MethGo Documentation

Release 1.0

Wen-Wei Liao

Apr 24, 2017

Contents

DNA methylation is a major epigenetic modification regulating several biological processes. A standard approach in the study of DNA methylation is bisulfite sequencing (BS-Seq). BS-Seq couples bisulfite conversion of DNA with next generation sequencing to provide a genome wide profile of DNA methylation at single base resolution. The analysis of BS-Seq data involves the use of customized aligners for mapping reads and additional bioinformatic pipelines for downstream data analysis. While most post-alignment programs generate methylation calls, MethGo carries out subsequent genomic and epigenomic analyses to comprehensively explore BS-Seq datasets.

MethGo is a simple and effective tool designed for the analysis of data from whole genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS). MethGo provides 5 major modules:

- COV: Coverage distribution of each cytosine
- MET: Both global and gene-centric cytosince methylation levels
- TXN: Cytosine methylation levels at transcription factor binding sites (TFBSs)
- SNP: Single nucleotide polymorphism (SNP) calling
- CNV: Copy number variation calling

Contents:

CHAPTER 1

Installation

1. Obtain Python 2.7 and virturalenv.

   Note: MethGo depends on SAMtools and BEDtools, so please make sure you already have them on your server.

2. Create a virtual environment somewhere on your disk, and then activate it.

```
$ virtualenv --no-site-packages --python=python2.7 methgo_env
$ source methgo_env/bin/activate
```

3. Download the source code and install the requirements.

```
$ git clone https://github.com/paoyangchen-laboratory/methgo.git
$ pip install -r methgo/requirements/base.txt
$ pip install -r methgo/requirements/addition.txt
```

   Note: If you're using Mac and the OS version is larger than 10.8, run the following line before you install the requirements:

```
$ export CFLAGS=-Qunused-arguments
```

   pip will install the following packages:

   - NumPy
   - SciPy
   - matplotlib
   - pandas
   - PySAM (0.8.0)

- Biopython
- pyfasta
- Cython
- pybedtools

4. Add your MethGo path to the PATH environment variable.

CHAPTER 2

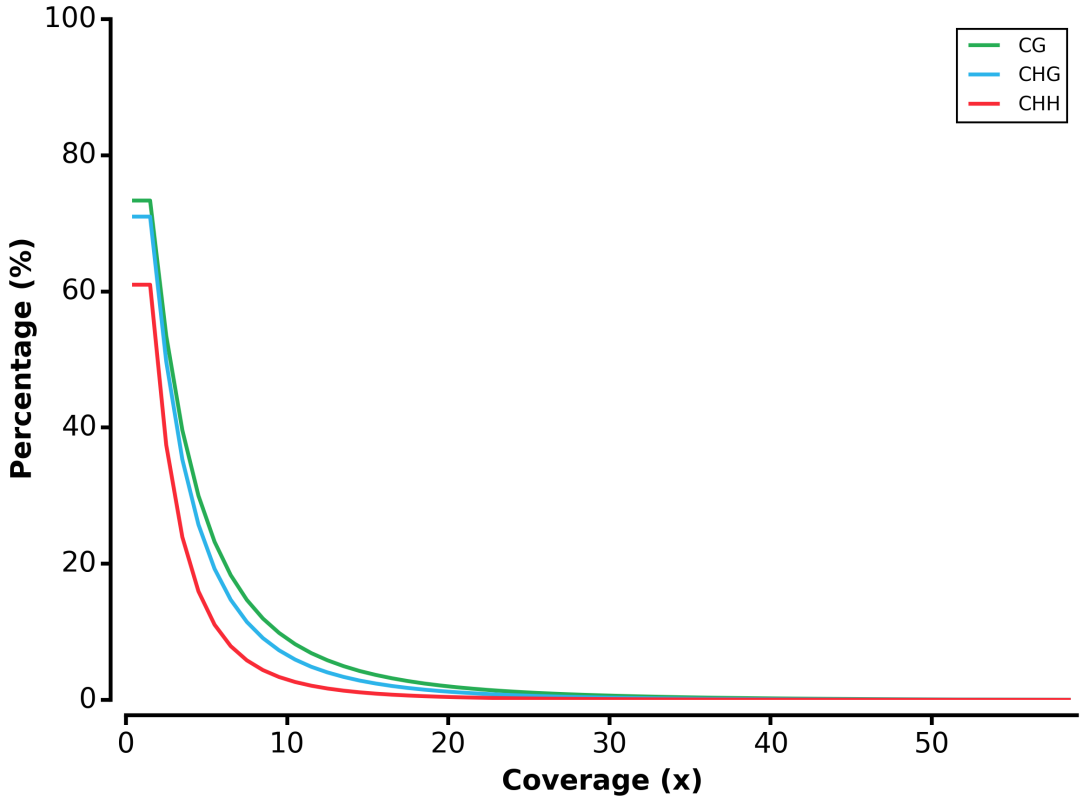
Tutorial

1. Download the sample input file

```
$ curl -O http://paoyang.ipmb.sinica.edu.tw/public_data/methgo_demo.tar.gz
$ tar xvfz methgo_demo.tar.gz
$ cd methgo_demo/data
```

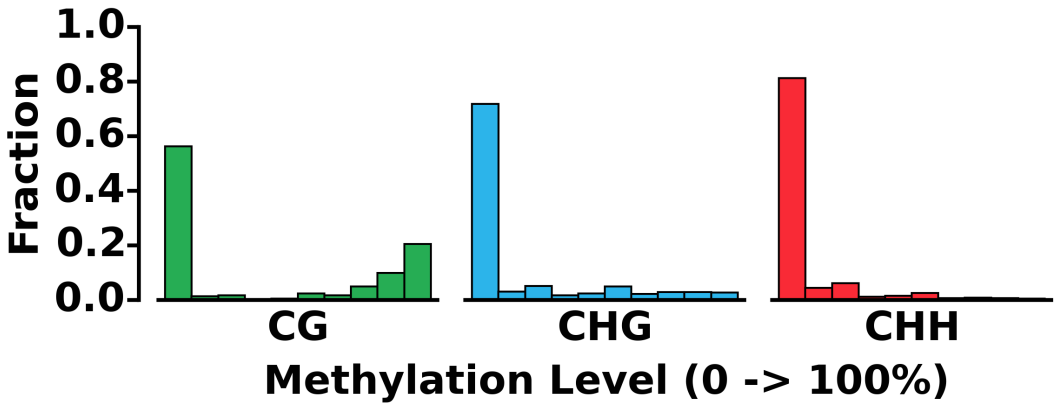
2. Run COV module:

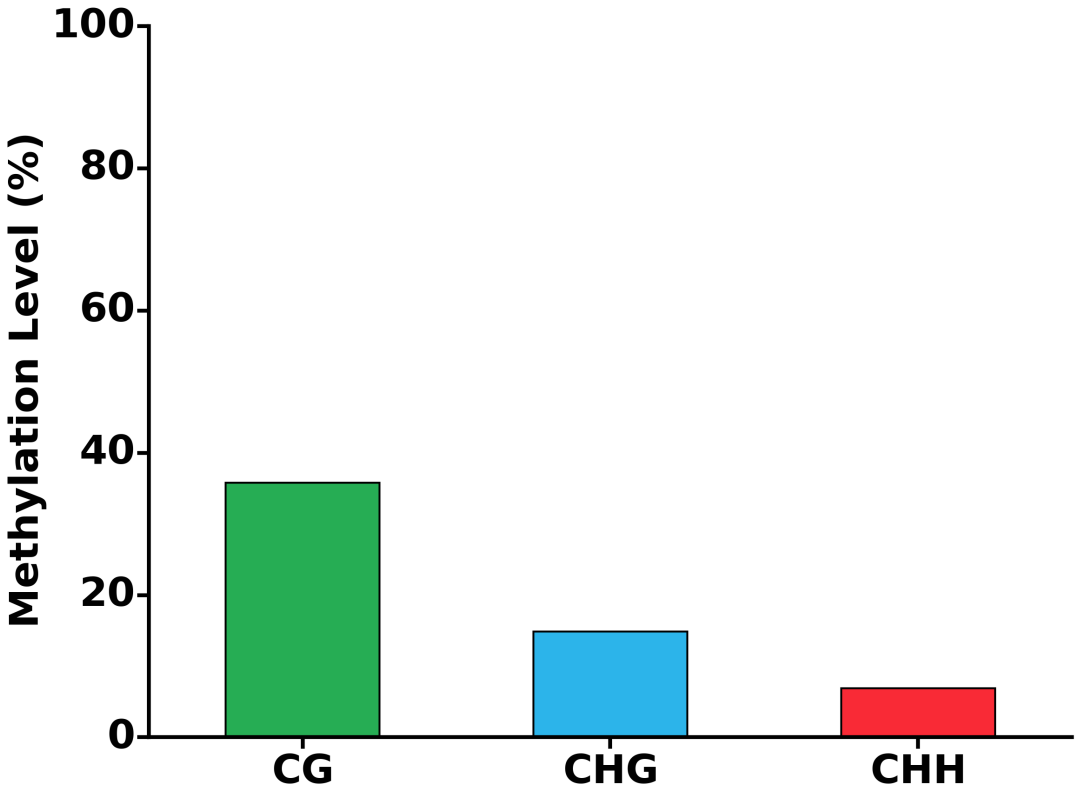
```
$ methgo cov genome.fa demo.CGmap
```

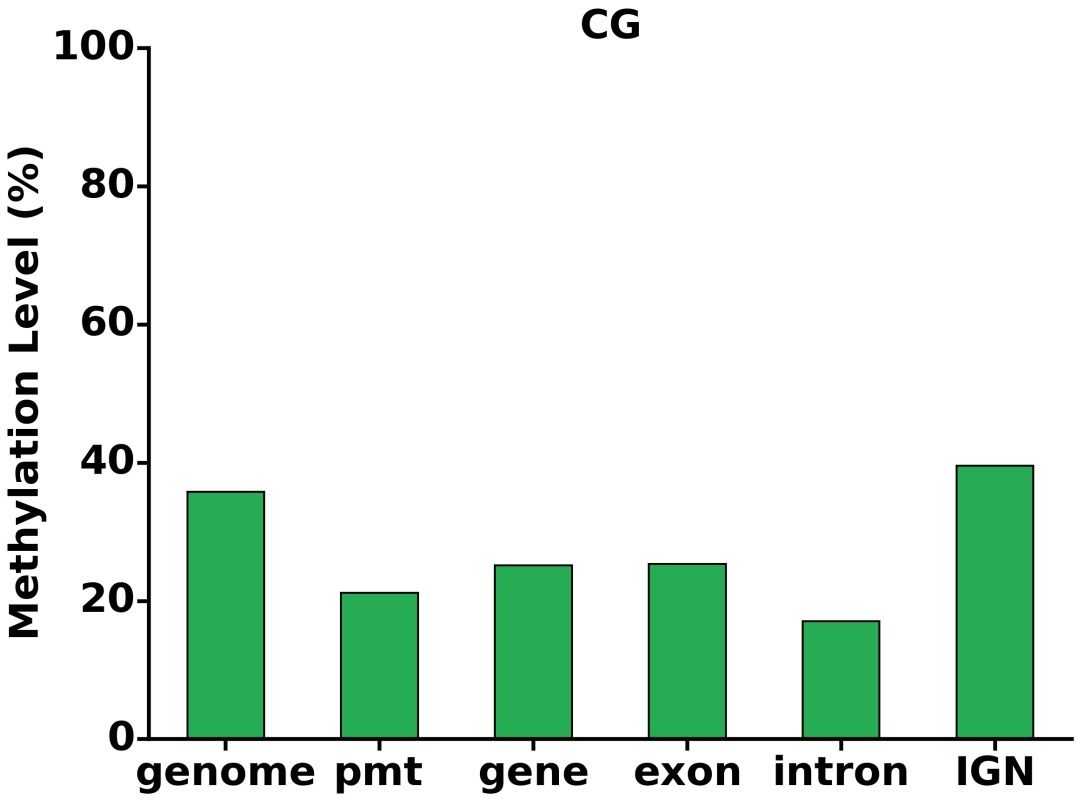


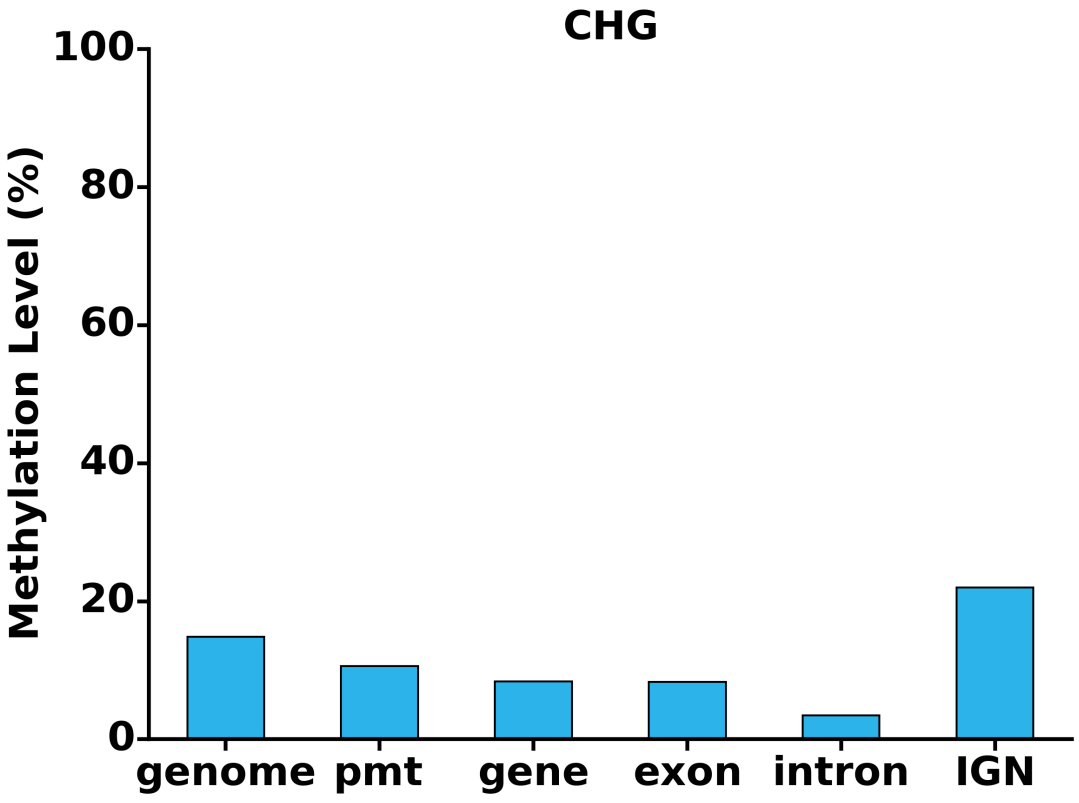
3. Run MET module:

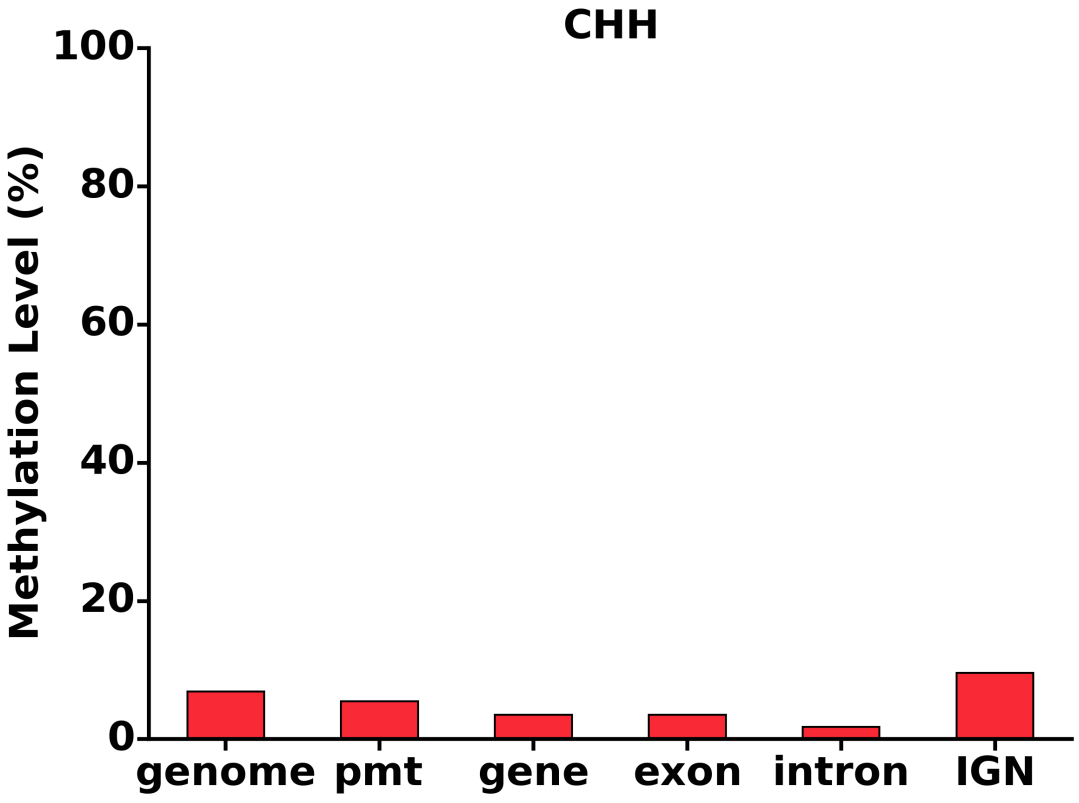
```
$ methgo met genes.gtf genome.fa demo.CGmap
```

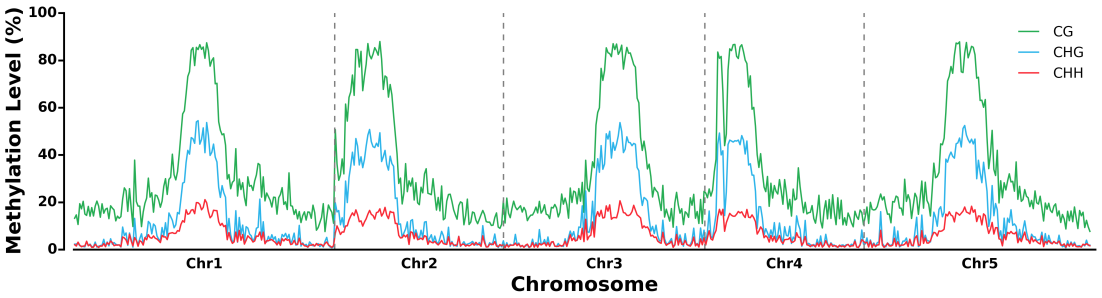



Methylation Level (%)
100
80
60
40
20
0
CG
CHG
CHH


CG
Methylation Level (%)
100
80
60
40
20
0
genome
pmt
gene
exon
intron
IGN

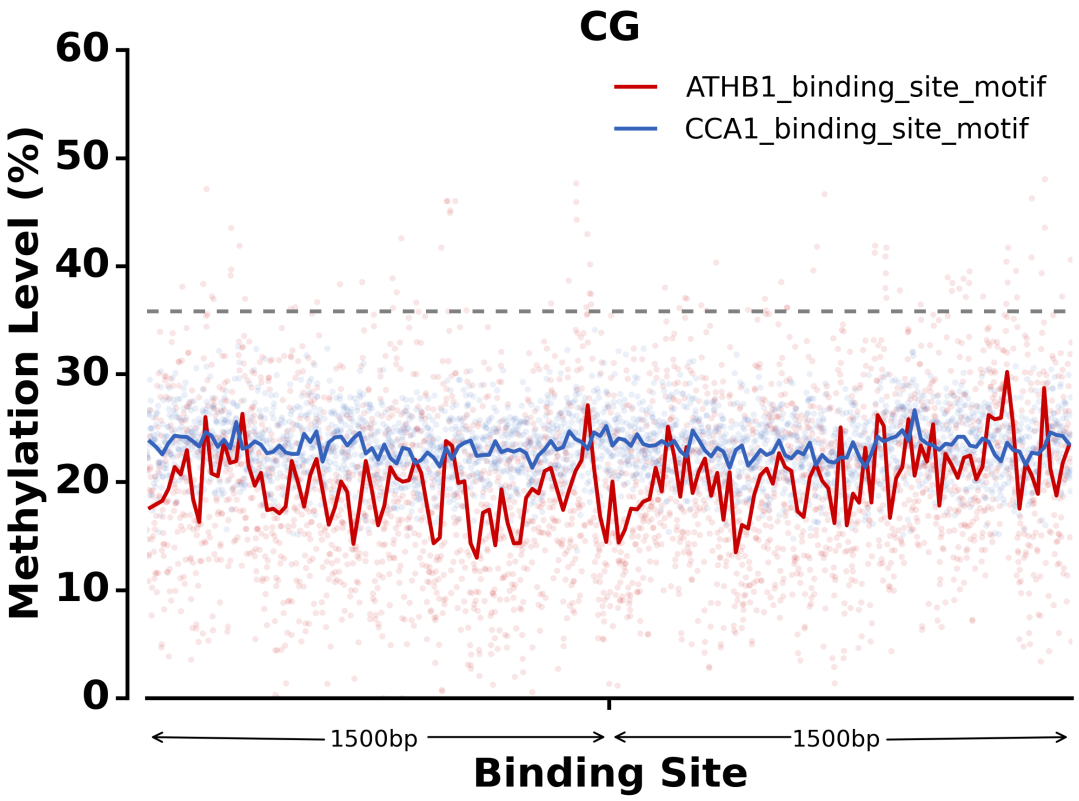

CHG
Methylation Level (%)
100
80
60
40
20
0
genome
pmt
gene
exon
intron
IGN

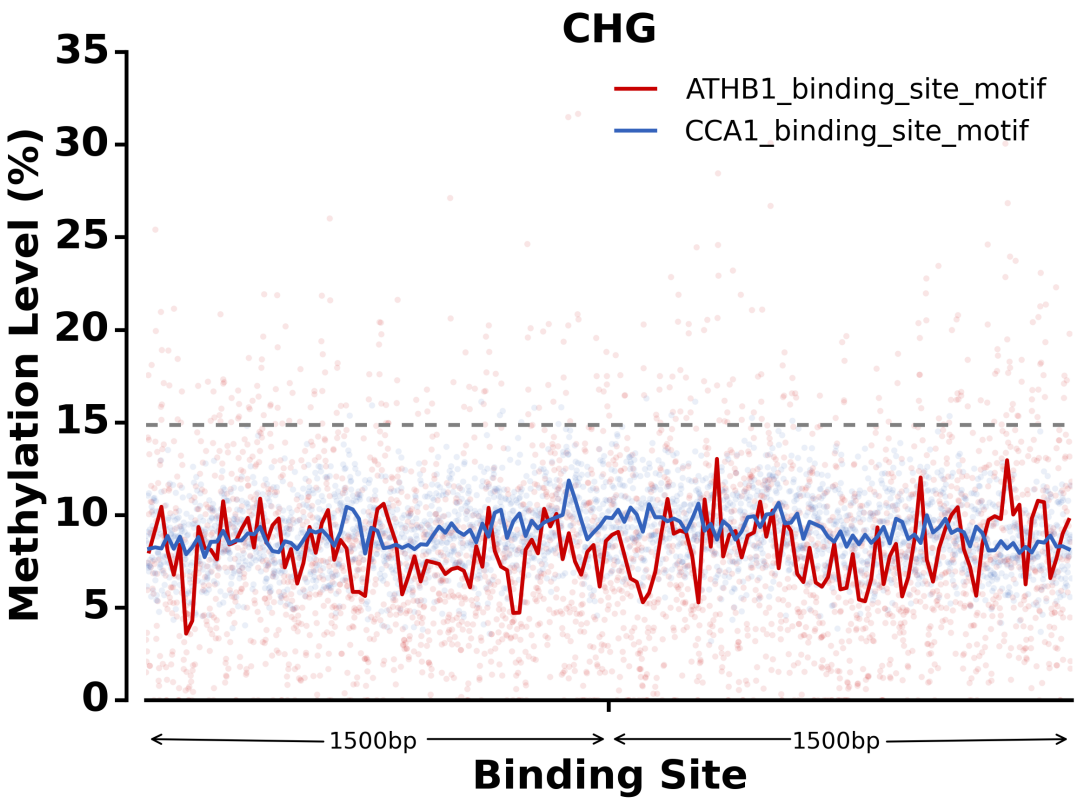

CHH
Methylation Level (%)
100
80
60
40
20
0
genome
pmt
gene
exon
intron
IGN

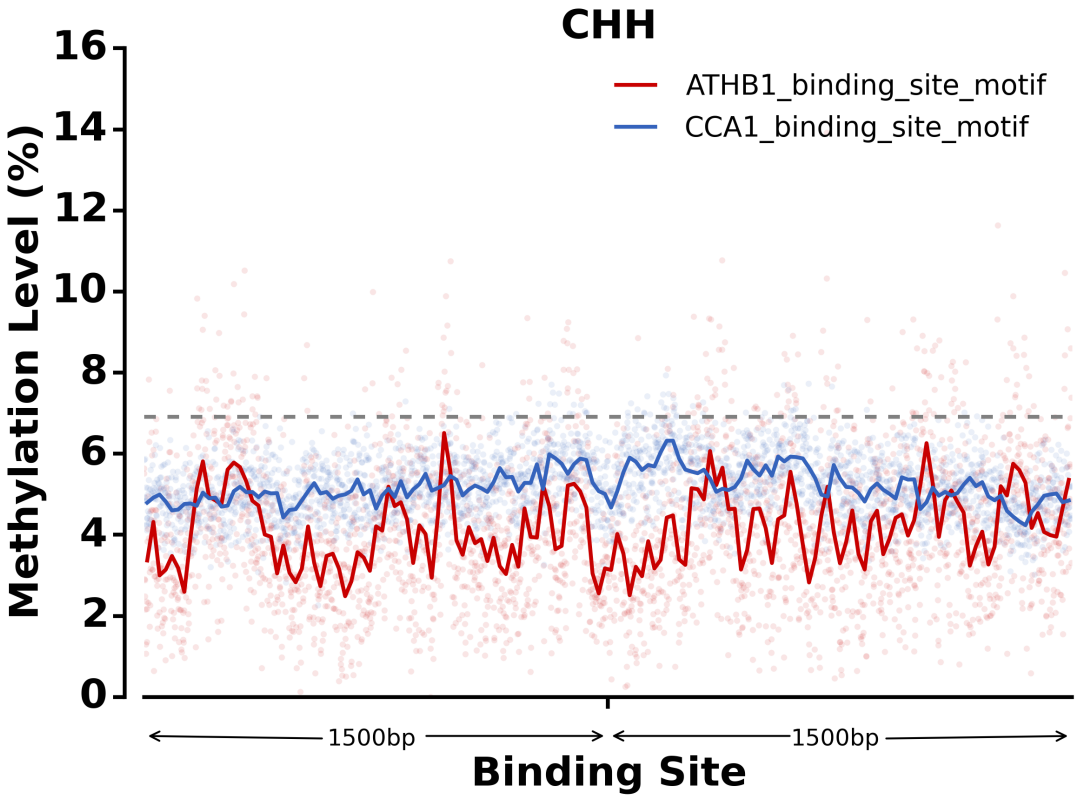


4. Run TXN module:

```
$ methgo txn -t methgo/scripts/txn/tair10_txn -l ATHB1_binding_site_motif,
↪CCA1_binding_site_motif -c demo.CGmap
```




CHG
ATHB1_binding_site_motif
CCA1_binding_site_motif
Methylation Level (%)
35
30
25
20
15
10
5
0
1500bp
1500bp
Binding Site

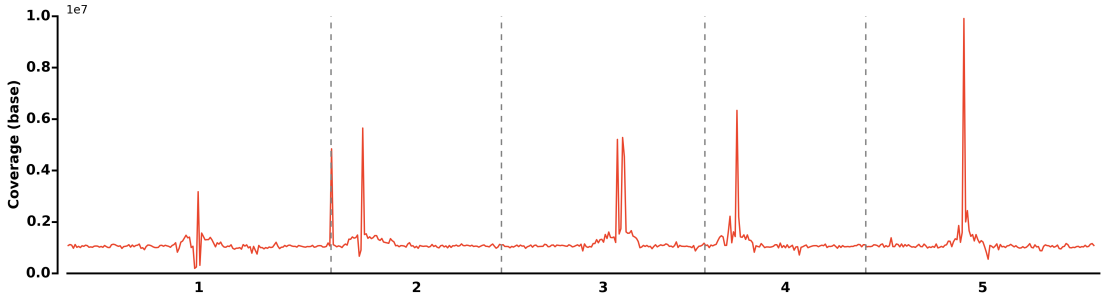

CHH
ATHB1_binding_site_motif
CCA1_binding_site_motif
Methylation Level (%)
16
14
12
10
8
6
4
2
0
1500bp
1500bp
Binding Site

5. Run SNP module:

```
$ methgo snp -g genome.fa demo.sorted.bam
```

6. Run CNV module:

```
$ methgo cnv genome.fa.fai demo.sorted.bam
```



CHAPTER 3

User's Guide

COV

The COV module uses information from methylation calls in the form of a CGmap to calculate coverage of all methylation sites in three contexts. The inputs are a CGmap, and reference fasta. For each cytosine position, the coverage is determined, the first to identify all the cytosine sites from reference genome, and second to extract coverage information from CGmap file. The reverse cumulative plot for coverage distributions is reported.

Input

fasta

Input reference genome FASTA file

cgmap

Input CGmap file

Arguments

-h, –help

Show the help message and exit

MET

The MET module uses information from methylation calls in the form of a CGmap and gene annotations to identify gene methylation levels and their respective promoter methylation levels. The inputs are a CGmap, gene annotation

file, and reference fasta. For each chromosome in the reference genome, two dictionaries are generated, the first to mark the positions of methylation for each methylation contexts, and the second to note the methylation levels at each point of methylation. A list of genes and their bounds are generated from the gene annotation file in General Transfer Format (GTF). For each gene, all the methylation levels of 3 contexts within the gene bounds are averaged to produce its gene body methylation level. Promoters with a size specified by the user are then analyzed in the same manner to produce promoter methylation levels for each gene. Based on methylation contexts, results of gene methylation levels sorted by their gene ID are outputted in a text file.

Input

gtf

Input GTF file

fasta

Input reference genome FASTA file

cgmap

Input CGmap file

Arguments

-d, –depth <INT>

Minimum depth of reads desired, default is 4

-p, –pmtsize <INT>

Size of promoter, default is 1,000

-w, –winsize <INT>

Size of sliding window, default is 200,000

-h, –help

Show the help message and exit

TXN

The TXN module plots the methylation levels adjacent to the transcription factor binding sites (TFBSs) from methylation calls in the form of a CGmap and TFBS annotations. The inputs are a CGmap. For each TFBS, the methylation levels of sites within 1,500 bp are averaged over tiling windows (30 bp). The methylation levels distributions are reported in a scatter plot with smooth curve.

Input

-t, –txnfiles <PATH>

Path to the preprocessing TXN files

-l, –txns <LABELS>

List of TXN labels

-c, –cgmap <FILE>

Path to CGmap file

Arguments

-h, –help

Show the help message and exit

-m, –mincov <INT>

Set the minimum coverage

CNV

The CNV calling module investigates the number of copies of genes in the genotype of an individual to find areas in the genome likely to have large-scale genome rearrangement. Inputs include a sorted BAM file and reference genome index. The PySAM pileups method is used in order to obtain the number of bases for reads at each position in the reference genome. Using the window size given by the user (default is 200,000), all the bases of the reads at the positions within each window are summed up. The standard Z score is calculated and converted to a P-value for each window. If there are 3 windows in a row (user can change this default setting) and all their P-values are smaller than a given threshold (default is 0.05), then this region is considered a CNV.

Input

refindex

Input rference genome index file

bamfile

Input BAM file

Arguments

-w, –winsize <INT>

Size of tiled window, default is 20,000 bp

-p, –pvalue <FLOAT>

P-value to be considered a possible region, default is 0.05

-s, –succession <INT>

Number of successive possible regions to be considered a CNV, default is 3

-h, –help

Show the help message and exit

SNP

The PySAM pileups method is used in order to obtain the alleles for reads at each position in the reference genome. Allele counts are then determined for each position, and if the coverage, or number of reads present at that position, exceeds a given amount (default is 5), the alleles at that position are analyzed for the presence of a SNP. Homozygous SNPs are considered to have occurred at positions in which an allele exists with a frequency higher than the given major allele frequency (default is 0.9) in the reads that overlap at that position. Additionally, the major allele needs to differ from the allele in the reference genome at that position. Heterozygous SNPs are considered to have occurred when two alleles occur with frequencies in the reads within a range close to 0.5. A buffer is set (default is 0.1) around 0.5 for the frequencies of the two alleles to be within (so default frequencies are 0.4-0.6) to be considered a heterozygous SNP.

Input

bamfile

Input BAM file

-g, –genomeFile <FILE>

Input reference genome FASTA file

Arguments

-c, –coverage <INT>

Coverage or minimum number of reads desired, default is 5

-m, –majorAlleleFreq <FLOAT>

Frequency to be considered homozygous allele, default is 0.9

-b, –buffer <FLOAT>

Buffer on either side of 0.5 to be considered heterozygous allele, default is 0.1

-h, –help

Show the help message and exit

CHAPTER 4

Indices and tables

- genindex
- modindex
- search