Methy-Pipe Manual (v2.02)

The structure of directory hierarchy for Methy-Pipe and analysis output.

(1) Methy-Pipe pipeline structure.

methy-pipe2

```
|-methy-pipe2.pl          # main script used to run Methy-Pipe
|-cpp_prog/               # binary programs compiled from C++
|-perl_prog/              # perl programs
|-R_prog/                 # R scripts (requires ggplot2 and gridExtra packages)
|-bed_files/              # bed files such as TSS regions frequently used to calculate
                          # the regional methylation density
|-utils/                  # extra utilities such as DMRs mining etc.
  |-2bwt-builder/         # scripts can be used to build the BWT index for a reference genome
  |-split_meth_call/      # scripts can be used to split the methylation call files based on
                          # each chromosome.
  |-DMR/                  # programs can be used to identify the differentially methylated regions
                          # based on above split methylation call files
  |-bed_files/            # containing regions of interest in bed format such as TSS regions,
                          # gene regions, etc
  |-database/             # containing reference genome
```

(2) Methy-Pipe output structure.

Methy-Pipe_output

```
|-outprefix_alignment/    # BS-Seq alignment results
|-outprefix_meth_call/    # methylation calling results
|-outprefix_meth_density/ # methylation density profiling by using a fixed window size
|-outprefix_summary/      # summary for the Methy-Pipe
|-outprefix_logs/         # intermediate results that can be deleted by users
|-DMRs/                   # analyzing the differentially methylated regions
```

How to use Meth-Pipe

I. If you want to quickly start Methy-Pipe, please use the following shell scripts in dataset folder. It will show some key instructions of how to run it. For the detailed implementation, please refer to the following section and manuscript.

```
#IMPORTANT:
#Please install the ggplot2 (http://ggplot2.org/) and gridExtra
(http://cran.r-project.org/web/packages/gridExtra/) libraries for R before
starting Methy-Pipe.
#uncompress the files:
tar xjf methy-pipe2.full.tar.bz2
tar xjf dataset_light.tar.bz2
cd dataset_light
#create a makefile for BS-Seq alignment and methylation calling:
./wk.sh
#change to the output result folder:
cd Methy-Pipe_output
#Use makefile to run the Methy-Pipe:
make
#if users want to use 2 parallel computing nodes to run
#Methy-Pipe based on SGE platform, run the following command:
#./qsub.sh 2 makefile
#Split the methylation call by each chromosome:
./demo_split_call.sh
#change to DMRs identification folder
mkdir DMRs
cd DMRs
./demo_DMRs.sh
```

II. Two test datasets are accompanied with the released Methy-Pipe software:

(1) dataset_light.tar.bz2:

This folder contains the raw data in fastq format, which can be used to test the Methy-Pipe. Since the depth of this dataset is very low (less than 5 fold on average), the DMR detection probably is not very accurate to offer biological significance. Nevertheless, it is a good example to illustrate how to use Methy-Pipe for the following purposes:

- trimming the raw reads with low-quality bases or sequencing adaptors.
- aligning the BS-seq reads.
- calculating the mappability and sequencing coverage.
- summarizing the results in a "summary.html" file.

(2) dataset_full.tar.bz2:

This dataset contains raw data in fastq format, which can be used to test Methy-Pipe as said in (1). In additional, users can perform the DMR mining.

III. The configuration file for the Methy-Pipe is illustrated in ***CONF***. It is easy to modify this standard ***CONF*** to analyze new dataset. The following is one example of CONF (Please modify the path in blue accordingly).

```
#FOMART: KEY<TAB>VALUE
# path to the statistics program R
R       /path-binary-R/R
#reference genome index for the BSAaligner
BS_INDEX       /path-to-BSAlignerIndex/hg19
#mismatch allowed for each end
MISMATCH 2
#minimal insert size allowed for paired-end reads
MIN_INS 0
#maximal insert size allowed for paired-end reads
MAX_INS 600
#each chromosome length
LIST_CHR_LEN   /path-to-BSAlignerIndex/hg19.size
#Watson strand reference (fasta)
GENOME_W_FA    /path-to-BSAlignerIndex/hg19.W.ori.fa
#Crick strand reference (fasta)
GENOME_C_FA    /path-to-BSAlignerIndex/hg19.C.ori.fa
#frequency for each 3mer in reference genome
HG_3MER /path-to-BSAlignerIndex/hg19.3mer
#windows around TSS (ucsc reference gene)
TSS     /path-to-BSAlignerIndex/TSS.win.bed
#sequencing data in fastq format
SEQ_FORMAT     fq
#prefix for each output result
OUT_PREFIX     test
#sequencing mode in a paired-end manner (PE) or single-end manner (SE)
SEQ_MODE       PE
#how many first cycles supposed to be used
#for example, 75 means the cycles after 75th would be omitted
USED_CYCLES    75
#how many threads supposed to be used for the BSAligner
THREAD 20
#whether to merge the all of alignments in this batch
#0 mean don't merge; 1 means merge
MERGE   0
#window size to profile the methylation density across the genome
#only the CpG sites are considered
BIN_SIZE_CPG   100e3
#window size to profile the methylation density across the genome
#only the CpG sites are considered
BIN_SIZE_NONCPG        100e3
#how many total cycles expected to be used (read1+read2).
SEQUENC_TOT_CYCLE      150
#a separated file recording the path of fastq files as well as the
#sample names to be analyzed (see Part IV)
INFO    ./info
```

IV. The ***info*** file is required for ***CONF***. It records the location of raw data as well as sample information.

```
#sample lane    description           path-to-read1.fq       [path-to-read2.fq
for paired-end reads]
PW396w  7       PW396w  /path-to/PW396w.read1.fq      /path-to/PW396w.read2.fq
CVS396  8       CVS396  /path-to/CVS396.read1.fq      /path-to/CVS396.read2.fq
```

V. If you need to perform DMR identification, you should first split the methylation call by each chromosome using the following commands.

```
#Please change to Methy-Pipe output directory,
#then type in the following commands:

../../methy-pipe2/utils/anno/split_call.sh \
       test.CVS396_8.W.call test.CVS396_8.C.rev.call CVS396_8_split CVS396_8
../../methy-pipe2/utils/anno/split_call.sh \
       test.PW396w_7.W.call test.PW396w_7.C.rev.call PW396_7_split PW396w_7

#or users can directly run the demo_split_call.sh in Methy-Pipe output
directory.
```

VI. You can use the following commands to identify DMRs:

```
#Please change to Methy-Pipe output directory,
#then type in the following commands:

mkdir DMRs
cd DMRs
../../../methy-pipe2/utils/DMR_calling/auto_DMR.biomarker.sh \
       ../CVS396_8_split ../PW396_7_split CVS396_refto_PW396w

#or users can directly run the demo DMRs.sh in Methy-Pipe output directory.
```

VII. You can further annotate DMRs to closest genes using the following script in DMRs directory by:

```
perl ../../../methy-pipe2/utils/DMR_anno/dmr_anno.pl \
       ../../../methy-pipe2/bed_files/iGenome.hg19.revised.gff3 \
      all.hyper.call.filtered > all.hyper.call.filtered.anno2gene.xls
```

VIII. The methylation in any arbitrary region can be calculated by following script:

```
perl utils/regional_meth_density/calc_regional_met_density.pl \
      region.bed sample.W.CpG.call sample.C.rev.CpG.call > output
```

IX. Examples of Methy-Pipe output.

a. Alignment results (*.bsalign)

Read ID	Read	Quality	No. of hits	Read 1/2	Read length	Strand	Chr	Position	No. of mismatches	mismatch tracking	CIGAR string	alignment tracking	Watson /Crick
HWI-ST328:7:1101:12	AGAATTTATGTTA	BJPJ`Z^S`b	1	a	62	+	chr4	1.2E+08	0		62M	62	C
HWI-ST328:7:1101:12	AATAGTTTATGAT	fehdba^N\	1	b	62	-	chr4	1.2E+08	1	G->9T24	62M	9G52	C
HWI-ST328:7:1101:11	GGATATTGTATG	JJ\c`cSc^\[	1	a	62	+	chr1	2.9E+07	0		62M	62	W
HWI-ST328:7:1101:11	GGATATTGTATG	^[a[fff^^X<	1	b	62	-	chr1	2.9E+07	0		62M	62	W

b. Methylation calling (*.call)

chr	Position	Base in reference	Total depth	Cytosine counts	Thymine counts	Sequence context
chr1	11310	c	1	0	1	c:a:g
chr1	11315	c	1	0	1	c:c:c
chr1	11316	c	1	0	1	c:c:t
chr1	11317	c	1	0	1	c:t:c
chr1	11319	c	1	0	1	c:t:t

c. Methylation density (*.density)

Chr	start	end	Cytosine counts	Thymine counts	Methylation density (%)
chr1	0	1000000	1008	556	64.5
chr1	1000000	2000000	4166	1459	74.1
chr1	2000000	3000000	2384	966	71.2
chr1	3000000	4000000	2101	502	80.7
chr1	4000000	5000000	2164	660	76.6
chr1	5000000	6000000	2143	453	82.6
chr1	6000000	7000000	3577	1045	77.4
chr1	7000000	8000000	2692	649	80.6

d. Regional methylation density calculation

Chr	Start	End	Description	Cytosine counts	Thymine counts	Methylation density
chr3	184375797	184376100	AluSx,SINE/Alu	7	1	87.5
chr7	29320335	29320416	MIRb,SINE/MIR	0	1	0.0
chr14	45211549	45211841	AluJb,SINE/Alu	6	1	85.7
chr12	28050879	28051088	MIR3,SINE/MIR	0	2	0.0
chr22	23351240	23351896	PABL_A,LTR/ERV1	3	1	75.0

e. DMR identification (DMRs/all.hypo.filtered)

				Test		Control		Methylation			CpG number	
Chr	Start	End	hypo/hyper	Cytosine counts	Thymine counts	Cytosine counts	Thymine counts	Test	Control	P-value	Test	Control
chr10	38816800	38818300	hypo	132	191	279	52	40.87	84.29	2.91E-08	30	25
chr10	42383000	42396900	hypo	33927	28610	46262	8388	54.25	84.65	0.00E+00	474	434
chr10	42596500	42598500	hypo	8944	6579	11626	1476	57.62	88.73	0.00E+00	61	57
chr10	42598700	42600700	hypo	13163	11847	18054	4509	52.63	80.02	0.00E+00	69	60
chr10	1.28E+08	1.28E+08	hypo	8	118	82	85	6.35	49.1	2.93E-09	21	29
chr14	19640700	19642900	hypo	62	434	195	92	12.5	67.94	7.36E-28	68	48
chr11	65779200	65779700	hyper	17	18	0	35	48.57	0	1.63E-04	7	7
chr9	66455500	66456000	hyper	20	33	0	41	37.74	0	2.23E-04	9	6

f. DMR annotation

				Test		Control		Methylation			CpG number		
Chr	Start	End	hypo/hyper	Cytosine counts	Thymine counts	Cytosine counts	Thymine counts	Test	Control	P-value	Test	Control	Gene associated regions
chr10	5093000	5093500	hypo	20	37	51	12	35.09	80.95	8.49E-03	6	5	AKR1C3:intron
chr10	5094500	5096400	hypo	182	111	322	39	62.12	89.2	2.88E-03	27	28	AKR1C3:intron
chr10	5117500	5118000	hypo	8	27	29	2	22.86	93.55	1.93E-03	5	5	AKR1C3:intron
chr10	5210900	5211400	hypo	18	47	59	9	27.69	86.76	2.71E-04	5	6	AKR1CL1:intron
chr10	5237400	5237900	hypo	29	39	52	3	42.65	94.55	6.41E-03	6	6	AKR1C4:promoter
chr10	5241000	5242000	hypo	62	94	221	36	39.74	85.99	9.31E-06	15	17	AKR1C4:intron
chr10	5405100	5408700	hypo	280	408	640	90	40.7	87.67	4.22E-18	70	66	UCN3:5UTR
chr10	5410400	5417300	hypo	596	640	1279	172	48.22	88.15	5.70E-22	126	125	UCN3:5UTR
chr10	5436800	5437800	hypo	146	125	276	19	53.87	93.56	2.89E-05	19	19	TUBAL3:CDS:3
chr10	5441300	5441800	hypo	9	27	53	9	25	85.48	2.32E-03	6	6	TUBAL3:intron
chr10	5486800	5487300	hypo	54	46	91	3	54	96.81	8.87E-03	7	5	NET1:promoter

X. If you have any other question, please contact jiangpeiyong@cuhk.edu.hk