the e1071(svm) package

3. Input parameters

The R function “mctp” is included in the file “methylation_cross_tissue_prediction.R” and stands for Methylation Cross Tissue Prediction. Five input parameters are required, corresponding to file names for methylation data from surrogate and target tissue of the training dataset, surrogate tissue of the samples being predicted, output file names for predicted target tissue methylation by SVM and LM. In each methylation data file, rows correspond to sample and columns for methylation probes.

The detailed parameters for the mctp() function are:

- (1) x is file name of surrogate tissue of training data
- (2) y is file name of target tissue of training data
- (3) z is file name of surrogate tissue of testing data
- (4) u is output file name of predicted target tissue data by SVM
- (5) v is output file name of predicted target tissue data by LM

4. How to run the mctp() function and get output predicted value

- (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 file names of output predicted target tissue data, one is based on SVM model and the other based on linear model, for example, "Predicted_target_svm.txt" is the predicted target tissue methylation value using SVM, "Predicted_target_lm.txt" is the predicted target tissue methylation value using linear model. Rows and columns of the output files are corresponding to the rows and columns of the surrogate tissue of testing data file "Surrogate_testing_data.txt".
- (3) Call mctp() function in your R program, for example,

```
source("methylation_cross_tissue_prediction.R") #load the R program

mctp("Surrogate_training_data.txt","Target_training_data.txt","Surrogate_testing_data.txt","Predicted_target_svm.txt","Predicted_target_lm.txt")
```

The first three file names are corresponding to surrogate training data, target training data and surrogate testing data. After running this function, we will get two output predicted target tissue data files: "Predicted_target_svm.txt" and "Predicted_target_lm.txt".