Chapter 32

Bioinformatics Analysis of DNA Methylation Through Bisulfite Sequencing Data

Fei Sang

Abstract

DNA methylation plays an important role in the regulation of gene expression as one of the epigenetic modifications. The bisulfite sequencing is widely used to determine the patterns of genomic methylation as a gold standard technology allowing conversion of the unmethylated cytosines to uracils that are represented as Ts in the sequencing reads. This chapter introduces the methodology for analyzing bisulfite sequencing data using various bioinformatics tools.

Key words DNA Methylation, Bisulfite Sequencing, Bioinformatics

1 Introduction

Epigenetic regulation of gene expression has a vital role in development and differentiation [1]. One of the major epigenetic modifications, DNA methylation (5-methylcytosine, 5mC) usually occurs at the cytosines of CpG dinucleotides in mammals. Usually, 5mC in the promoter regions is associated with the suppression of gene expression via recruitment of a series of binding proteins and histone modifiers [2]. DNA methylation in other regions including gene bodies, intergenic regions and CpG islands can also affect gene expression via different underlying mechanisms [3–5]. Bisulfite sequencing provides a way to study the DNA methylation landscape across the whole genome at the single base resolution [6]. This method relies on conversion of the unmethylated cytosines (Cs) to uracils, while the methylated cytosines remain unchanged. As the converted uracils are turned into thymines (Ts) during PCR amplification, all the Cs in the corresponding sequencing reads represent 5mC in the sample DNA. Normally, the bisulfite sequencing requires amplification of the sample DNA by PCR. However, PCR may induce bias during the sequencing, since amplification efficiency of methylated and unmethylated molecules

Alexey Ruzov and Martin Gering (eds.), *DNA Modifications: Methods and Protocols*, Methods in Molecular Biology, vol. 2198, https://doi.org/10.1007/978-1-0716-0876-0_32,

are different, and this bias often appears strand-specific [7]. In order to avoid this bias in the analysis, the optimal PCR conditions need to be determined by selecting specific primers, and deduplication is necessary in the downstream analysis. Analyzing bisulfite sequencing data can be simply summarized as the detection of single-nucleotide variants (SNPs) occurred due to the conversion of unmethylated cytosines.

In this chapter, we focus on describing the bioinformatics strategies for analysis of the bisulfite sequencing data. The whole methodology can be divided into the following steps (Fig. 1):

- Quality control of sequencing reads.
- Read trimming and filtering.
- Read alignment onto the reference genome.
- Methylation calling.
- Methylation annotation.
- Differential methylation.

2 Bioinformatics Tools

2.1 Quality Control

The quality of raw reads needs to be determined first, to ensure the experimental setup and sequencing are successful. This can be done by *FastQC* [8] or *fastp* [9]. Both these tools can create a list of reports including those on read quality, GC content, the distribution of read length, duplication level, adaptor contamination, and so on.

2.2 Read Trimming

Several different tools are available for read trimming based on both the adaptor sequences and sequence quality. *Trimmomatic* [10] can trim both paired- and single-end reads using both these parameters. *Skewer* [11] is a fast trimming tool for the NGS short reads, especially for the Illumina paired-end reads, which can conduct the competitive trimming accuracy and can trim adaptor sequences in both 5′ and 3′ of PE reads. *Trim Galore* [12] is not only a quality and adaptor trimming tool but also has special functions to remove the bias of the methylated positions in reads obtained by Reduced Representation Bisulfite Sequencing (RRBS).

2.3 Alignment and Methylation Calls

Bismark [13] is a software widely used for the analysis of BS-Seq data, including aligning reads, removing duplication and detecting cytosine methylation. *Bismark* can use *Bowtie2* [14] or *HISAT2* [15] for the alignment of short reads. *Picard* [16] tools and *samtools* [17] are required by *Bismark* to remove duplicates and manipulate bam files.

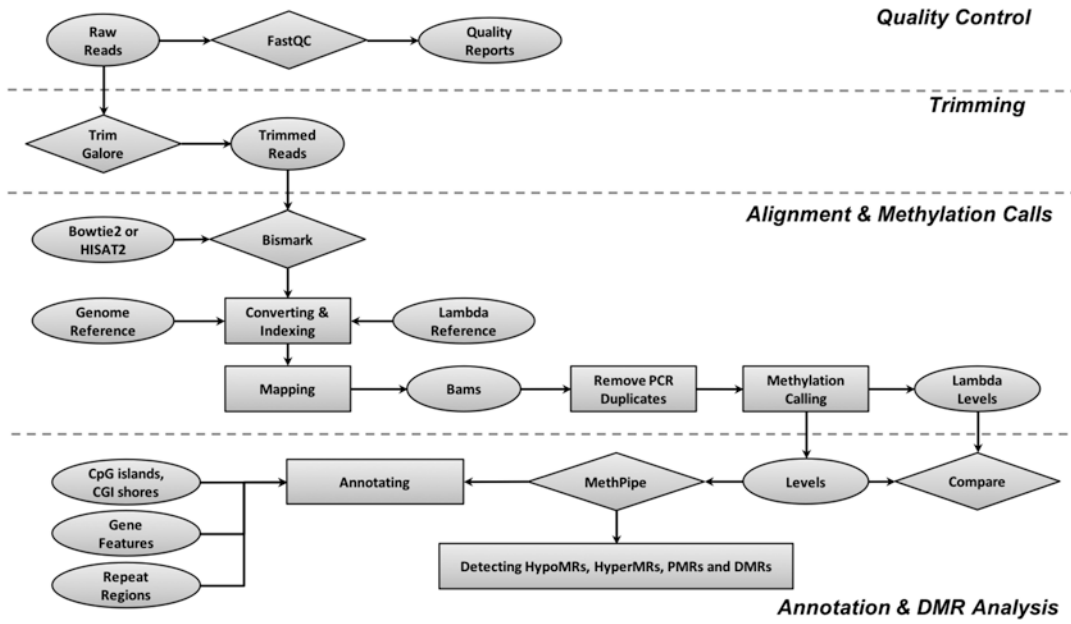

Fig. 1 Pipeline for the analysis of bisulfite sequencing data. The steps include quality control, trimming, alignment, methylation calls, annotation and downstream analysis

2.4 Methylation Annotation

MethPipe [18] provides a set of tools to process the BS-Seq data including the alignment, removing duplicates, detecting methylated cytosines, and other downstream analysis.

2.5 Differential Methylation

A number of tools can be used to find the differential methylated regions. *MethPipe* is recommended for the analysis.

3 Methods

In this section we describe the methods of analyzing BS-Seq data step-by-step. All the commands in the following subsections require the Linux environment.

3.1 Quality Control

The first step, before the downstream analysis, is to examine the sequencing quality and potential contaminations occurred during the experiment, to avoid mistakes and reduce biases. *FastQC* can take the compressed fastq files of raw reads as the inputs and can be run in parallel for the quality check. The simple *FastQC* command to analyze multiple fastq files is:

```
fastqc-t 12 pair1.fastq.gz pair2.fastq.gz
```

FastQC also provides some useful options to manipulate various outputs. Moreover, the user-friendly graphic interface has been

developed for *FastQC* to facilitate the processing step without command lines. The outputs return an HTML summary report file and a compressed zip file containing the figures for each fastq file. *FastQC* reports have 10 different parts including those on base quality, duplication level, adapter contamination, and overrepresented sequences.

Basic Statistics report provides a brief summary of BS-Seq reads. *Per base sequence quality* report shows the Phred quality score of each base along the read length. A score more than 30 can be considered as a good one, which means the error rate of this position is less than 0.001. Normally a decreasing trend of the quality score can be observed toward the 3′ end; therefore, the trimming step is required to polish the reads. *Per sequence quality scores* report summarizes the average Phred qualities of the reads. *Per base sequence content* report describes the base component of each position across the reads and, thus, can indicate if the bisulfite conversion is successful. The proportions of each bases are nonbias and similar in an ideal random library. The Cs have been converted to Ts in the bisulfite library, therefore, their proportion must be lower. The unsuccessful bisulfite conversion is observed if the nonbias proportions appear in this report. *Per sequence GC content* report shows the distribution of GC content across sequencing reads compared with a normal GC distribution. BS-Seq reads usually show the abnormal GC content in this part due to conversion of Cs to Ts. *Per base N content* report refers to the proportion of unidentified bases represented by N. Usually the level of Ns should be very low, nearly zero, if the sequencing is of good quality. *Sequence Length Distribution* report calculates the distribution of read lengths. *Sequence Duplication Levels* report should be considered as a warning, as PCR duplicates often appear in the BS-Seq with a high proportion, albeit can be removed during the downstream analysis. *Overrepresented sequences* report shows the overrepresented sequences in the reads. *Adapter Content* report identifies the potential adapter sequences in the reads.

3.2 Read Trimming

Trimming is a necessary step of processing the raw reads before the alignment, because, inevitably, some adaptor sequences are sequenced along with raw reads, resulting in the low alignment rate and false C–T conversions.

Trim Galore is not only a quality and adaptor trimming tool but also has special functions to remove the bias of the methylated positions in RRBS reads. In addition, *Trim Galore* can autodetect the type of adapter sequences. There are several important parameters that can be specified during the trimming in *Trim Galore*. To set the trimming cutoff of read quality *-q*; to set the quality type *--phred33/−−phred64*; to set the mode of paired-end library *--paired*; to set the mode of RRBS library *--rrbs*; to set the length cutoff of discarding short reads *--length*; to remove the Ns

at the both ends of reads *--trim-n*. Below is an example of running *Trim Galore* to trim the paired-end reads from the 3′ end with a quality cutoff 30 using phred33 score. Before running this command, *cutadapt* needs to be installed first as a prerequisite of *Trim Galore*. *-j* option is set for speeding up *cutadapt* to run in parallel.

```
trim_galore -j 12 --paired --phred33 -q 30 pair1.fq.gz
pair2.fq.gz
```

Two output files can be created after trimming. One is the trimmed fastq file, and the other one is the summary report containing the overview of trimmed reads.

3.3 Alignment and Methylation Calls

Correct determination of the methylation status of different genomic sequences depends on the proper read alignment against the reference genome. BS-Seq converts Cs to Ts, therefore, we use *Bismark*, which has been widely utilized as an effective aligner for BS-Seq reads. It consists of a series of comprehensive tools that align reads with *Bowtie2* or *HISAT2*, remove PCR duplicates, detect the genomic methylated positions, generate the genomic coverages of methylated bases, and so on.

First, *Bismark* needs to be used to build the indexes and conversions of genomic sequences before the initial alignment, to convert Cs to Ts and Guanines (Gs) to Adenine (As) in the reference. Then the converted reference is indexed by the aligner *Bowtie2* or *HISAT2*. Here *Bowtie2* was used as an example:

```
bismark_genome_preparation --path_to_aligner bowtie2 --ver-
bose reference_folder
```

Then *Bismark* can use *Bowtie2* to map the trimmed reads and specify the additional parameters of *Bowtie2* to improve the alignment. To decide on a valid pair of a paired-end read, the proper insert size needs to be configured by *–I,* the minimum insert size, and *–X,* the maximum insert size. Furthermore, it is important to choose the correct library option depending on the sequencing library, *––non_directional* or *--pbat*. *-1* and *– 2* represent the forward and reverse reads for the paired-end reads. *Bismark* also provides parallel options to speed up the alignment (*see* **Note 1**). To improve the mapping efficiency, the minimum alignment score needs to be specified carefully. The default function of the score is $f(x) = 0 + (-0.2 \times x)$ represented by *--score_min* L, 0, −0.2. The lower minimum score can allow more reads to be mapped to the reference, but it may also increase the rate of false positive alignments (*see* **Note 2**).

```
bismark --bowtie2 --genome reference_folder --pbat --score_
min L,0,-0.4 -I 0 -X 1000 \
```

```
-p 4 --parallel 12 -o output -1 pair1.trimmed.fq.gz -2
pair2.trimmed.fq.gz
```

The next step is to remove the potential PCR duplicates, which is recommended in the analysis of the whole-genome bisulfite sequencing but not in RRBS, amplicon- or other target enrichment-type libraries. *Bismark* examines the alignments, the chromosomes, the strands, the start positions of paired-end reads for the duplication and can take the UMIs (unique molecular identifiers) or barcodes into account for the deduplication.

```
deduplicate_bismark -p --output_dir output --bam align-
ment.bam
```

The detection of methylated cytosines is crucial for the whole analysis. *Bismark* can simultaneously detect the cytosines in CpG (5mC/C followed by G), CHG and CHH (H corresponds to A, T, or C/5mC) contexts, and also create the strand-specific outputs, OT (original top strand), CTOT (complementary to original top strand), OB (original bottom strand), CTOB (complementary to bottom strand), if the alignment running with *--directional* option. The genome-wide cytosine report summarizes the whole methylation status, including chromosome, position, strand, methylated and unmethylated counts, C-context, and trinucleotide context. The bedGraph counts report the level of methylation for every single CpG in the genome. The M-Bias plot shows the proportion of methylation for each position across the read:

```
bismark_methylation_extractor -p -o output --parallel
12 --cytosine_report --gzip \

--bedGraph --genome_folder reference_folder alignment.
deduplicated.bam
```

bismark2report can create an HTML report combining Bismark alignment, deduplication, methylation extraction reports as well as M-bias files.

The conversion rate can be estimated from the methylation levels of *Bismark* reports in case the experiment contains the unmethylated spike-in control (*see* **Note 3**). All the Cs of the control should be converted to Ts. The conversion rate can be calculated based on the number of Cs and Ts in the sequencing. Incomplete conversion results in the false positive methylation calls due to the existence of unchanged cytosines. The ideal conversion rate should be expected as 100% which cannot be achieved in most experiments. A typical good rate needs to be higher than 95% and close to 100%.

3.4 Methylation Annotation

This analysis can be divided into three aspects: annotation of the methylation levels according to (1) the locations of CpG islands

and CGI shores; (2) the genomic features (e.g., genes, promoters, exons, introns, intergenic regions, and transcript start and transcript end sites); and (3) by the repeat density, including repeat-poor regions, repeat-rich regions, and nonrepetitive regions. The annotation results can help to investigate the landscape of genome methylation (*see* **Note 4**).

The module *roimethstat* of *MethPipe* is used to calculate the methylation level for each region. It can take *Bismark's* output—the methylated level of every single CpG, as the input (*see* **Note 5**). *-P* option is to keep the regions aligned without any reads in the output:

```
roimethstat -P -o output regions.bed input
```

Apart from the DNA methylation landscape, *MethPipe* can identify the profiles of various methylation characteristics, such as hypermethylated regions (HyperMR), hypomethylated regions (HypoMRs), partially methylated regions (PMRs), and regions with allele-specific methylation (AMRs).

First, *MethPipe* needs to convert the methylation level to the symmetric one, which can be done by the command below. Then the methylation levels of HypoMRs can be called by *hmr* function:

```
symmetric-cpgs -o symmetric_level bismark_level
hmr -o hypomr_level symmetric_level
```

The levels of HyperMRs can also be called by *hmr* function, but with an extra step:

```
awk '{$5=1-$5; print $0}' symmetric_level > inverted_level
hmr -o hypermr_level inverted_level
```

3.5 Analysis of Differential Methylation

The analysis of differentially methylated regions (DMRs) can reveal the methylation changes between two methylomes from the same reference, which may have important implications for the regulation of gene expression. The identification of DMRs provides a comprehensive investigation of epigenetic changes among the samples. *MethPipe* provides two methods to calculate DMRs for different types of datasets. Thus, *methdiff* and *dmr* are designed for the small groups of methylomes, while the beta-binomial regression implemented by *radmeth* is appropriate for the datasets composed of a larger number of methylomes with multiple replicates.

For small datasets, *methdiff* takes the methylation level of every single base as input, and, in the output, creates a table containing the probability that the methylation level at each given site is lower in the dataset1 than in the dataset2, which is calculated by one-directional version of Fisher's exact test. Then, *dmr* can use HMRs

of both datasets to find the DMRs, and its two outputs show the DMRs with lower methylation of both datasets separately compared with each other:

```
methdiff -o output_methdiff data1_level data2_level

dmr output_methdiff data1_hmr data2_hmr mr_data1_lt_data2
dmr_data2_lt_data1
```

For large datasets, *MethPipe* recommends using at least three replicates in the analysis based on the beta-binomial regression. First, *merge-methcounts* merges all the datasets into one table, and the user needs to create a sample sheet to describe the conditions. Second, raw DMRs are calculated by *radmeth regression*. However, the *p*-value is not ideal for the test. Therefore, the next step is to calculate the adjusted *p*-value that is applied to capture the differential methylation sites. Finally, the DMRs are determined by *radmeth merge*:

```
merge-methcounts -t data1_rep1_level data1_rep2_level
data1_rep3_level \

data2_rep1_level data2_rep2_level data2_rep3_level >
merged.table

radmeth regression -factor case sample_sheet.txt merged.
table > cpgs.bed

radmeth adjust -bins 1:200:1 cpgs.bed > cpgs.adjusted.bed

awk '$5 < 0.01 "{ print $0; $}"' cpgs.adjusted.bed > dm_
cpgs.bed

radmeth merge -p 0.01 cpgs.adjusted.bed > dmrs.bed
```

The analysis of DMRs for large datasets by *MethPipe* is complicated due to combination of different commands. Other alternative tools are available to facilitate the analysis, such as a novel computational pipeline *DMRfinder* [19] and an R package *methylAction* [20].

4 Notes

1. As a time-consuming step, *Bismark* alignment can be parallelized by dividing the alignment according to the subgroups of reads or the chromosomes.
2. *Bismark* can be successfully used for the alignment of majority of BS-Seq datasets. However, if the mapping efficiency of paired-end BS-Seq reads is not good enough, different mapping strategies may be applied instead. *--score_min* can be set to a low value to increase the number of aligned reads. Then, after

normal mapping of paired-end reads, the unmapped reads can be left and realigned in the single-end nondirection mode as inputs. This can rescue some misaligned paired-end reads appeared due to the incorrect insert size resulting from the narrow sequencing area. The results from both the paired-end alignment and the single-end alignment can be merged for the methylation calls.

3. The unmethylated lambda phage DNA where all Cs should, ideally, be converted into Ts by bisulfite treatment, is recommended to be included in the analysis. Thus, the efficiency of bisulfite conversion in the samples can be estimated based on the proportion of the unconverted Cs in lambda phage DNA. Generally, the rate of conversion in the spike-in control should be above 99%.
4. Annotations of CpG islands, genes, promoters, and repeat regions can be downloaded from UCSC and Ensembl databases. Promoter regions are usually defined as sequences located between 1000 bp upstream and 500 bp downstream of a transcription start site. Promoters with high-CpG content (HCP) contain a 500-bp region with a CpG ratio larger than 0.75 and a GC content larger than 55%. Promoters with low-CpG content (LCP) do not contain any 500-bp regions with a CpG ratio larger than 0.48. Intermediate-CpG promoters (ICPs) are neither HCP nor LCP [21].
5. CpGs of the reference may, in some cases, contain common SNPs. Removing known C/T SNPs is recommended before starting the methylation calls. The sequences of known SNPs can be obtained from NCBI dbSNP database.

References

1. Deng X, Song X, Wei L et al (2016) Epigenetic regulation and epigenomic landscape in rice. Natl Sci Rev 3:309–327
2. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33:245–254
3. Schulz WA, Steinhoff C, Florl AR (2006) In: Doerfler W, Böhm P (eds) Methylation of endogenous human Retroelements in health and disease BT - DNA methylation: development, genetic disease and cancer. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp 211–250
4. Ball MP, Li JB, Gao Y et al (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol 27:361–368
5. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9:465–476
6. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet 11:191–203
7. Warnecke PM, Stirzaker C, Melki JR et al (1997) Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. Nucleic Acids Res 25:4422–4426
8. Simon A (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
9. Chen S, Zhou Y, Chen Y, Gu J (2018) Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890

10. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120
11. Jiang H, Lei R, Ding SW, Zhu S (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15:1–12
12. Krueger F (2012) Trim Galore: a wrapper script to automate quality and adapter trimming. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore
13. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. Bioinformatics 27:1571–1572
14. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9:357–359
15. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements Daehwan HHS public access. Nat Methods 12:357–360
16. Broad Institute (2019) Picard Toolkit. https://github.com/broadinstitute/picard
17. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
18. Song Q, Decato B, Hong EE et al (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One 8:e81148
19. Gaspar JM, Hart RP (2017) DMRfinder: efficiently identifying differentially methylated regions from MethylC-seq data. BMC Bioinformatics 18:1–8
20. Bhasin JM, Hu B, Ting AH (2016) MethylAction: detecting differentially methylated regions that distinguish biological subtypes. Nucleic Acids Res 44:106–116
21. Weber M, Hellmann I, Stadler MB et al (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39:457–466