Step1 : trim_galore --clock 옵션을 이용해 barcode 제거 (Perl 파일을 이용해서 barcode 제거한 것과 동일한 과정)

- 우리는 paired-end data를 가지고 있기 때문에 --paired 옵션을 지정해줘야 된다.

```
(Genomic_data_science) pjw@DESKTOP-39ALU06:~/PJW_Workspace/Homework/test_sample$ trim_galore --clock --paired SRR5195663_R1.fastq.gz SRR5195663_R2.fastq.gz
Multicore support not enabled. Proceeding with single-core trimming.
Path to Cutadapt set as: 'cutadapt' (default)
Cutadapt seems to be working fine (tested command 'cutadapt --version')
Cutadapt version: 3.7
single-core operation.

IT'S TIME FOR CLOCK PROCESSING!!!                                                          [pun intended]

Writing dual trimmed version of the input file 'SRR5195663_R1.fastq.gz' to 'SRR5195663_R1.clock_UMI.R1.fq.gz'
Writing dual trimmed version of the input file 'SRR5195663_R2.fastq.gz' to 'SRR5195663_R2.clock_UMI.R2.fq.gz'
               ---
Processed 1000000 sequences so far...
Processed 2000000 sequences so far...
Processed 3000000 sequences so far...
Processed 4000000 sequences so far...
Processed 5000000 sequences so far...
Processed 6000000 sequences so far...
Processed 7000000 sequences so far...
Processed 8000000 sequences so far...
Processed 9000000 sequences so far...
Processed 10000000 sequences so far...
Processed 11000000 sequences so far...
Processed 12000000 sequences so far...
Processed 13000000 sequences so far...
Processed 14000000 sequences so far...
Processed 15000000 sequences so far...
Processed 16000000 sequences so far...
Processed 17000000 sequences so far...
Processed 18000000 sequences so far...
Processed 19000000 sequences so far...
Processed 20000000 sequences so far...
Processed 21000000 sequences so far...
Processed 22000000 sequences so far...
Processed 23000000 sequences so far...
Processed 24000000 sequences so far...
Processed 25000000 sequences so far...
Processed 26000000 sequences so far...
Sequences processed in total: 26991432
thereof had fixed sequence CAGT in both R1 and R2:      3572903 (13.24%)
     ~~~~~~~~~~~~~~~~~~~~~~~~~~~~


Pre-processing finished...

Please run Trim Galore again to remove adapters, poor quality bases as well as UMI/fixed sequences from the 3'-end of the reads.
A sample command for this is:

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
trim_galore --paired --three_prime_clip_R1 15 --three_prime_clip_R2 15 *.clock_UMI.R1.fq.gz *.clock_UMI.R2.fq.gz
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Step2 : trim_galore를 이용해 trimming

- 우리는 paired-end data를 가지고 있기 때문에 --paired 옵션을 지정해줘야 된다.

```
(Genomic_data_science) pjw@DESKTOP-39ALUO6:~/PJW_Workspace/Homework/test_sample$ trim_galore --paired --three_prime_clip_R1 15 --three_prime_clip_R2 15 --fastqc *.clock_UMI.R1.fq.gz *.clock_UMI.R2.fq.gz
Multicore support not enabled. Proceeding with single-core trimming.
Path to Cutadapt set as: 'cutadapt' (default)
Cutadapt seems to be working fine (tested command 'cutadapt --version')
Cutadapt version: 3.7
single-core operation.
No quality encoding type selected. Assuming that the data provided uses Sanger encoded Phred scores (default)



AUTO-DETECTING ADAPTER TYPE
===========================
Attempting to auto-detect adapter type from the first 1 million sequences of the first file (>> SRR5195663_R1.clock_UMI.R1.fq.gz <<)

Found perfect matches for the following adapter sequences:
Adapter type    Count   Sequence        Sequences analysed      Percentage
Illumina        76335   AGATCGGAAGAGC   1000000 7.63
Nextera 5       CTGTCTCTTATA    1000000 0.00
smallRNA        0       TGGAATTCTCGG    1000000 0.00
Using Illumina adapter for trimming (count: 76335). Second best hit was Nextera (count: 5)

Writing report to 'SRR5195663_R1.clock_UMI.R1.fq.gz_trimming_report.txt'

SUMMARISING RUN PARAMETERS
==========================
Input filename: SRR5195663_R1.clock_UMI.R1.fq.gz
Trimming mode: paired-end
Trim Galore version: 0.6.7
Cutadapt version: 3.7
Number of cores used for trimming: 1
Quality Phred score cutoff: 20
Quality encoding type selected: ASCII+33
Adapter sequence: 'AGATCGGAAGAGC' (Illumina TruSeq, Sanger iPCR; auto-detected)
Maximum trimming error rate: 0.1 (default)
Minimum required adapter overlap (stringency): 1 bp
Minimum required sequence length for both reads before a sequence pair gets removed: 20 bp
All Read 1 sequences will be trimmed by 15 bp from their 3' end to avoid poor qualities or biases
All Read 2 sequences will be trimmed by 15 bp from their 3' end to avoid poor qualities or biases
Running FastQC on the data once trimming has completed
Output file(s) will be GZIP compressed

Cutadapt seems to be fairly up-to-date (version 3.7). Setting -j 1
Writing final adapter and quality trimmed output to SRR5195663_R1.clock_UMI.R1_trimmed.fq.gz


>>> Now performing quality (cutoff '-q 20') and adapter trimming in a single pass for the adapter sequence: 'AGATCGGAAGAGC' from file SRR5195663_R1.clock_UMI.R1.fq.gz <<<
```

Step3 : bismark를 이용해서 alignment 진행 (Reference genome : Mus_musculus.GRCm38.69 toplevel.fa)

- bismark_genome_preparation --verbose ../DNA_methylation/Step3_Alignment/Genomes/ 명령어를 먼저 진행한 뒤 실행! (Default로 bowtie2로 실행됨 , Reference genome 파일 경로 확인하고 진행)

```
(Genomic_data_science) pjw@DESKTOP-39ALUO6:~/PJW_Workspace/Homework/test_sample$ bismark --genome ./Genome/ -1 SRR5195663_R1.clock_UMI.R1_val_1.fq.gz -2 SRR5195663_R2.clock_UMI.R2_val_2.fq.gz
Bowtie 2 seems to be working fine (tested command 'bowtie2 --version' [2.4.5])
Output format is BAM (default)
Alignments will be written out in BAM format. Samtools found here: '/home/pjw/anaconda3/envs/Genomic_data_science/bin/samtools'
Reference genome folder provided is ./Genome/   (absolute path is '/mnt/d/Lab_wsl/PJW_Study/Homework/test_sample/Genome/)'
FastQ format assumed (by default)

Input files to be analysed (in current folder '/mnt/d/Lab_wsl/PJW_Study/Homework/test_sample'):
SRR5195663_R1.clock_UMI.R1_val_1.fq.gz
SRR5195663_R2.clock_UMI.R2_val_2.fq.gz
Library is assumed to be strand-specific (directional), alignments to strands complementary to the original top or bottom strands will be ignored (i.e. not performed!)
Setting parallelization to single-threaded (default)

Summary of all aligner options: -q --score-min L,0,-0.2 --ignore-quals --no-mixed --no-discordant --dovetail --maxins 500
Current working directory is: /mnt/d/Lab_wsl/PJW_Study/Homework/test_sample

Now reading in and storing sequence information of the genome specified in: /mnt/d/Lab_wsl/PJW_Study/Homework/test_sample/Genome/

chr 1 (195471971 bp)
chr 10 (130694993 bp)
chr 11 (122082543 bp)
chr 12 (120129022 bp)
chr 13 (120421639 bp)
chr 14 (124902244 bp)
chr 15 (104043685 bp)
chr 16 (98207768 bp)
chr 17 (94987271 bp)
chr 18 (90702639 bp)
chr 19 (61431566 bp)
chr 2 (182113224 bp)
chr 3 (160039680 bp)
chr 4 (156508116 bp)
chr 5 (151834684 bp)
chr 6 (149736546 bp)
chr 7 (145441459 bp)
chr 8 (129401213 bp)
chr 9 (124595110 bp)
chr X (171031299 bp)
chr Y (91744698 bp)
```

Step4 : UmiBam을 이용해 Deduplication 진행

- 여기서 -p 옵션 없이 진행하면 해당 input bam 파일은 single-end로 인식되어서 M-bias plot에서 read1 밖에 그려지지 않음.
- 우리는 paired-end 데이터를 가지고 alignment를 진행했기 때문에 반드시 –p 옵션을 해야 된다.

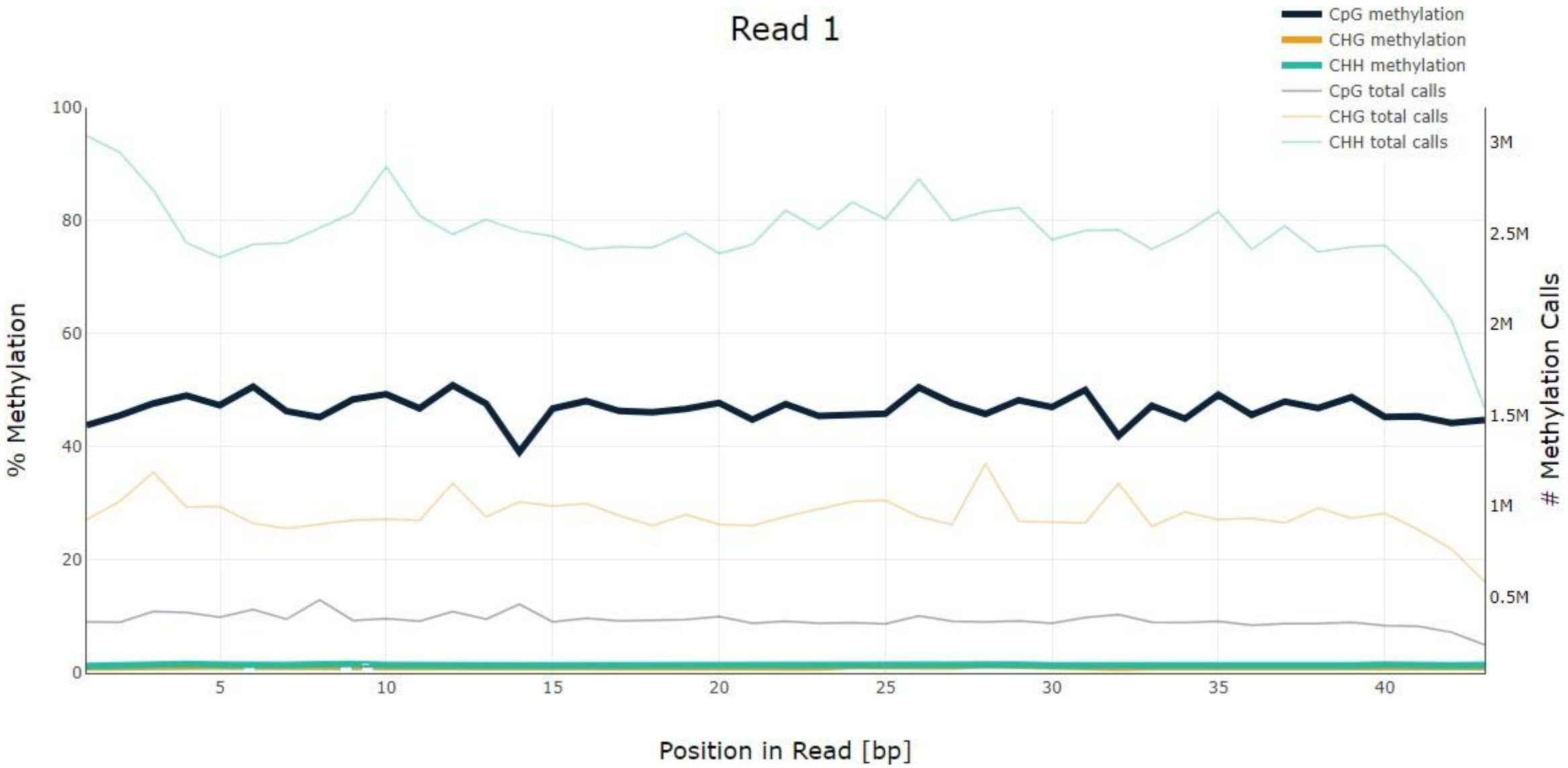
```
(Genomic_data_science) pjw@DESKTOP-39ALUO6:~/PJW_Workspace/Homework/test_sample$ UmiBam -p --bam SRR5195663_R1.clock_UMI.R1_val_1_bismark_bt2_pe.bam --double_umi
Setting --umi as well
Processing paired-end Bismark output file(s) (SAM format):
SRR5195663_R1.clock_UMI.R1_val_1_bismark_bt2_pe.bam


If the input is a multiplexed sample with several alignments to a single position in the genome, only alignments with a unique UMI will be chosen

Checking file >>SRR5195663_R1.clock_UMI.R1_val_1_bismark_bt2_pe.bam<< for signs of file truncation...



Now testing Bismark result file SRR5195663_R1.clock_UMI.R1_val_1_bismark_bt2_pe.bam for positional sorting (which would be bad...)      ...passed!
Running in >>> UMI-mode <<< (no mismatches in UMI tolerated)
```

Step5 : bismark_methylation_extractor를 이용해 methylation call 진행

```
(Genomic_data_science) pjw@DESKTOP-39ALUO6:~/PJW_Workspace/Homework/test_sample$ bismark_methylation_extractor --gzip --bedGraph --buffer_size 10G --cytosine_report --genome_folder ./Genome/ -p --ignore 5 --ignore_r2 5 SRR5195663_R1.clock_UMI.R1_val_1_bismark_bt2_pe.UMI_deduplicated.bam

 *** Bismark methylation extractor version v0.23.1 ***

Setting option '--no_overlap' since this is (normally) the right thing to do for paired-end data


Setting core usage to single-threaded (default). Consider using --multicore <int> to speed up the extraction process.

Summarising Bismark methylation extractor parameters:
============================================================
Bismark paired-end SAM format specified (default)
Number of cores to be used: 1
First 5 bp will be disregarded when processing the methylation call string of Read 1
First 5 bp will be disregarded when processing the methylation call string of Read 2
Output will be written to the current directory ('/mnt/d/Lab_wsl/PJW_Study/Homework/test_sample')
```

```
Storing all covered cytosine positions for chromosome: 1
Writing cytosine report for chromosome 1 (stored 257727 different covered positions)
Writing cytosine report for chromosome 10 (stored 215332 different covered positions)
Writing cytosine report for chromosome 11 (stored 303367 different covered positions)
Writing cytosine report for chromosome 12 (stored 179795 different covered positions)
Writing cytosine report for chromosome 13 (stored 174048 different covered positions)
Writing cytosine report for chromosome 14 (stored 156121 different covered positions)
Writing cytosine report for chromosome 15 (stored 184138 different covered positions)
Writing cytosine report for chromosome 16 (stored 129362 different covered positions)
Writing cytosine report for chromosome 17 (stored 193924 different covered positions)
Writing cytosine report for chromosome 18 (stored 129108 different covered positions)
Writing cytosine report for chromosome 19 (stored 130503 different covered positions)
Writing cytosine report for chromosome 2 (stored 317496 different covered positions)
Writing cytosine report for chromosome 3 (stored 196823 different covered positions)
Writing cytosine report for chromosome 4 (stored 298165 different covered positions)
Writing cytosine report for chromosome 5 (stored 299238 different covered positions)
Writing cytosine report for chromosome 6 (stored 204965 different covered positions)
Writing cytosine report for chromosome 7 (stored 266562 different covered positions)
Writing cytosine report for chromosome 8 (stored 243416 different covered positions)
Writing cytosine report for chromosome 9 (stored 222571 different covered positions)
Writing cytosine report for chromosome X (stored 108145 different covered positions)
Writing cytosine report for last chromosome Y (stored 8737 different covered positions)
Finished writing out cytosine report for covered chromosomes (processed 21 chromosomes/scaffolds in total)

Now processing chromosomes that were not covered by any methylation calls in the coverage file...
All chromosomes in the genome were covered by at least some reads. coverage2cytosine processing complete.
```

M-bias plot results



M-bias plot results