

# Genomewide methylation heterogeneity and differential heterogeneity analysis

## Contents

<b>1</b>	<b>Overview</b>	<b>1</b>
<b>2</b>	<b>Methylation heterogeneity profiling (MeH profiling)</b>	<b>2</b>
2.1	Obtaining inputs for genome screening . . . . .	3
2.2	Parameters for genome screening . . . . .	4
2.3	Output from genome screening . . . . .	5
2.3.1	Default output . . . . .	5
2.3.2	Optional output . . . . .	5
<b>3</b>	<b>Differential heterogeneous (methylation) regions (DHR) analysis</b>	<b>6</b>
3.1	Windows to bins of 400bp . . . . .	6
3.2	DHR Identification . . . . .	6
3.2.1	Statistical tests . . . . .	7
3.2.2	Annotation to input . . . . .	7

## 1 Overview

The overview of the analysis of methylation heterogeneity can be found in [Figure 1](#). The analysis is divided into two main steps: genome screening and differential heterogeneity analysis. Genome screening includes obtaining information necessary for the estimation of methylation heterogeneity (MeH) from its raw data. It extracts bisulfite converted reads and the genomic locations they're mapped to to summarise the compositions of methylation patterns within windows of multiple cytosines (of a given context; i.e., CG). It then evaluates MeH given the window has enough reads and outputs the results (MeH and the starting CpG position) as a row in an excel file. This process is done within a python program called `genome_scr.py`. The program not only evaluates MeH genome-wide but also preprocess (split the bam files for multiprocessing) for computational efficiency and merge results from several files (samples) after genome screening. After the results from several samples are merged, differential heterogeneous regions can be identified using DHR.R.

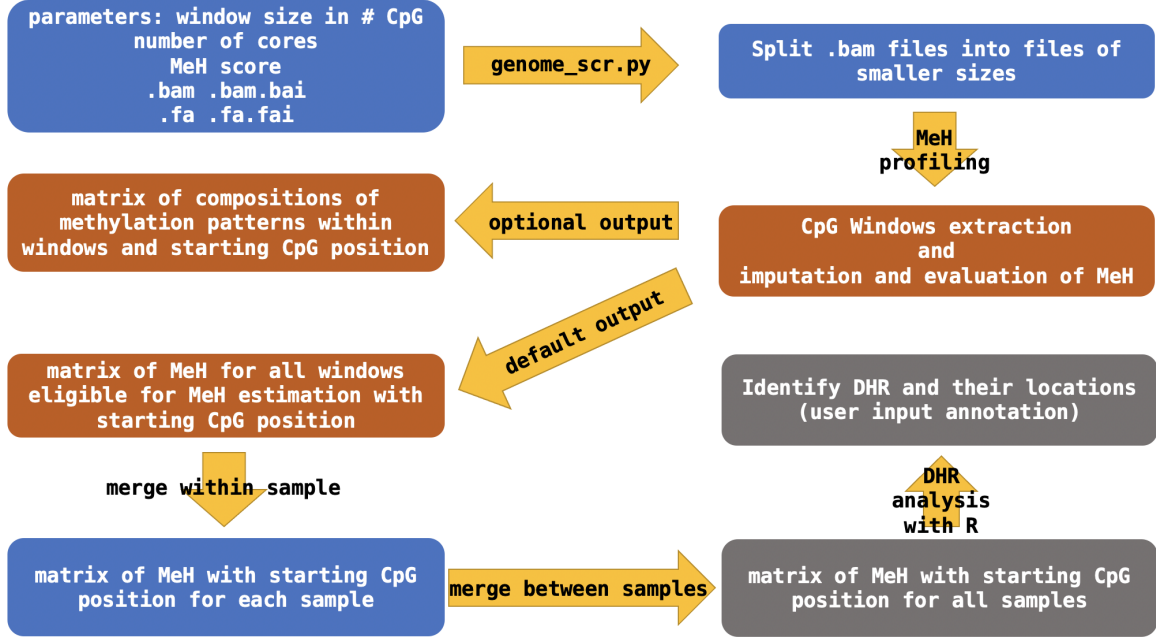


Figure 1: Overview of using the program to analyse MeH.

## 2 Methylation heterogeneity profiling (MeH profiling)

Genome screening is the process of obtaining information necessary for the estimation of methylation heterogeneity, estimating methylation heterogeneity and recording the results. It is done by sliding windows of 1 cytosine and recording the bisulfite converted reads mapped to the genomic location. Once enough cytosines are gone through, they are imputed if eligible and evaluated for MeH if there are enough reads. There are a few choices offered for users. They are window size  $w$  or the number of cytosines for the evaluation of 1 methylation heterogeneity score, the default is 4, minimum number of reads for the evaluation of MeH, the default is  $2^{w-2}$ , number of cores  $c$  to be used for multiprocessing, the default value is 4, what methylation contexts to consider, the default is CG sites, and whether the compositions of methylation patterns will be outputted, i.e., CG, CHG or CHH (or any combination of these). Since information such as methylation patterns including methylation statuses and genomic position mapped to for each base pair of every read is required, the program needs both the original alignment files and the reference genome, that are given by ".bam" and ".fa". In order to obtain ".bam", a standard workflow is also suggested in [Figure 2](#).

## 2.1 Obtaining inputs for genome screening

Before MeH can be profiled, users need to obtain .bam files, that is, mapped bisulfite converted reads. The standard workflow is summarised in [Figure 2](#).

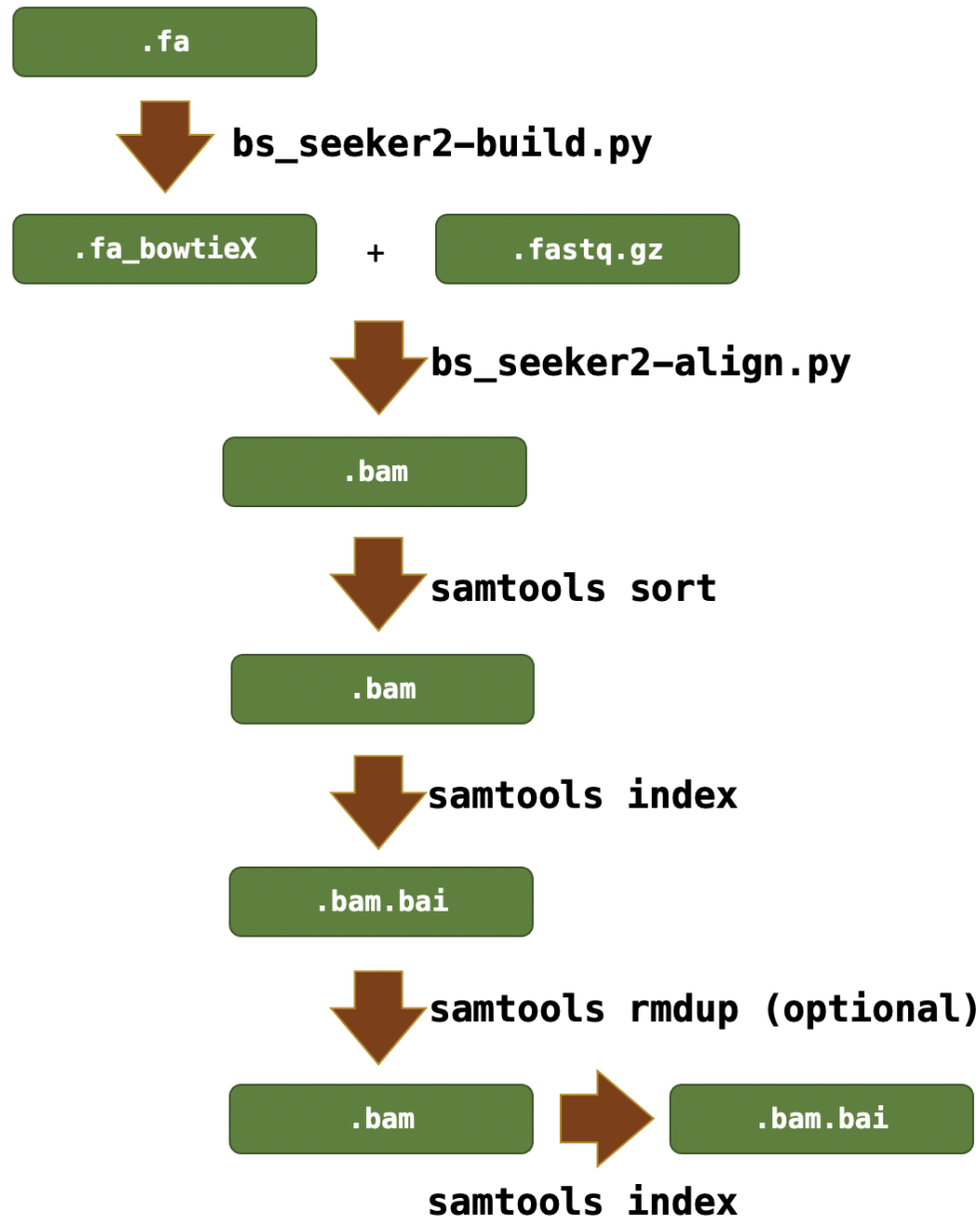


Figure 2: Preprocessing of data before inputting into the program.

Once all the bam files (and their corresponding index files) are obtained, they need to be placed in a folder named "MeHdata" together with the reference genome and its index file.

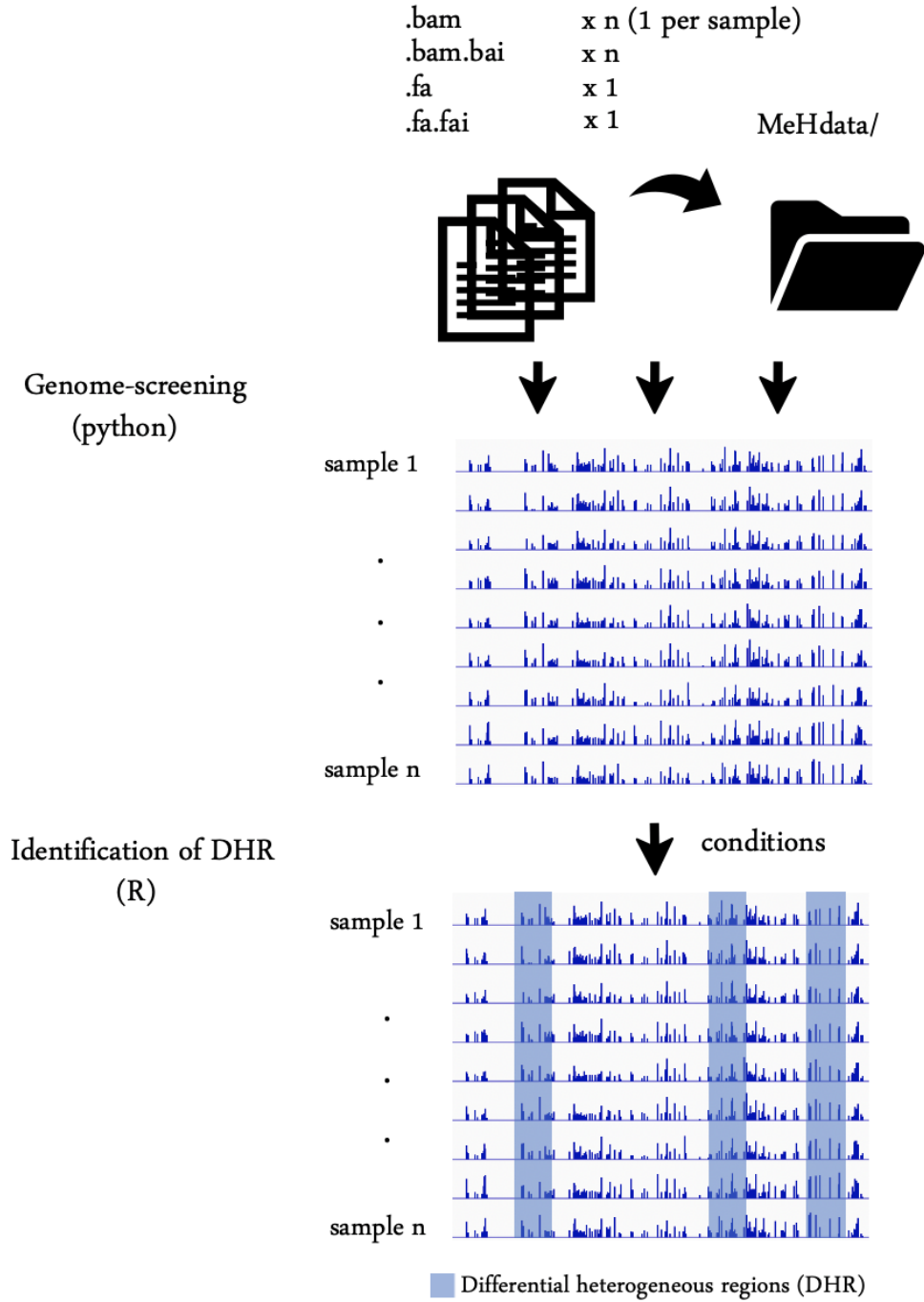


Figure 3: Visualisation of the process.

## 2.2 Parameters for genome screening

- -w, window size in numbers of cytosines, default is 4

- -m, scores for evaluating methylation heterogeneity, default is 2: pairwise-similarity-based method, other options include 1: Phylogeny-based method, 2: Abundance-based method, 4: Entropy and 5: Epipolymorphism.
- -d, distance used to calculate pairwise distances between methylation patterns, default is 1: Hamming distance, other option include 2: weighted degree kernel. Only required for Phylogeny-based method and pairwise-similarity-based method.
- -c, number of cores for parallel processing, default is 4
- -CG, to evaluate context 'CG', default is False
- -CHG, to evaluate context 'CHG', default is False
- -CHH, to evaluate context 'CHH', default is False

## 2.3 Output from genome screening

### 2.3.1 Default output

When the program `genome_scr.py` is run, all bam files in the folder will be processed and results from all samples will be merged into one .csv file containing the MeH of each sample and their genomic locations.

	A	B	C	D	E	F	G
1	chrom	pos	strand	C0031test1234	C0033test1234	C0035test1234	C0037test1234
2	1	553	r	1.41421	0	2.96614	0
3	1	575	r	1.41421	0	1.41421	0
4	1	595	r	1.41421	0	2.00904	0
5	1	637	r	0	0	1.41421	0
6	1	643	r	0	1.41421	1.41421	1.41421

Figure 4: Visualisation of the default output.

### 2.3.2 Optional output

The intermediate product, that is, the compositions of methylation patterns for each window can be outputted if user chooses to. It will also be in .csv with each row containing the compositions of a window and its starting (cytosine) genomic location.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	ML	MeH	chrom	depth	dis	p01	p02	p03	p04	p05	p06	p07	p08	p09	p10	p11	p12	p13	p14	p15	p16	pos	sample
2	1	1.41421	chr1	7	47	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	5	12270	PC3
3	1	1.41421	chr1	7	53	0	0	0	0	0	2	0	5	0	0	0	0	0	0	0	0	12277	PC3
4	0.714	0	chr1	7	50	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	12304	PC3
5	0.815	2.647	chr1	81	87	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	7	14887	PC3
6	1	2.95271	chr1	23	27	0	0	0	3	0	0	0	16	0	0	0	1	0	0	0	3	24905	PC3

Figure 5: Visualisation of the optional output.

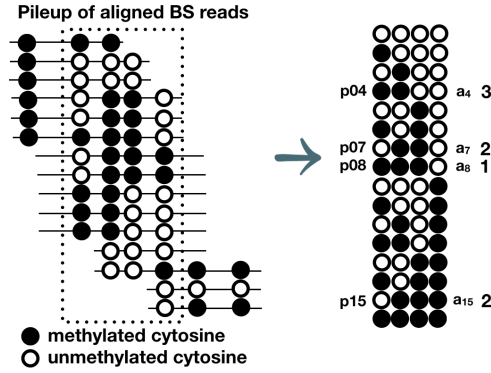


Figure 6: Explanation of how to obtain the abundances.

### 3 Differential heterogeneous (methylation) regions (DHR) analysis

DHR analysis is done in R with example scripts given as an .R file. The users need to load the output into R themselves.

#### 3.1 Windows to bins of 400bp

The outputs from genome\_scr.py contain MeH of every window eligible for MeH estimation. This result can be merged into bins of 400bp where MeH will be averaged within the bins and set to be located at the center of the bin.

#### 3.2 DHR Identification

The users need to give conditions to each sample; i.e., wildtype and mutant or stages 1 to 3 ...etc, and comparisons can be done for user specified conditions; i.e., stage 3 versus stage

1 or stage 2 versus stage 1. The output from this comparison will be a matrix consisting of difference in mean MeH between the conditions specified, p-values given by t-test, mean MeH of samples from the second condition, and mean MeH of samples from the first condition.

### **3.2.1 Statistical tests**

Student's t-test is used for the comparison between conditions specified by the user under the null hypothesis of there is no difference between the mean MeH of the conditions specified and alternative hypothesis of the difference in the mean MeH of the conditions is nonzero. After the comparisons are done, the users can select regions (bins) based on p-value and delta (difference in mean MeH between the conditions specified) cutoffs of interest.

### **3.2.2 Annotation to input**

If user chooses to provide annotation with the name of the genes, chromosome, transcription starting position and ending position then a list of differentially heterogeneous genes can be selected based on criteria given (how many bp upstream and downstream from TSS).