Computationally expanding Infinium HumanMethylation450 BeadChip array data to reveal distinct DNA methylation patterns of rheumatoid arthritis

Shicai Fan[1,2] ,Chengzhe Li[1],Rizi Ai[2], Gary S. Firestein[3,*], Wei Wang[2,*]

[1]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China;

[2]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA, USA;

[3]Department of Medicine, University of California San Diego School of Medicine, La Jolla, CA, USA;

*To whom correspondence should be addressed.(wei-wang@ucsd.edu)

Before starting

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 and randomForest

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:

```
[scfan@skat DownloadData]$ ls
DemoData          PredictWith450K.R  SeqFeatureWGBSFlank5K5CpG
ModelwithoutSeq   RatioFile          Top50SeqFeatureIndex
ModelwithSeq      RFSeqFeatSelect.R  WGBSWith450KFlank5K5CpG
Normed450K14Tis   SeqFeature450K
[scfan@skat DownloadData]$
```

We provided two categories of models: models with sequence features (in the folder of ModelwithSeq) or models without sequence features (in the folder of ModelwithoutSeq). models with sequence features are the models we used in our project, however, the process of selection of 50 most selected sequence features and training with sequence features are quite time consuming and memory consuming. Therefore, we also provided the models without sequence features (with only x_1 and x_2 in the flowchart), these models are much faster and only reduce a little performance (less than 4%).

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('PredictWith450K.R')
InputDataFolder <- 'DemoData';
OutputDataFolder <- 'PredictResult';
ChromArray <- c(20,21);
SeqModel <- 0;
PredictWith450K(InputDataFolder, OutputDataFolder, ChromArray, SeqModel)
```

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.