**Procedure for data analysis:**

1. Please use Bismark software for bisulfite sequence pre-processing and mapping.
2. Merge CpG site information (merge\_CpG\_site.pl)

### merge\_CpG\_site.pl

Description

This script is used to parse Bismark result to get the methylation

pattern for each CpG dyads. Both the Bismark result and CpG coordinate

need to be provided as input. The information for the two paired Cs in

the same CpG dyad are merged as one based on the CpG coordnate file.

Usage

perl merge\_CpG\_site.pl <cpg coordnate file> <bismark output> > output

Example

perl merge\_CpG\_site.pl hg19.CPG CpG\_context.hg19.fetal.sam.txt >

CPG\_context.hg19.fetal.sam.txt.site

Version

Ver 1.1, Ming-an Sun, May 18, 2014 (for multiple chr)

Ver 1.0, Ming-an Sun, Sep 04, 2013 (for individual chr)

1. Get 4CG seed information (extract\_seed\_info.pl) and convert the format (convert\_seed\_format.pl)

### extract\_seed\_info.pl

Description

This script is used to extract 4CG seed information from merged CpG

results. The input should be sorted by chrs and coordinates already (the

output of merge\_CpG.pl can be used directly).

Usage

perl extract\_seed\_info.pl <detailed siteInfo file> > output

Example

perl extract\_seed\_info.pl hg19.fetal.site

Version

Ver 1.0, Ming-an Sun, May 19, 2014

### convert\_seed\_format.pl

Description

This script is used to convert segmetn format. The input format is with

rich information which also include patterns with missing info (such as

??11). The output is only for patterns with full information, and also

shows the calculated ML and ME for each segment.

Usage

perl convert\_seed\_format.pl <detailed seed file>

Version

v1.0, Ming-an Sun, Aug 12, 2014

1. CSM prediction (csm\_detector.m)

The matlab script csm\_detector.m is designed for CSM prediction. Make sure to change the codepath (line 7) before used. The input file name also need to be changed.

## Input:

The input is file with methylation pattern of format below. It is got from 4-CG sliding window (4-CG seed). The format is “methylationPattern:count;”. For example, 0000:3 means 0000 occurs 3 times. Before prediction, make sure to make a input of this format.

The other thing is that CSM can only be predicted for 4-CG windows with both 0000 and 1111. Discard lines without both 0000 and 1111 before prediction.

0000:3;0001:10;0011:2;0101:2;0111:1;1000:1;1111:1;

0000:1;1011:1;1111:11;

0000:4;0111:1;1111:13;

0000:4;1110:2;1111:12;

## Output:

It produces two output files. One is the distance between two clusters; the other is p-values. Currently, we predict a 4-CG seed as CSM when the dist>=0.8 and pval<=0.01.

1. dist.d5t5.txt
2. pval.d5t5.txt