# MHap

## *Maximum likelihood inference of methylation haplotypes*

### *General information*

MHap is used to infer methylation haplotypes and to estimate methylation haplotype frequencies as well as methylation entropy from DNA methylation data derived from short-read sequencing.

Genomic regions of interested are given by the user. MHap uses a sliding window to select windows with n number of CpG with intermediate methylation. For each genomic region, a set of windows is retrieved and the analysis is performed in each one of them.

Making use of the Expectation-Maximization (EM) algorithm, MHap identifies all methylation haplotypes consistent with the sequence reads and estimates the frequency of each haplotype within the windows. Once the algorithm has converged, the inferred haplotype frequencies are used to calculate the Shannon entropy. Also the methylation fraction of individual CpGs is calculated using the methylation states provided by the sequencing reads and using the haplotype frequencies.

MHap is freely available at https://github.com/bzmartinelli/MHap.

### *Input*

MHap takes as input two files:

1) file containing the genomic regions of interest in a .bed format.
   Example:

| chromosome | start | end |
|---|---|---|
| chr21 | 9412099 | 9415796 |
| chr15 | 20000038 | 20009331 |

2) file containing the methylation data. MHap accepts the output file from Bismark (generated by bismak_methylation_extractor), and the output from BisSNP (the CpG reads file).

### *Usage*

```
python MHap.py -gr <genomic_regions_file> -data <cpg_reads_file>
-data_from <bissnp or bismark> [options]
```

## Arguments

*Required*

-gr / --genomic_regions
File in .bed format containing the genomic regions of interest.

-data / --cpg_reads_data
File containing the methylation data.

-data_from / --data_from
Where the input data is from. Two options are accepted: bissnp or bismark, which is passed to the command line as *-data_from bissnp* or *-data_from bismark.*

*Optional*

-ncpgs / --number_of_cpgs
Number of CpGs per window (default = 5).

-mmin / --meth_min
Minimum methylation percentage of each CpG included in the analysis.

-mmax or --meth_max
Maximum methylation percentage of each CpG included in the analysis.

-max_iter / --max_iterations
Maximun number of iterations for the EM algorithm (default = 1000)

-conv / --convergence_threshold
The threshold for the termination of the algorithm (default = 0.000001)

-freq_cutoff / -- min_freq_cutoff
Minimum frequency to display the haplotype frequencies (default = 0.001)

-initial_freq / --initial_freq_em
The initial frequencies for the EM algorithm. A uniform distribution is used as default, where the initial frequencies are $f^{(0)}$ = 1/number of haplotypes. Optionally, the algorithm can start with random frequencies from a symmetric Dirichlet distribution by passing the argument *-initial_freq random* to the command line.

## Output

Three output files are generated:

1) File containing the inferred methylation haplotypes and their frequencies. The haplotypes are composed by a combination of 1 and 0, representing a methylated and unmethylated CpG, respectively. The output file looks like this (tab separated):

| Genomic_region | Window_start | Window_end | Haplotype | Frequency |
|---|---|---|---|---|
| chr15:20000038-20001331 | 20001199 | 20001244 | 11111 | 0.142213596065 |
| chr15:20000038-20001331 | 20001199 | 20001244 | 10101 | 0.139995097475 |
| chr15:20000038-20001331 | 20001199 | 20001244 | 10111 | 0.083239206257 |
| chr15:20000038-20001331 | 20001221 | 20001331 | 11110 | 0.121290509717 |

2) File containing the estimated methylation entropy of each window. It looks like this (tab separated):

| Genomic_region | Window_start | Window_end | Estimated_Entropy |
|---|---|---|---|
| chr15:20000038-20001331 | 20001199 | 20001244 | 0.575174578809 |
| chr15:20000038-20001331 | 20001200 | 20001245 | 0.601738171849 |
| chr15:20000038-20001331 | 20001221 | 20001331 | 0.550343857703 |

3) File containing the proportion of CpGs in a methylated state based on the calculation using the sequence reads or the haplotype frequencies. The file looks like this (tab separated):

| Genomic_region | Window_start | Window_end | CpG_position | From_reads | From_haplotypes |
|---|---|---|---|---|---|
| chr15:20000038-20001331 | 20001199 | 20001244 | 20001200 | 0.454545454545 | 0.451095028167 |
| chr15:20000038-20001331 | 20001199 | 20001244 | 20001221 | 0.6 | 0.630378466441 |
| chr15:20000038-20001331 | 20001199 | 20001244 | 20001244 | 0.736842105263 | 0.7271302952 |

The output files are generated in a folder named MHap_output_(+ date and time), which is created in the same directory where you run MHap.

## Example

Example files to demonstrate the usage of MHap are available. A typical command to run MHap using this data looks like this:

```
python MHap.py -gr genomic_regions_example.bed -data cpg_reads_example.bed -data_from bissnp
```

or including some optional arguments:

```
python MHap.py –gr genomic_regions_example.bed –data cpg_reads_example.bed -data_from bissnp -mmin 20 -freq_cutoff 0.01 -initial_freq random
```

## Contact

Please, feel free to get in touch by email martinelli.bz@gmail.com and https://github.com/bzmartinelli/MHap/issues.