Integrative analysis of TCGA data with expanded DNA methylation data reveals common patterns in cancers

Shicai Fan[1,2,4,5], Nan Li[2], Ying Zhao[2], Rizi Ai[2], Mengchi Wang[2], Chunguo Wu[5], Wei Wang[2,3*]

1. School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China;
2. Department of Chemistry and Biochemistry, 3. Department of Cellular and Molecular Medicine, University of California, San Diego, CA 92093-0359.
4. Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China;
5. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China;
*Correspondence: wei-wang@ucsd.edu

Before starting

The EAGLING model should be implemented on 64-bit version of Linux platform installed with R.

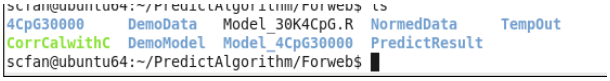
Install R

You will need to install the correct version of R for your operating system. In order to do this visit the R mirror that is closest to your location from: http://www.r-project.org/

Required packages: include but are not limited to preprocessCore, e1071.

QuickStart

You will need to download the trained model, source code, required data, extract them and MUST locate them in the same directory. It is like the situation below:



Analyzing your data

Your data should be in the correct format:

1) The given 450K data should be provided as one file for each chromosome (the file name need to be named as like *chr1.txt, see examples in the demo data),

2) The given 450K data file has two columns separated with '\t'. The first column would be the location of C in the version of hg19 (needs to be sorted in increasing order), and the second column is the methylation value ranging from 0 to 1.

Then in the R platform, you can just simply type in the command line with the commands similar to the example below.

```
source('Model _30K4CpG.R')
InputDataFolder <- 'DemoData';
OutputDataFolder <- 'PredictResult';
ModelFolder <- 'DemoModel';
ChromArray <- c(20,21);
Model _30K4CpG (InputDataFolder,OutputDataFolder,DemoFolder,ChromArray)
```

The definition and possible values about parameters of the function could be found in the R script.

Output

The predicted methylation values for all the expanded CpG sites of each chromosome would be output into the specified output folder. In each output file, there are two columns, the first column is the location, and the second column is the predicted methylation value.