Chapter: Bioinformatics Analysis of DNA Methylation

Pei-Yu Lin[1] [0000-0003-0255-5467], Kuan-Lin Chen[1], Guan-Jun Lin[1,2], Shiang-Chin Huang[1], and Pao-Yang Chen[1] [0000-0002-7402-3075]

[1] Institute of Plant and Microbial Biology, Academia Sinica, 115, Taipei, Taiwan

[2] Genome and Systems Biology Degree Program, Academia Sinica and National Taiwan University, 106, Taipei, Taiwan

paoyang@gate.sinica.edu.tw

Abstract.

DNA methylation is a crucial epigenetic modification that influences gene expression and plays a role in various biological processes. High-throughput sequencing techniques, such as bisulfite sequencing (BS-seq) and enzymatic methyl sequencing (EM-seq), are widely used to investigate DNA methylation patterns at a genome-wide level. In this chapter, we present a bioinformatics pipeline for analyzing genome-wide DNA methylation. We outline the step-by-step process of the essential analyses, including aligning the converted reads, DNA methylation calling, differential methylation region (DMR) identification, data visualization, and post-alignment analyses. To illustrate the application of BS-seq and EM-seq, we demonstrated a case study on analyzing Arabidopsis *met1* methylome. This shows that genetic alteration of the DNA methyltransferase MET1 leads to disrupted DNA methylation patterns at CG sites, influencing various aspects of plant development and gene regulation. Overall, our pipeline

for methylome analysis can be applied to investigate the DNA methylation patterns of any genome, facilitating the identification of specific methylation profiles and their potential regulatory implications.

Keywords: DNA methylation, BS-seq, EM-seq, NGS, DMR, bioinformatic

1 Introduction

Epigenetics refers to alterations in gene expression that do not involve any change in the underlying DNA sequences. Such modifications can be inherited and are often reversible [1]. Among all epigenetic factors, DNA methylation is the most studied epigenetic regulator; it refers to the mechanism by which a methyl group is transferred to the C5 position of cytosine to form 5-methylcytosine (5mC) via DNA methyltransferases (DNMTs). DNA methylation occurs in the contexts of symmetric CG and CHG as well as asymmetric CHH sites, where H represents A, C, or T.

DNA methylation can silence genes or transposable elements by changing the chromatin structure or interfering with transcription factor binding [2] to regulate several biological processes. Due to the importance of DNA methylation in biological processes, several experimental approaches have been developed to profile genome-wide DNA methylation. The most popular method is next-generation sequencing (NGS), for example, reduced-representation bisulfite sequencing (RRBS) [3], whole-genome bisulfite sequencing (WGBS) [4], and enzymatic methyl sequencing (EM-seq) [5]. These NGS-based approaches can determine the methylation status of DNA sequences at single-base resolution and measure DNA methylation levels digitally. In

bisulfite sequencing (RRBS, BS-seq or WGBS), bisulfite conversion is the key step during sodium bisulfite chemical conversion of an unmethylated C into uracil (U) and eventual conversion into thymine (T) in subsequent PCR, while 5mC remains unchanged (Fig. 1a). Such treatment can result in approximately 84-96% DNA degradation, causing the loss of DNA material and induction of sequence bias, therefore affecting the accuracy of the analyses [6]. To improve from the bisulfite treatment in BS-seq, EM-seq is performed to reduce DNA damage and produce higher-quality libraries for detecting 5mC from approximately 400-fold smaller amounts of DNA. It uses two sets of enzymatic reactions, methylcytosine dioxygenase 2 (TET2) and T4-phage beta-glucosyltransferase (T4-BGT), to convert 5mC and 5hmC into products that cannot be deaminated by apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A). Then, APOBEC3A deaminates unmodified C to generate U, which is eventually converted into T during PCR (Fig. 1a) before the final library is sequenced. Compared to BS-seq, EM-seq offers a higher yield and better genome coverage with fewer PCR cycles required [7]. Unlike bisulfite libraries, EM-seq libraries do not exhibit biased AT-rich, GC-poor sequence representation since the absence of bisulfite treatment-induced DNA damage [5]. Moreover, low-input EM-seq libraries provide similar results to high-input libraries; for instance, a 0.5-ng input of EM-seq covers more CpGs than the 200 ng input used in BS-seq, highlighting the higher sensitivity of EM-seq [5].

Profiling genome-wide DNA methylation can be computationally intensive [8]. The general workflow for such bioinformatics analysis usually includes assessment of read quality, removal of duplicated reads, alignment of reads, quantification of DNA

methylation levels, identification of differentially methylated regions (DMRs), visualization of the methylome, and other post-alignment analyses (Fig. 1b).

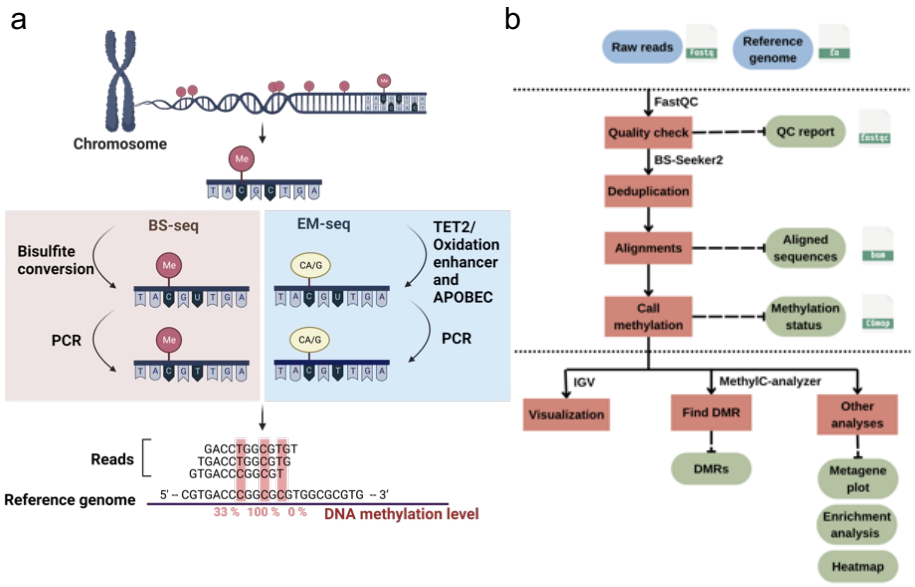

Fig. 1 DNA methylation and bioinformatics pipelines overview. (a) The library construction of the EM-seq and BS-seq. The panel was created using BioRender (http://biorender.com/). (b) The workflow of bioinformatics pipelines for the DNA methylation analysis. The blue color represents the input data, while the green color is the output. Red boxes are the steps for analysis, and the suggested tools are listed above the box.

Read alignment

Aligning reads to the reference genome is a critical first step in identifying methylated DNA sites from DNA methylation sequencing data. It can be carried out by commonly used bisulfite-read aligners with two types of algorithms: wild-card aligners [9] and three-letter aligners [10]. Wild-card aligners, such as BSMAP [11], replace Cs in the reference genome with the wild-card letter Y, which can match both Cs and Ts in the bisulfite-converted reads. This method offers higher genomic coverage, but it can introduce a bias towards higher methylation levels [10]. On the other hand, three-letter aligners, such as Bismark [12] and BS-Seeker2 [13], have higher mapping accuracy but lower coverage compared to the wild-card aligners [10], as they convert all Cs into Ts in the reads for both strands of the sequence. Bismark is more accurate than BSMAP but its mapping rate and accuracy may decrease with high read error rates in longer reads [14]. BS-Seeker2 is more capable of mapping reads from problematic libraries [13] and it is only slightly affected by read error rates [14]. Overall, among these tools, BSMAP offers the fastest alignment speed and minimal memory usage, while BS-Seeker2 provides the highest mapping accuracy [15]. Additionally, there is BS-Seeker3 [16] which is developed to improve BS-Seeker2, providing higher accuracy and mappability with a shorter processing time. The aligners output the alignments as BAM [17] or SAM flies [18] and the methylation calling information of each C base with sequence context information as CGmap files [19].

Cytosine methylation level information from CGmap files can be utilized for identifying DMRs. It refers to genomic regions with significant differences in DNA methylation levels between two groups of methylomes (e.g., experimental and control). The

genomic locations of DMRs may be further linked to specific biological meaningful features, such as promoters, genes, CpG islands, or other user-defined regions [20, 21].

Differential methylation region identification

Several tools have been developed for DMR detection, including HOME [22], MethylC-analyzer [23], and Bicycle [24] (Table 1). These tools can be divided into three different approaches: machine learning-based, statistical-based, and model-based methods. By implementing the machine learning algorithm, HOME utilizes a trained support vector machine (SVM) model to score each cytosine by specific features computed by weighted logistic regression using methylation level differences and p values between two groups. The tool groups cytosines into DMRs based on scores and distances to their neighboring cytosines [22]. The prebuilt SVM model in HOME has primarily been designed for analyzing mammalian (mainly human) DNA methylation data and therefore incorporates assumptions that may not account for the unique genetic regulation in nonmammalian species [25]. DMRs found by HOME are predicted by a precise delineation of the boundaries, and the lengths of the DMRs can vary widely. Statistical DMR identification tools, such as MethylC-analyzer, identify DMRs by comparing the average methylation levels of the genomic regions between the two groups. It offers users a choice between three statistical methods, the Student's t-test, the Kolmogorov–Smirnov test, and the Mann–Whitney U test, for detecting DMRs with significant differences (p value < 0.05) [23]. These statistical tests may have limitations due to certain assumptions they require; for example, the Stu-

dent's t-test assumes data to be approximately normally distributed [26], and the credibility of research findings may be affected when sample sizes are small [27]. Users can customize the length of DMRs using the MethylC-analyzer, which will be consistent within each genomic region. As a model-based DMR finding tool, Bicycle compares methylation levels of user-defined regions between two groups and identifies DMRs using the likelihood ratio test based on beta-binomial models with considerations for sensitivity and specificity [24]. Using beta-binomial models was claimed to decrease the false-positive rate in DMR identification. The tool selection can be based on data type and analysis requirements, as different tools employ different approaches to defined DMRs with diverse lengths and characteristics.

Data visualization

After the reads are aligned, the methylome data can be visualized by Integrative Genomic Viewer (IGV) [28] or the UCSC Genome Browser [29]. Users can customize the tracks on both the IGV and UCSC Genome Browser for a better understanding of the global DNA methylation pattern and compare it with other genome features ranging from single-nucleotide to megabase scales. IGV is a user-friendly desktop application that allows users to visualize methylation sites on the genome easily by importing files such as wiggle (WIG) files [29], which are commonly used for plotting quantitative genomic data such as methylation levels at cytosines. With IGV, we can directly view the methylation levels of identified DMRs and explore the adjacent genomic region that may be the potential regulatory targets of identified DMRs.

Post-alignment analyses

Post-alignment analyses aim to associate genomic regions with identified DMRs and explore the roles of these DMRs in genomic regulatory mechanisms where various toolkits can be applied to such analyses. The R package methylKit [30] can identify DMR proportions in various genetic elements, such as promoters, exons, or enhancers. MethGO [31] provides several modules for analyzing the correlation between methylation level and genomic features, including transcription factor-binding sites (TFBSs). MethylC-analyzer [32] provides an easy-to-use pipeline following the DMR identification step and includes several common analyses, such as enrichment analysis and metagene analysis. Enrichment analysis can assess the preferential localization of DMRs within genomic features across the genome, and metagene analysis is able to show the distribution of methylation levels along the gene body and adjacent regions.

2 Materials

2.1 Software

- SRA Toolkit 3.0.5 (https://github.com/ncbi/sra-tools) [33]
- bowtie2 v2.26 (https://bowtie-bio.sourceforge.net/bowtie2/index.shtml) [34]
- BS-Seeker2 v2.0.8 (https://github.com/BSSeeker/BSseeker2) [13]
- HOME v1.0.0 (https://github.com/ListerLab/HOME) [22]
- MethylC-analyzer (https://github.com/RitataLU/MethylC-analyzer) [23]
- bicycle v1.8.2 (http://www.sing-group.org/bicycle) [24]
- IGV Desktop v2.16.0 (https://igv.org/) [28]

2.2 Genome-wide DNA methylation Dataset

To demonstrate the methylation analysis pipeline, we downloaded and processed *Arabidopsis thaliana* (GSE122394) BS-seq datasets [35], including wild-type (wt) strains as controls and *met1* mutant strains in which DNA methyltransferase 1 (MET1) functions primarily to maintain CG methylation [36]. Each group contained three biological replicates.

The data for project GSE122394 are available on Gene Expression Omnibus (GEO) and can be accessed using the provided accession codes. The raw reads for each sample are stored in the Sequence Read Archive (SRA) listed in the GEO. To obtain the data, you can use SRAToolkit [33] to download the file using `prefetch` and then convert it into the FASTQ format (.fastq) for analysis by `fast-dump`.

```
$ prefetch SRR8180314 ## download SRA data
$ fast-dump SRR8180314 ## transfer into fastq file
```

3 Methods

3.1 Processing methylomes

To provide useful guidance, a bioinformatics pipeline is introduced below, and the tools used in the protocol are listed in the materials section. In the following demonstration, BS-Seeker2 is used.

3.1.1 Alignments of methyl-seq read

1. Use bowtie2 to create a reference genome index file (*see* **Note 1**). for the *Arabidopsis thaliana* TAIR10 version in the aligner and save it as `BS2_bt2_Index`.

```
$ bs_seeker2-build.py -f genome.fa --aligner=bowtie2 -d ./BS2_bt2_Index
```

2. Align raw reads of wild-type replicate 1 to the reference genome using the align function and save it as a BAM file named `wt_r1_align.bam`. The input data is suggested to undergo the quality check before start analyzing (*see* **Note 2**).

```
$ bs_seeker2-align.py -i wt_r1.fastq -g genome.fa  --aligner=bowtie2 -o
wt_r1_align.bam
```

3.1.2 Call methylation

1. Use call methylation script to calculate the methylation level.

```
$ bs_seeker2-call_methylation.py -i wt_r1_align.bam -o wt_r1.CGmap -d
/BS2_bt2_Index/genome.fa_bowtie2
```

2. View the methylation call output (CGmap). The file with each row represents a single CpG site.

```
$ zless wt_r1.CGmap.gz
```

Each CpG site contains the following information: chromosome, nucleotide on Watson strand, position, context, dinucleotide context, methylation level, number of methylated cytosines (#C), and the total number of all cytosines (#C+T) (Fig. 2.)

Chromosome	Nucleotide on Watson (+) strand	Position	Context	Dinucleotide context	Methylation level	Methylated cytosines	All cytosines
1	C	917	CHH	CA	0.0	0	1
1	C	919	CHH	CA	0.0	0	1
1	C	927	CHG	CT	1.0	1	1
1	G	929	CHG	CA	0.0	0	3
1	C	935	CHH	CA	0.0	0	1
1	C	938	CHH	CT	0.0	0	1
1	C	940	CHH	CA	0.0	0	1
1	C	946	CHH	CC	0.0	0	1
1	C	947	CHG	CC	0.0	0	2
1	C	948	CG	CG	0.5	1	2

Fig. 2 Example of CGmap file generated with the BS-seeker2 call methylation script. The figure displays a snapshot of the ten rows from the `wt_r1.CGmap.gz` file.

3.1.3 Conversion rate

In regard to methyl-seq (EM-seq and BS-seq) analysis, the estimation conversion rate [37], which measures how effectively bisulfite or enzyme treatment can convert unmethylated cytosines to uracil in DNA samples, is required for evaluation. By comparing the unmethylated bacteriophage lambda genome as a reference to our bisulfite/enzyme treatment genomes, the percentage of successfully converted cytosines can be estimated. It is simply calculated by dividing the number of converted cytosines ($\#T$) by the total number of cytosines ($\#T + C$) and multiplying by 100. Typically, a conversion rate of 95% or above is preferred because it shows more reliable and accurate results [38].

1. The first step for the conversion rate is the same as above but changes the input reference genome to the lambda genome.

```
  $ bs_seeker2-build.py -f lambda_genome.fa --aligner=bowtie2 -
d ./BS2_lambda_Index

  $ bs_seeker2-align.py -i wt_r1_rmdup.fastq -g lambda_genome.fa  --
aligner=bowtie2 -o wt_r1_lambda.bam -m 3 -d BS2_lambda_Index

  $ bs_seeker2-call_methylation.py -i wt_r1_lambda.bam -o wt_r1_lambda -
d BS2_bt2_Index/genome.fa_bowtie2/
```

2. The conversion rate is calculated by the R script (*see* **Note 3**) with the formula:

$$Conversion\ rate = \frac{\#T}{\#T+\#C} \times 100\ .$$

```
  $ Rscript coversion_rate.R  wt_r1_lambda.CGmap.gz
  [ 03:13:46 AM ] Calculating bisulfite conversion rate
  [ 03:13:46 AM ] Bisulfite conversion rate: 97.01493 %
```

In our example, the conversion rate for the wt_r1 methylome is 97.01%, which means that 97.01% of the unmethylated cytosines in the DNA sample have been successfully converted to uracil.

3.2 DMR identification

Here, MethylC-analyzer is selected to demonstrate how to find DMRs from the aligned methylation data output. To prevent environmental conflicts, the docker image provided by the software is utilized (*see* **Note 4**).

3.2.1 Searching DMR

1. The command `DMR` is used along with the input `samples_list.txt` file that

listed all CGmap file names (`wt_r1.CGmap.gz`, `wt_r2.CGmap.gz`, `wt_r3.CGmap.gz`, `met1_r1.CGmap.gz`, `met1_r2.CGmap.gz`, and `met1_r3.CGmap.gz` in our case) and the description of each input sample (wt and *met1* in our case), and a `gene.gtf` file. GTF files can be downloaded from UCSC (https://hgdownload.soe.ucsc.edu/downloads.html); and is a file format containing information about the genomic features of genes, such as exons, introns, coding sequences, and untranslated regions (UTRs) [39]. The default minimum depth for CpG sites and the number of sites within a region are both set to four. The default size of the DMR is 500 base pairs (bp). The default p value cutoff for Student's t-test for identifying DMRs is 0.05. These arguments can be adjusted by users.

```
$ docker run --rm -v $(pwd):/app peiyulin/methylc:V1.0 python /MethylC-analyzer/scripts/MethylC.py DMR samples_list.txt gene.gtf /app/ -a met1 -b wt
```

The output consists of all, hyper, and hypo DMRs as text files. Here, we found 3,282 DMRs in CG methylation between the wt and *met1* groups.

3.2.2 Analyzing DMRs

In the comparison of different DMR identifiers, we only discussed the difference between HOME and MethylC-analyzer since Bicycle requires its specific file format from its own pipeline. For a fair comparison, the regions of DMR require at least four Cs when applying both tools.

MethylC-analyzer discovered 3,282 DMRs in fixed regions of 500 bp, which were subsequently merged into 2,785 DMRs by combining contiguous DMRs (size range from 500 to 14,500 bp). HOME identified 16,185 DMRs in regions of varying lengths, with the longest being 36,721 bp and the shortest being 50 bp. HOME identified more DMRs and covered 94.5% of the DMRs found by MethylC-analyzer (Fig. 3a). Moreover, it can be observed that the DMRs identified by HOME are much wider and span a large region, even extending across multiple genes (Fig. 3b red box) and those small size DMRs tend to spread out in the intergenic regions (IGR) (Fig. 3b yellow boxes). To sum up, HOME identifies more DMRs than MethylC-analyzer, while HOME is more sensitive to the changes between two groups, and MethylC-analyzer may be more precise by pinpointing the smaller regions.

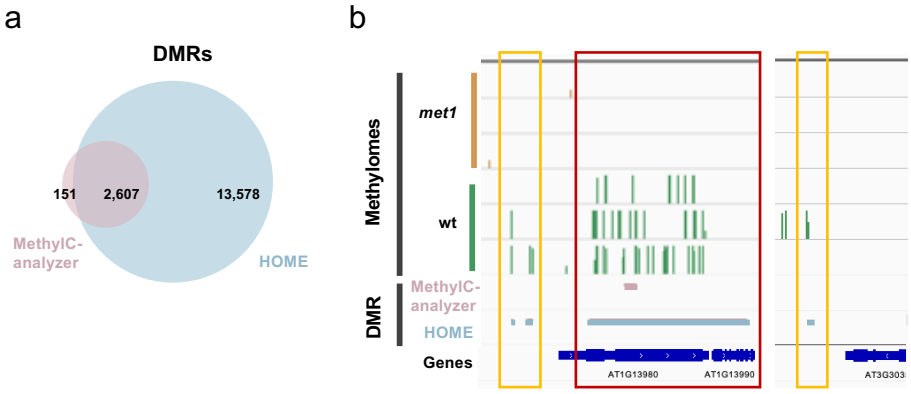

Fig. 3 DMRs found by MethylC-analyzer and HOME. (a) The Venn diagram shows the number of overlapping DMRs between HOME and MethylC-analyzer. The criteria for DMR identification were a minimum of 4 cytosines within a DMR, delta meth-

ylation level cutoff = 0.1 and p value < 0.05. (b) Comparison of identified DMRs between HOME and MethylC-analyzer in IGV. The cross genes DMR is highlight in red and the intergenic DMRs are in yellow.

3.3 Data visualization

3.3.1 Genome browser

1. Download and activate the IGV Desktop application according to the operating system (*see* **Note 5**). This application supports operating systems including MacOS, Windows, and Linux.

2. Select the reference genome from the dropdown list. Here, we chose *A. thaliana* (TAIR10) as a reference genome. Additional reference genomes can be downloaded by clicking *More* or can be loaded from the local path (in FASTA format).
3. Convert the file from the WIG file to the suggested track formats, BigWig [40] or TDF files, by running IGVtools (Click *Tools>Run IGVtools*).

4. Select *File>Load from File* to load data into the track panel. Right-click the panel to adjust the graphic type or other settings.

5. Use the dropdown list and search box at the top panel to select the chromosome and region shown. Click +/- on the top panel to zoom in/out. Clicking or dragging on the track of the chromosome can also adjust the region shown.

6. Click *File>Save session* or *File>Save Image* to save the visualization result (Fig. 4a).

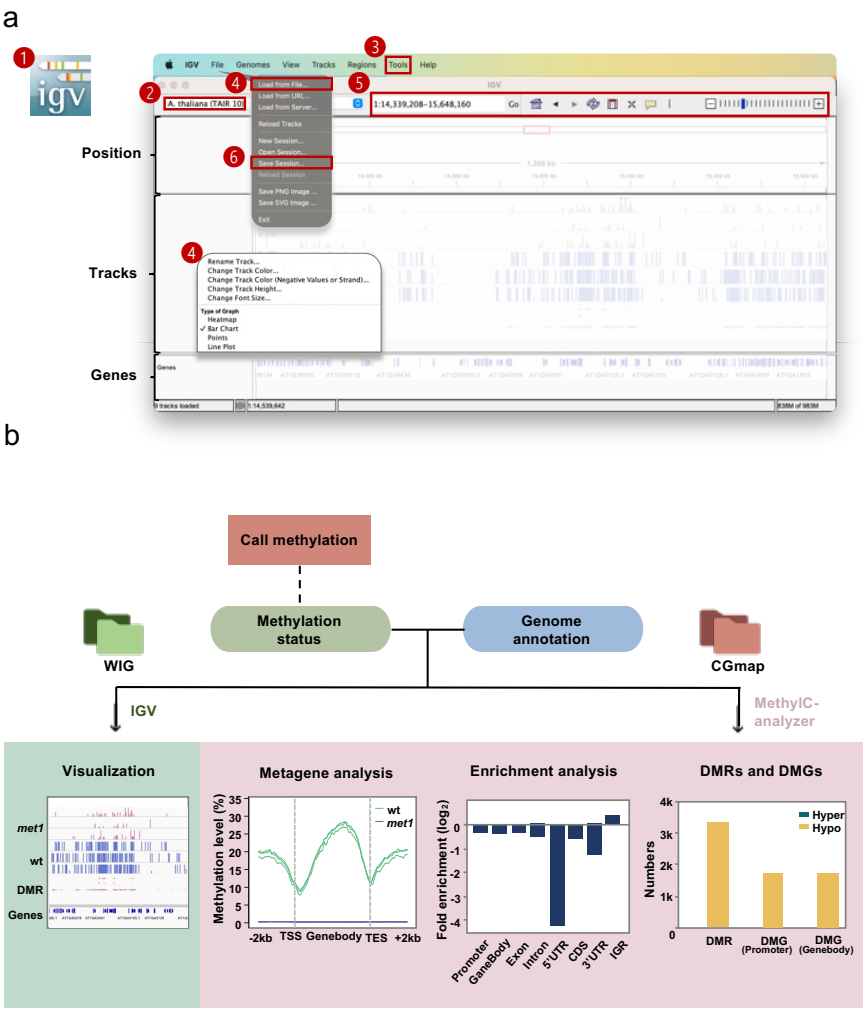

Fig. 4 Schematic for post-alignment analysis and visualization. (a) The interface of IGV Desktop on a Mac system. The steps and the operating areas are in red (see method 3.3). The main steps shown in the figure: ❶ open IGV; ❷ select the reference

genome; ❸ convert WIG to BigWig (Tools > Run IGVtools); ❹ load tracks (File > Load from File) and adjust tracks; ❺ select the region of interst; ❻ save session (File > Save session). (b) Overview of the post-alignment analyses. Analyses in the right panel (pink) are performed by MethylC-analyzer, which requires CGmap and genome annotation GTF file as input. Visualization by IGV is in the left panel (green). It allows the WIG file from aligners, as well as BigWig and BED files from MethylC-analyzer.

3.4 Post-alignment analyses

For a better interpretation of methylation data, post-alignment analyses like enrichment analysis or metagene analysis are commonly carried out for explaining the methylation profiles. Enrichment analysis calculates the fold change in genomic region enrichment in identified DMRs compared to the whole genome. Metagene analysis represents the average methylation level along the gene body and adjacent regions at normalized length. In this section, MethylC-analyzer is applied to perform enrichment and metagene analyses (*see* **Note 6**).

3.4.1 Enrichment analysis

1. Use the `Fold_Enrichment` command to generate the enrichment result.

This module generates output files, including `CG_Fold_Enrichment.pdf` and

```
$ docker run --rm -v $(pwd):/app peiyulin/methylc:V1.0 python /MethylC-analyzer/scripts/MethylC.py Fold_Enrichment samples_list.txt gene.gtf /app/ -a met1 -b wt
```

multiple BED files, such as `CommonRegion_CG.txt.bed`. The BED format provides the information like the positions of common methylated regions across samples. The BED file can be visualized by using IGV.

DMRs exhibit a positive fold enrichment value in the IGR, suggesting a higher likelihood of DMRs being located in IGRs (Fig. 4b).

3.4.2 Metagene analysis

1. Use the `Metaplot` command to generate the Metaplot result. This module generates two types of metagene plots: one represents the average methylation level in

```
$ docker run --rm -v $(pwd):/app peiyulin/methylc:V1.0 python /MethylC-analyzer/scripts/MethylC.py Metaplot samples_list.txt gene.gtf /app/ -a met1 -b mt
```

   two groups (metaplot_CG.pdf), and the other shows the difference between the two groups (metaplot_delta_CG.pdf). The former illustrates the methylation pattern along the gene body and adjacent region, while the latter directly represents the difference in distribution between wt and *met1*. This module also generates BigWig files (met1_r1_CG.bw) to record methylated C sites in metagene analysis, and these BigWig files can be visualized by IGV.

In our case, the wt samples exhibit a standard CG methylation pattern [41] with a lower methylation level at the transcription start site (TSS) and transcription end site (TES). The *met1* samples show a consistently low methylation level along the gene body, reflecting the dysfunction of the methyltransferase (Fig. 4b).

4 Note

1. The reference genome can be downloaded from iGenomes (https://support.illumina.com/sequencing/sequencing_software/igenome.html) [42], which offers a collection of reference sequences and annotation files for commonly studied organisms.
2. Before alignment, the methyl-seq reads should undergo quality control (QC) to remove low-quality reads and duplicate sequences generated by PCR amplification and adapter sequences. The suggested tool for QC is FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [43], and BS-Seeker2 also provides the command lines for removing duplicated reads.

```
# quality control
$ fastqc wt_r1.fastq

# remove duplicate
$ FilterReads.py -i wt_r1.fastq -o wt_r1_rmdup.fastq > FilterReads.log
```

3. The conversion rate script can be viewed or downloaded on the GitHub page: (https://github.com/beritlin/NGS_analyses/blob/main/DNA_Methylation_Analyses/coversion_rate.R) [44].
4. As different tools require specific environmental settings to run properly, using a docker image can prevent environmental conflict issues.
5. The UCSC Genome Browser provides web-based track hubs, which are convenient for users to quickly find and visualize public genome-wide datasets. Users

who are looking for more detailed genomic information on well-studied genomes (e.g., the human genome hg38) are recommended to use the UCSC Genome Browser for visualization.

6. MethylC-analyzer provides an all-in-one process to perform multiple analyses for the same dataset in one command to save running time. The command is shown below:

```
$ docker run --rm -v $(pwd):/app peiyulin/methylc:V1.0 python /MethylC-analyzer/scripts/MethylC.py all samples_list.txt gene.gtf /app/ -a met1 -b wt
```

Acknowledgments

This work was supported by grants from Academia Sinica and the Ministry of Science and Technology of Taiwan (111-2927-I-001-505- and 111-2311-B-001-030-), NTU-AS Innovative Joint Program (AS-NTU-112-12) and VGH-TSGH-AS Joint Research Program (VTA112-T-3-2) to P.-Y. C. We also extend acknowledgment to BioRender (http://biorender.com/) for the creation of images.

References

1. Richards CL, Bossdorf O, Verhoeven KJ (2010) Understanding natural epigenetic variation. New Phytol 187 (3):562-564. doi:10.1111/j.1469-8137.2010.03369.x
2. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13 (7):484-492. doi:10.1038/nrg3230
3. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res 33 (18):5868-5877. doi:10.1093/nar/gki901

4. Cokus SJ, Feng S, Zhang X et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452 (7184):215-219. doi:10.1038/nature06745
5. Vaisvila R, Ponnaluri VKC, Sun Z et al. (2021) Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Res 31 (7):1280-1289. doi:10.1101/gr.266551.120
6. Grunau C, Clark SJ, Rosenthal A (2001) Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Res 29 (13):E65-65. doi:10.1093/nar/29.13.e65
7. Feng S, Zhong Z, Wang M, Jacobsen SE (2020) Efficient and accurate determination of genome-wide DNA methylation patterns in Arabidopsis thaliana with enzymatic methyl sequencing. Epigenetics Chromatin 13 (1):42. doi:10.1186/s13072-020-00361-9
8. Yong WS, Hsu FM, Chen PY (2016) Profiling genome-wide DNA methylation. Epigenetics Chromatin 9:26. doi:10.1186/s13072-016-0075-3
9. Schbath S, Martin V, Zytnicki M, Fayolle J, Loux V, Gibrat J-F (2012) Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 19 (6):796-813. doi:10.1089/cmb.2012.0022
10. Bock C (2012) Analysing and interpreting DNA methylation data. Nature Reviews Genetics 13 (10):705-719. doi:10.1038/nrg3273
11. Xi Y, Li W (2009) BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10:232. doi:10.1186/1471-2105-10-232
12. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27 (11):1571-1572. doi:10.1093/bioinformatics/btr167
13. Guo W, Fiziev P, Yan W et al. (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics 14:774. doi:10.1186/1471-2164-14-774
14. Lee J-H, Park S-J, Kenta N (2015) An integrative approach for efficient analysis of whole genome bisulfite sequencing data. BMC Genomics 16 (12):S14. doi:10.1186/1471-2164-16-S12-S14
15. Gong W, Pan X, Xu D et al. (2022) Benchmarking DNA methylation analysis of 14 alignment algorithms for whole genome bisulfite sequencing in mammals. Computational and Structural Biotechnology Journal 20:4704-4716. doi:10.1016/j.csbj.2022.08.051
16. Huang KYY, Huang Y-J, Chen P-Y (2018) BS-Seeker3: ultrafast pipeline for bisulfite sequencing. BMC Bioinformatics 19 (1):111. doi:10.1186/s12859-018-2120-7
17. Genome Browser BAM Track Format.
18. Li H, Handsaker B, Wysoker A et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25 (16):2078-2079. doi:10.1093/bioinformatics/btp352

19. Guo W, Zhu P, Pellegrini M, Zhang MQ, Wang X, Ni Z (2018) CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. Bioinformatics 34 (3):381-387. doi:10.1093/bioinformatics/btx595
20. Piao Y, Xu W, Park KH, Ryu KH, Xiang R (2021) Comprehensive Evaluation of Differential Methylation Analysis Methods for Bisulfite Sequencing Data. Int J Environ Res Public Health 18 (15). doi:10.3390/ijerph18157975
21. Chen D-P, Lin Y-C, Fann CSJ (2016) Methods for identifying differentially methylated regions for sequence- and array-based data. Briefings in Functional Genomics 15 (6):485-490. doi:10.1093/bfgp/elw018
22. Srivastava A, Karpievitch YV, Eichten SR, Borevitz JO, Lister R (2019) HOME: a histogram based machine learning approach for effective identification of differentially methylated regions. BMC Bioinformatics 20 (1):253. doi:10.1186/s12859-019-2845-y
23. Lu RJ-H, Lin P-Y, Yen M-R, Wu B-H, Chen P-Y (2023) MethylC-analyzer: a comprehensive downstream pipeline for the analysis of genome-wide DNA methylation. Botanical Studies 64 (1):1. doi:10.1186/s40529-022-00366-5
24. Graña O, López-Fernández H, Fdez-Riverola F, González Pisano D, Glez-Peña D (2017) Bicycle: a bioinformatics pipeline to analyze bisulfite sequencing data. Bioinformatics 34 (8):1414-1415. doi:10.1093/bioinformatics/btx778
25. Huther P, Hagmann J, Nunn A et al. (2022) MethylScore, a pipeline for accurate and context-aware identification of differentially methylated regions from population-scale plant whole-genome bisulfite sequencing data. Quant Plant Biol 3:e19. doi:10.1017/qpb.2022.14
26. Massey Jr FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association 46 (253):68-78
27. De Winter JC (2013) Using the Student's t-test with extremely small sample sizes. Practical Assessment, Research, and Evaluation 18 (1):10
28. Robinson JT, Thorvaldsdóttir H, Winckler W et al. (2011) Integrative genomics viewer. Nature Biotechnology 29 (1):24-26. doi:10.1038/nbt.1754
29. Kent WJ, Sugnet CW, Furey TS et al. (2002) The human genome browser at UCSC. Genome Res 12 (6):996-1006. doi:10.1101/gr.229102
30. Akalin A, Kormaksson M, Li S et al. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biology 13 (10):R87. doi:10.1186/gb-2012-13-10-r87
31. Liao W-W, Yen M-R, Ju E, Hsu F-M, Lam L, Chen P-Y (2015) MethGo: a comprehensive tool for analyzing whole-genome bisulfite sequencing data. BMC Genomics 16 (12):S11. doi:10.1186/1471-2164-16-S12-S11
32. Lu RJ, Lin PY, Yen MR, Wu BH, Chen PY (2023) MethylC-analyzer: a comprehensive downstream pipeline for the analysis of genome-wide DNA methylation. Bot Stud 64 (1):1. doi:10.1186/s40529-022-00366-5
33. Team STD (2023) SRA Toolkit 3.0.5. GitHub. https://github.com/ncbi/sra-tools.

34. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature methods 9 (4):357-359. doi:10.1038/nmeth.1923
35. Choi J, Lyons DB, Kim MY, Moore JD, Zilberman D (2020) DNA Methylation and Histone H1 Jointly Repress Transposable Elements and Aberrant Intragenic Transcripts. Mol Cell 77 (2):310-323 e317. doi:10.1016/j.molcel.2019.10.011
36. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11 (3):204-220. doi:10.1038/nrg2719
37. Hong SR, Shin KJ (2021) Bisulfite-Converted DNA Quantity Evaluation: A Multiplex Quantitative Real-Time PCR System for Evaluation of Bisulfite Conversion. Front Genet 12:618955. doi:10.3389/fgene.2021.618955
38. Bock C, Reither S, Mikeska T, Paulsen M, Walter J, Lengauer T (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. Bioinformatics 21 (21):4067-4068. doi:10.1093/bioinformatics/bti652
39. Kent WJ, Sugnet CW, Furey TS et al. (2002) The human genome browser at UCSC. Genome research 12 (6):996-1006
40. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26 (17):2204-2207. doi:10.1093/bioinformatics/btq351
41. Feng S, Zhong Z, Wang M, Jacobsen SE (2020) Efficient and accurate determination of genome-wide DNA methylation patterns in Arabidopsis thaliana with enzymatic methyl sequencing. Epigenetics & Chromatin 13 (1):42. doi:10.1186/s13072-020-00361-9
42. Tabata S, Kaneko T, Nakamura Y et al. (2000) Sequence and analysis of chromosome 5 of the plant Arabidopsis thaliana. Nature 408 (6814):823-826. doi:10.1038/35048507
43. Simon A FastQC A Quality Control tool for High Throughput Sequence Data.
44. Lin P-Y (2023) NGS_analyses. GitHub. https://github.com/beritlin/NGS_analyses/blob/main/DNA_Methylation_Analyses/coversion_rate.R.

Table

Table 1. Comparison of three DMR tools.

Features	HOME	MethylC-analyzer	Bicycle
Version	1.0.0	-	1.8.2
Language	Python, R	Python, R	java
Environment	CLI/	CLI/Docker	CLI
Available context	CG, CHG, CHH	CG, CHG, CHH	CG, CHG, CHH
Testing method	Weighted logistic regression, support vector machine	Student's t-test, Kolmogorov–Smirnov test, Mann–Whitney U test	Likelihood ratio of beta-binomial models
User-defined DMR length	not available	available	available

CLI: command line interface, GUI: graphical user interface