Analysis of plant DNA methylation profiles using R

Catoni, Marco; Zabet, Nicolae Radu

DOI:
10.1007/978-1-0716-1134-0_21

Analysis of plant DNA methylation profiles using R

Marco Catoni[1] and Nicolae Radu Zabet[2]

Affiliation:

[1]School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[2]School of Life Sciences, University of Essex, Colchester, CO4 3SQ, United Kingdom

Correspondence:

Marco Catoni: m.catoni@bham.ac.uk

Radu Zabet: nzabet@essex.ac.uk

Running head:

Analysis of plant DNA methylation profiles using R

Abstract

DNA methylation is a transgenerational stable epigenetic modification able to regulate gene expression and genome stability. The analysis of DNA methylation by genome-wide bisulfite sequencing become the main genomic approach to study epigenetics in many organisms; leading to standardisation of the alignment and methylation call procedures. However, subsequent steps of the computational analysis should be tailored to the biological questions and the organisms used. Since most bioinformatics tools designed for epigenetic studies are built using mammalian models, they are potentially unsuitable for organisms with substantially different epigenetic regulation, such as plants. Therefore, in this chapter we propose a computational workflow for the analysis, visualisation and interpretation of data obtained from alignment of whole genome bisulfite sequencing of plant samples. Using almost exclusively the R working environment we will examine in depth how to tackle some plant-related issues during epigenetic analysis.

Keywords: Plant Epigenetics, Differentially Methylated Regions, cytosine methylation

Introduction

DNA methylation is an inheritable epigenetic mark found in many eukaryotic organisms, consisting of the addition of a methyl group to the carbon-5 position of the cytosine~~s~~ ring. Although this molecular mark leaves the DNA sequence unaltered, it influences many biological processes, including transposable element~~s~~ (TEs) silencing, gene expression and genome stability amongst others [1].

DNA methylation can be studied by treating the DNA with sodium bisulfite, a chemical that deaminates unmethylated cytosines into uracile, whil~~e~~ methylated cytosines are protected during the reaction. Consequently, in downstream sequencing reactions methylated cytosines remain unchanged and unmethylated cytosines are converted to thymines allowing definition of the DNA epigenetic profile at a single base resolution [2]. The use of next generation sequencing associated with bisulfite treatment allowed the development of Bisulfite sequencing (BS-Seq) or Whole Genome Bisulfite Sequencing (WGBS) protocols; which can be used for mapping the epigenetic profile of an entire genome [3, 4]. These methods are routinely applied to many organisms and are considered the gold standard for epigenetic studies.

The analysis of the data produced by these strategies can be divided in two parts. The first part includes well-established protocols of alignment to a reference genome, followed by calling methylation levels at each cytosine position by comparing the number of covering reads supporting the presence of methylation (read as cytosines) and the absence of methylation (read as thymine) [5–7].

The second part of the analysis is more variable and mostly dependent on the experimental design and the studied model. For example, while in mammals DNA methylation occurs almost uniquely at cytosines in CG context (cytosine followed by a guanine), in plants all cytosines can be methylated and at least three contexts are described, namely CG, CHG and CHH (where H represents any nucleotide except guanine) [4]. Although methylation in CG and non-CG contexts appear to be at least partially functionally linked [8], the methylation in each context depends on the affinity of specific methyltransferases, which can be directly linked to an epigenetic pathway [9]. Consequently, for epigenetic analysis involving plants, it is normal to inspect the three contexts independently.

Furthermore, there are at least two main issues with analysing methylation data at single cytosine levels. Firstly, independent of genome-wide sequencing depth, there are always cytosines for which the read coverage is too low and this can prevent accurate detection of changes in methylation levels [10, 11]. Secondly, methylation data needs to be interpreted in relation to functional features (e.g. TEs, enhancers, genes, promoters), which contain stretches of cytosines that consistently change their methylation level. In most cases, a change in methylation state of a single cytosine is not sufficient to trigger a biological effect. Due to these issues, the interpretation of DNA methylation data is challenging when individual cytosines are considered. However, taking into account that cytosine methylation levels display high spatial correlation (at least in CG and CHG contexts) [10, 12], one possibility is to consider methylation of neighbouring cytosines together, thus reducing the noise generated by the independent use of single positions. This solution is implemented in most DNA methylation analysis workflows, and it is a common procedure to merge DNA methylation information in regions of annotated features.

Here, we describe a protocol for the analysis of WGBS data applied to the study of plants DNA methylation profiles. All steps are associated to examples implemented using the popular R programming language [13], in order to facilitate users to adapt the scripts to their own analysis.

Materials

Cytosine methylation report

The protocol described here assumes that a genome-wide cytosine report file (CX_report) has been generated for each sample considered for the analysis. CX_report is the most complete output of Bismark [6], a popular tool used for genome-wide alignment and methylation call of DNA reads obtained by high-throughput sequencing of bisulfite converted DNA libraries (*see* **Note 1**).

The CX_report is generated as tab-delimited text file containing information for each cytosine in the genome, with the following format:

<chromosome> <position> <orientation> <count methylated> <count unmethylated> <context> <trinucleotide context>

Here is an example:

```
3     417   +     13    3     CG    CGT
3     418   -     6     0     CG    CGC
3     421   -     2     5     CHH   CAA
3     427   -     6     1     CHH   CAA
3     428   -     2     5     CHH   CCA
3     429   +     1     19    CHH   CCT
3     430   +     11    9     CHG   CTG
3     432   -     5     4     CHG   CAG
3     433   +     15    3     CG    CGT
3     434   -     9     2     CG    CGC
```

It is important to note that the protocol requires this seven-column text file and not a file specifically generated by Bismark. This means that other tools can be used to perform the methylation call, such as BS-Seeker [14, 15] or BSMAP [7], as long as the output of those tools is then converted to a text file with the seven columns described above.

DMRcaller

The R package 'DMRcaller' is designed to analyse DNA methylation data starting with the Bismark CX_report files or any other tab-delimited file formatted accordingly [10]. DMRcaller implements three different methods for identification of Differentially Methylated Regions (DMRs) in two samples or in two groups of biological replicates. In addition to its main task, DMRcaller also integrates a series of additional functions designed to facilitate analysis of WGBS experiments, including plotting functions.

Tools to export DMRs from R

Internally, DMRcaller stores the DMRs as GRanges objects [16]. There are several Bioconductor packages that can export GRanges to bed files. The most popular is 'rtracklayer' [17] which is designed for importing and exporting annotated data in various formats compatible with the main genome browsers. Alternatively, `genomation` package can also be used to export the DMRs to bed or bedGraph files that can be then loaded in genome browsers [18].

IGV

The Integrative Genomics Viewer (IGV) [19] is a tool design for the visualization and interactive exploration of large genomics datasets.

Workflow

Loading files

The DMRcaller function readBismark can be used to import CX_reports files directly in R, or any other cytosine methylation reports formatted accordingly to the Bismark output. DMRcaller imports CX_reports files and stores them as GRanges objects [16] with the following metadata columns:

- **context** – the context of the cytosine (CG, CHG or CHH).
- **readM** – the number of methylated reads (corresponding to the 'count methylated' field in the CX_report file).
- **readN** – the total number of reads (the sum of 'count methylated' and 'count unmethylated' fields in the CX_report file).
- **trinucleotide context** - the specific context of the cytosine (as the corresponding field reported in the CX_report file).

Calculate conversion rate

One important step in any epigenetic analysis that includes bisulfite conversion is the estimation of cytosine conversion rate. In theory, all unmethylated cytosines should be converted to uraciles but many variables can influence the efficiency of the bisulfite reaction, resulting in the retention of unmethylated cytosines. Unconverted unmethylated cytosines, if not taken into account, are wrongly considered methylated in downstream analysis, which can lead to data misinterpretation.

Methods to estimate bisulfite conversion efficiency are based on known unmethylated DNA regions, which are either naturally present in the sample or derived from synthetic DNA artificially incorporated before the bisulfite treatment. In many plants, chloroplast DNA has been found to display low or absent methylation [20, 21], and therefore represents a practical target to check bisulfite conversion efficiency. Chloroplast DNAs have been successfully used to estimate conversion rate in several plants, including Arabidopsis [3] , rice [22], tomato [23], soybean [24], eggplant[25], and many others.

Load the cytosine report:

```
CX_report <- DMRcaller::readBismark("CXreport.txt")
```

Extract the chloroplast methylation data:

```
PtDNA <- CX_report[seqnames(CX_report) == "KU682719"]
```

Calculate conversion:

```
conversion <- 1 - (sum(mcols(PtDNA)$readsM) / sum(mcols(PtDNA)$readsN))}
```

Correction for conversion rate

Once that the conversion rate is estimated, methylation levels can be adjusted by taking into account unconverted cytosines. Here we apply a method that decreases the number of reported methylated cytosine positions accordingly to the estimated conversion rate [4, 26].

The number of methylated reads is decreased at each cytosine position with the following function:

$$m^* = \lfloor \max(0, m - n(1 - c)) \rfloor$$

*m** = adjusted number of methylated reads per cytosine position.
m = original number of methylated reads per cytosine position.
n = total number of reads per cytosine position.
c = the conversion rate.

In R this can be implemented as:

```
CX_report_adjusted <- CX_report
CX_report_adjusted$readsM <- round(CX_report$readsM - CX_report$readsN * (1-
conversion))
CX_report_adjusted$readsM[CX_report_adjusted$readsM < 0 ] <- 0
```

Using this correction at each cytosine position, the total coverage should be decreased according to the conversion rate, which prevents overestimated coverage. This can be done with the simple function:

$$n^* = \lfloor nc \rfloor$$

*n** = adjusted number of total reads per cytosine position
n = original number of reads per cytosine position
c = estimated conversion rate

In R is implemented as:

```
CX_report_adjusted$readsN <- round(CX_report$readsN * conversion)
```

Then, a new CX report can be generated using DMRcaller:

```
DMRcaller::SaveBismark(CX_report_adjusted,"CX_report_adjusted.txt"))
```

Generate bedGraph for genome browser visualization

It is sometimes useful to visualise epigenetic data in a genome browser (e.g., IGV), which allows a visual interactive comparison of different samples in multiple tracks at any genomic location. The direct visualization of DNA methylation at specific genes can help to identify genomic areas under epigenetic regulation without running genome-wide unsupervised analysis (Figure 1). The cytosine report needs to be converted into a compatible file format as it cannot be directly loaded into a genome browser. It is important at this step to separate into different tracks the methylation of the different cytosine contexts (CG, CHG and CHH),

First, the CX_report should be loaded in R:

```
CX_report <- DMRcaller::readBismark("CXreport.txt")
```

Then, methylation in a specific context is selected (e.g., CG)

```
selection <- CX_report[which(CX_report$context=="CG")]
```

Optionally, cytosines with low coverage (e.g., less than 4 reads) might be excluded from the track to reduce the noise.

```
selection <- selection[selection$readsN >= 4]
```

The proportion of methylated reads for the selected cytosines can be calculated:

```
selection$score <- selection$readsM / selection$readsN
```

Finally, a bedgraph file can be generated using rtracklayer. Considering that bedgraph files are often very large, it may be useful to generate a bigwig file instead, which is compressed and can be loaded on IGV in a shorter time.

```
rtracklayer::export.bedGraph(selection, "CG_track.bedGraph")
rtracklayer::export.bw(selection, "CG_track.bw")
```

For example, Figure 1 shows how the direct comparison of DNA methylation profiles obtained from *Arabidopsis thaliana* and eggplant (*Solanum melongena*) can be useful to identify the most probable position of the DNA region controlling the *IBM1* gene splicing in eggplant using the *A. thaliana* functional annotation [27].

Computing the methylation frequency

In plants, at each cytosine context the methylation is maintained with a different degree of efficiency that depends on the specific epigenetic pathway involved [4]. Therefore, it is often informative to plot the distribution of methylation levels. This is usually done in intervals of 10%, using ten bins to cover methylation values from 0% to 100%.

It is important to consider that cytosines with low read depth are not informative in computing methylation frequency. Therefore, the data should be filtered to include only cytosines with a read depth that is higher than the number of bin used (e.g. if ten bins are used, only positions covered with more than 10 reads should be selected for this analysis).

```
CX_report <- DMRcaller::readBismark("CXreport.txt")
CX_report_cov <- CX_report[which(CX_report$readsN > 10)]
```

To exemplify this, we will show an example of how to calculate the proportion of methylated cytosines at each bin in all three cytosine contexts:

- Methylation percentage frequency for CG methylation

```
CX_report_CG <- CX_report_cov[CX_report_cov$context=="CG"]
CG_freq <- hist(100* CX_report_CG$readsM / CX_report_CG$readsN,
breaks=seq(0,100,by=10), plot=FALSE)
```

- Methylation percentage frequency for CHG methylation

```
CX_report_CHG <- CX_report_cov[CX_report_cov$context=="CHG"]
CHG_freq <- hist(100* CX_report_CHG$readsM / CX_report_CHG$readsN,
breaks=seq(0,100,by=10), plot=FALSE)
```

- Methylation percentage frequency for CHH methylation

```
CX_report_CHH <- CX_report_cov[CX_report_cov$context=="CHH"]
CHH_freq <- hist(100* CX_report_CHH$readsM / CX_report_CHH$readsN,
breaks=seq(0,100,by=10), plot=FALSE)
```

Then, the methylation frequencies can be visualized using standard R plot function (Figure 2):

```
cbbPalette <- c("#000000", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2",
"#D55E00", "#CC79A7")
```

```
bar_counts <- rbind(100*CG_freq$counts/sum(CG_freq$counts),
                    100*CHG_freq$counts/sum(CHG_freq$counts),
                    100*CHH_freq$counts/sum(CHH_freq$counts))
rownames(bar_counts) <- c("CG", "CHG", "CHH")
colnames(bar_counts) <- paste0(CG_freq$breaks[1:(length(CG_freq$breaks)-1)], "-",
                               CG_freq$breaks[2:length(CG_freq$breaks)])
barplot(bar_counts, xlab="% of methylation", beside = TRUE,
     ylab=paste0("% of Cs"), las=1, ylim=c(0,100), col=cbbPalette[c(7,6,4)])
legend("topright", rownames(bar_counts), fill=cbbPalette[c(7,6,4)], bty="n")
```

Figure 2 shows that majority of CHH sites display low or lack of methylation (< 10%), while, for CG sites , there is a large proportion of sites (approximately 40%) that display high level of methylation (>80%). Finally, majority of CHG sites are unmethylated but there is a small proportion of sites displaying intermediary and high level of methylation.

Coverage calculation and spatial correlation

The next step of the analysis consists of performing some preliminary analysis that will inform the selection of the DMR calling method. First, one needs to evaluate the coverage or the read depth of the libraries. To exemplify these steps, we can use a dataset from *A. thaliana* Col-0 2 weeks seedling in WT plants (GSM2384978) and *met1-1* plants (GSM2384979) [26] (*see* **Note 2**).

Once the files are downloaded, they can be loaded in R with DMRCaller as follow:

```
wt <- DMRcaller::readBismark("GSM2384978_wt_processed.txt.gz")
met1 <- DMRcaller::readBismark("GSM2384979_met1-1_processed.txt.gz")
```

Then, the proportion of cytosines with coverage above a customisable set of thresholds (in this example 1, 5, 10 and 15) can be computed and plotted (Figure 3) for each cystosine context using the following function:

```
DMRcaller::plotMethylationDataCoverage(wt, met1, breaks = c(1,5,10,15),
conditionsNames=c("WT","met1-1"), context = c("CG", "CHG", "CHH"), labels=LETTERS)
```

This step allows to evaluate the sequencing depth and setup strategies for the downstream analysis. In particular, for this dataset, we found that approximately 30-40% of the cytosines have at least 5-10 reads (Figure 3), which means that calling differentially methylated cytosines might have missed some true sites. Increasing the sequencing depth can partially solve this problem, but even highly sequenced libraries will not lead to all cytosines having at least 10 reads (*see* **Note 3**). There are several ways to computationally address this issue, and most of them assume merging several cytosines and pooling together the reads in those regions. We will discuss several options in the following sections below.

Calculate DNA methylation in features.

One popular approach is to determine if different genetic features display differential methylation. This approach consists of selecting an annotation file and pooling all methylated reads and unmethylated reads in each of the genomics features. DMRcaller supports this functionality by providing the *filterDMRs* function.

If DMRcaller package is installed, the bisulfite sequencing data and the annotation file can be loaded with:

```
data(methylationDataList)
data(GEs)
```

Note that the *methylationDataList* is a list object contains a subset of methylation data from [26]. Similar objects can be generated by using the `list` function and the imported CX_report files. In this example:

```
CX_WT <- DMRcaller::readBismark("CXreport_WT.txt")
CX_met1 <- DMRcaller::readBismark("CXreport_met1-3.txt")
methylationDataList <- list("WT" = CX_WT, "met1-3" = CX_met1)
```

The *GEs* is a GRanges object representing TAIR10 annotation of *Arabidopsis thaliana* genome, obtained by using the import function from rtracklayer:

```
GEs <- rtracklayer::import(
"https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_gff3
/TAIR10_GFF3_genes_transposons.gff")
```

Then, gene features can be filtered from the annotation object using the following command:

```
genes <- GEs[which(GEs$type == "gene")]
```

If we do not want to analyse the entire genome, a GRanges object should be created to select only the area of interest (in this example, 100 Kb DNA fragment of chromosome 3):

```
chr_local <- GRanges(seqnames = Rle("Chr3"), ranges = IRanges(5E5,6E5))
```

Finally, the *filterDMRs* function can be used to identify gene with statistical methylation differences between the two conditions:

```
DMRsGenesCG <- DMRcaller::filterDMRs(methylationDataList[["WT"]],
                          methylationDataList[["met1-3"]],
                          potentialDMRs = genes[overlapsAny(genes, chr_local)],
                          context = "CG", test = "score",pValueThreshold = 0.01,
                          minCytosinesCount = 4,minProportionDifference = 0.4,
                          minReadsPerCytosine = 3,cores = 1)
```

This can be very useful in identifying genes which are differentially methylated between two conditions. However, often changes in methylation can influence the expression of a gene even if they only partially overlap (or do not overlap at all) its transcribed sequence. Moreover, in plant most coding genes are not equally methylated along their sequence [4]. A methylation change between two conditions could be strongly underestimated if a single methylation value is estimated by averaging all cytosines in the gene coding sequence. In other words, the arbitrary definition of regions to test a difference in methylation does not necessarily correspond to the genomic area where the change in methylation occurred.

To visualize this issue with an example, we can plot the locus of the Arabidopsis gene AT3G02490 on chromosome 3 using the DMRcaller function *plotLocalMethylationProfile.*

We should select a 20 Kb location on the chromosome 3:

```
chr3Reg <- GRanges(seqnames = Rle("Chr3"), ranges = IRanges(510000,530000))
```

and then create a list with all genes differentially methylated identified in our previous analysis:

```
DMRsCGList <- list("genes" = DMRsGenesCG)
```

We can now use the function to generate the plot:

```
par(cex=0.9)
par(mar=c(4, 4, 3, 1)+0.1)
DMRcaller::plotLocalMethylationProfile(methylationDataList[["WT"]],
                            methylationDataList[["met1-3"]],
                            chr3Reg,DMRsCGList,
                            conditionsNames = c("WT", "met1-3"),
                            GEs,windowSize = 300,main="CG methylation")
```

In the plot (Figure 4), we can notice that only a small part of the gene displays methylation in wild type that is not present in *met1-3* mutant. Nevertheless, this difference is not enough to be statistically significant if the sequence of the entire gene is used to run the analysis.

In this case, it would be more appropriate to investigate differentially methylated regions (DMRs), independently from gene annotation. When a list of DMRs will be generated (as explained in the next section), one could check if genes (or other features) overlap with any DMRs. For example, one could do this by using the following commands (assuming that DMRs are listed in a GRanges object called *DMRs*)

```
DMGenes <- genes[overlapsAny(genes, DMRs)]
```

Call Differentially Methylated Regions (DMRs)

Call of DMRs is now an essential part of any WGBS analysis. In this analysis, genomic regions are determined and selected by the presence of differences in methylation between two samples. This approach avoids assumptions related to the use of predetermined features where methylation is expected to change (e.g. genes, promoters) and it is therefore preferred for unsupervised analysis.

In DMRCaller, the same function *computeDMRs* can be used to call DMRs with one of the three methods implemented (*see* **Note 4**); it is sufficient to specify the method of choice with the *method* parameter (possible choices are among *noise_filter*, *bins* and *neighbouring*, a full description of how these methods are implemented is provided in [10]).

An example of how to compute the DMRs in CG context with *noise_filter* method is:

```
DMRsNoiseFilterCG <- DMRcaller::computeDMRs(methylationDataList[["WT"]],
                      methylationDataList[["met1-3"]],
                      context = "CG", method = "noise_filter",
                      windowSize = 100, pValueThreshold = 0.01,
                      minCytosinesCount = 4, minProportionDifference = 0.4,
                      minGap = 200, minReadsPerCytosine = 4,
                      cores = 1)
```

Similarly, the DMRs in CHH context can be computed using *bins* method as follows:

```
DMRsBinsCHH <- DMRcaller::computeDMRs(methylationDataList[["WT"]],
               methylationDataList[["met1-3"]],
               context = "CHH", method = "bins", binSize = 100,
               pValueThreshold = 0.01, minCytosinesCount = 4,
               minProportionDifference = 0.1, minGap = 200,
               minReadsPerCytosine = 4, cores = 1)
```

The additional arguments of the function can be changes to adapt the analysis to the data structure. Here are following useful considerations for some of these parameters.

- `binSize/windowSize` (default = 100) can be changed depending by the desired output. Higher values will produce longer DMRs including more cytosines, while lower values are more efficient in detection of small DMRs. A previous study investigated how different value for this argument affect the DMR call [10].
- `regions` argument can be used to limit the DMR call to only a part of the genome. For example one can run a pilot analysis for parameter optimisation limiting the computational time only on one chromosome or a part of it.
- `minProportionDifference` controls the minimal differences between the methylation values in the two conditions which are considered significant. This threshold can be used to avoid calling DMRs with small changes of DNA methylation, under the assumption that small methylation changes between two conditions (even if statistically significant) are not biological relevant (*see* **Note 5**).
- `minGap` can be used to control how distant two DMRs should be merged together. This parameter affects the number of DMRs generated, but if set to 0 it will force the generation of not overlapping DMRs of identical length (equal to the `binSize`) when used in conjunction with *bins* method (*see* **Note 6**).
- `minCytosinesCount` controls the minimum number of cytosine in a DMR. Setting this as threshold will avoid calling significant differences in DMRs that are constituted by only one or few isolated cytosines (and therefore not properly defined as "regions") (*see* **Note 7**).
- `minReadsPerCytosine` is a threshold used to discard from the analysis DMRs with an average number of reads lower than this value. Higher values of this parameter ensure reliable results, but they also exclude proportional larger genomic area from the analysis, which is less covered.
- `cores` is the number of CPUs/cores used for the computation. More cores will lead to faster computations.

In many cases, it is possible to have access to datasets including biological replicates. One possible approach is to merge different biological replicates, but DMRcaller also allows treating the replicates independently (*see* **Note 8**).

First, the CX_reports files from each condition are loaded in R:

```
CX_CTR_rep1 <- DMRcaller::readBismark("CX_CTR_rep1.txt")
CX_CTR_rep2 <- DMRcaller::readBismark("CX_CTR_rep2.txt")
CX_TEST_rep1 <- DMRcaller::readBismark("CX_TEST_rep1.txt")
CX_TEST_rep2 <- DMRcaller::readBismark("CX_TEST_rep2.txt")
```

Then, the *joinReplicates* function is used iteratively to combine all data in the same object.

```
CX_all_data <- DMRcaller::joinReplicates(CX_CTR_rep1, CX_CTR_rep2)
CX_all_data <- DMRcaller::joinReplicates(CX_all_data, CX_TEST_rep1)
CX_all_data <- DMRcaller::joinReplicates(CX_all_data, CX_TEST_rep2)
```

A vector of labels should be generated to identify the samples:

```
condition_labels <- c("CTR", "CTR", "TEST", "TEST")
```

At this point, it is possible to call DMRs (in this example in CG context), using the beta regression test:

```
DMRs_CG <- DMRcaller::computeDMRsReplicates(CX_all_data, condition =
condition_labels, context = "CG",method = "bins")
```

Once the list of DMRs has been generated, it can be exported from R as txt file, or as annotation (bed or gff) file, by using rtracklayer.

```
write.table(as.data.frame(DMRs_CG),file="DMRs_CG.txt",sep="\t", quote=F)
rtracklayer::export(DMRs_CG, "DMRs_CG.gff3")
```

Call Differentially Methylated Cytosines (DMCs)

Although summarising DNA methylation information per features and call DMRs are a common procedure performed in WGBS analysis, in some conditions, the call of DMCs can also be informative (*see* **Note 9**).

With DMRcaller, DMCs can be simply calculated using the *computeDMRs* function and the *neighbouring* method, selecting a *minGap* value of zero. In this way single cytosines will be tested to be differentially methylated but not merged together to generate regions, and an output is provided as a list of single differentially methylated cytosines. The following is an example of how to run this analysis in R:

```
DMCs <- DMRcaller::computeDMRs(methylationDataList[["WT"]],
                    methylationDataList[["met1-3"]],
                    regions = chr_local, context = "CG",
                    method = "neighbourhood", test = "score",
                    pValueThreshold = 0.01, minCytosinesCount = 1,
                    minProportionDifference = 0.4,minGap = 0,
                    minSize = 1, minReadsPerCytosine = 4)
```

In this case, 1.5% of the CG sites at Chr3R:500,000-600,000 are detected to display differential methylation between WT and *met1-3* mutant. This method leads to correct identification of a small region inside AT3G02490 gene with a change in methylation, which could be missed when computing DMRs using a too large window size or the entire gene as feature (Figure 4).

Plot DMRs on chromosomes

Finally, when analysing mutants that lead to global changes in methylation or different conditions that could lead to significant global changes, one can compute and plot the low-resolution profiles on each chromosome using wide bins (e.g., 200 Kb). For example, if we perform this analysis, we could see that in *Arabidopsis thaliana* the highest methylation levels are located at pericentromeric regions and, in *met1-1* mutant, CG methylation is significantly lost globally although not completely depleted (Figure 5). This is what we would expect since MET1 is the main methyltransferase involved in CG methylation maintenance and *met1-1* mutation leads to partial loss of function [26].

To plot DMRs on chromosome 1, we first select this chromosome as range of the Arabidopsis genome:

```
chr1 <- GRanges(seqnames = "1", ranges = IRanges(1,30427671))
```

Then, the following code computes the average methylation in 200 Kb bins along chromosome 1 for both wild type and *met1-1* samples:

```
chr1_wt <- DMRcaller::computeMethylationProfile(wt, chr1,windowSize = 200000,
context = "CG")

chr1_met11 <- DMRcaller::computeMethylationProfile(met11, chr1,windowSize = 200000,
context = "CG")
```

Finally, the following code can be used to plot the averaged methylation data along the chromosome and to generate figure 5:

```
plot((start(chr1_wt) + end(chr1_wt))/2, 100*chr1_wt$Proportion, type="l", lty=1,
     lwd=2, col=cbbPalette[1], main="CG methylation on Chr 1", xlab="", xaxt="n",
     ylab="methylation percentage", ylim=c(0,100))
lines((start(chr1_met11) + end(chr1_met11))/2, 100*chr1_met11$Proportion, lty=1,
      lwd=2, col=cbbPalette[6])
legend("topright", c("WT", "met1-1"), col=cbbPalette[c(1,6)], bty="n", lty=1,
lwd=2)
```

Notes

1. CX_report files are generated in Bismark by running the *bismark_methylation_extractor* command and specifying the `--cytosine_report` and `--CX` options. For a detailed description of the use of Bismark please refer to the user manual [6].
2. The corrected CX_report files can be directly downloaded from ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM2384nnn/GSM2384978/suppl/GSM2384978_wt_processed.txt.gz and from ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM2384nnn/GSM2384979/suppl/GSM2384979_met1-1_processed.txt.gz
3. One might ask why we need more reads covering each cytosine. There are two answers to this: firstly, we need more reads to perform reliable statistical test to detect differential methylation and secondly ,the more reads we have the more robust we can call the actual methylation level when this has intermediary values (e.g., we need at least 4 reads to call a cytosine being methylated in 75% of the cases).
4. There are different methods to call DMRs and it appears that the method used should be selected depending on the methylation context, coverage and tissues used to generate the data. The DMRcaller tool implements three methods to call DMRs and the performances of each of them has been previously discussed [10].
5. If a binary methylation change is expected (i.e. regions pass from being highly methylated to a complete unmethylated status) as often happens for methylation in CG context in plants, a higher value of this parameter helps to reduce noise generated by random changes. By contrary, limited variations in methylation (more common for CHH context) require a lower value of this parameter to allow detection of small changes.
6. This setting applied to the `minGap` parameter can be useful in case of multiple sample comparisons, due to the fact that the number of DMRs found in each comparison is directly informative of the portion of genome with methylation difference. Therefore, if `minGap` is set to 0, the DMR lists would not be required to be normalised by their length when compared across samples.

7. Although a high value of this argument ensures robustness of the identified methylation difference (because more positions contribute to calculate the methylation value of each region), it should be increased with caution because it could generate artefacts. For example, for small bin sizes and less frequent contexts (CG and CHG), a high value of this parameter can bias the DMRs call toward genome area with high CG content.
8. Biological replicates can be used to distinguish between true differences in methylation levels and noise. We observed that, for large difference in methylation levels, the use of biological replicates does not improve significantly the results [10]. Nevertheless, for small differences in methylation (lower than 20%), biological replicates are critical to distinguish between the noise affecting the data and true differences between biological samples.
9. For example, the methylation at single cytosine positions has proved informative to study the cytosine context specificity of plant methyltransferases [4, 8, 28], or to estimate epigenetic mutation rate in Arabidopsis [29].

Acknowledgment

We thank Ms. Jessica Scivier (University of Birmingham, UK) for critical reading and proofreading of the chapter draft.

555 **References**

556 1. Zhang H, Lang Z, Zhu J-K (2018) Dynamics and function of DNA methylation in plants. Nat Rev
557 Mol Cell Biol 19:489. https://doi.org/10.1038/s41580-018-0016-z

558 2. Frommer M, McDonald LE, Millar DS, et al (1992) A genomic sequencing protocol that yields a
559 positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci
560 89:1827–1831. https://doi.org/10.1073/pnas.89.5.1827

561 3. Cokus SJ, Feng S, Zhang X, et al (2008) Shotgun bisulphite sequencing of the Arabidopsis
562 genome reveals DNA methylation patterning. Nature 452:215–219.
563 https://doi.org/10.1038/nature06745

564 4. Lister R, O'Malley RC, Tonti-Filippini J, et al (2008) Highly Integrated Single-Base Resolution
565 Maps of the Epigenome in Arabidopsis. Cell 133:523–536.
566 https://doi.org/10.1016/j.cell.2008.03.029

567 5. Chen P-Y, Cokus SJ, Pellegrini M (2010) BS Seeker: precise mapping for bisulfite sequencing.
568 BMC Bioinformatics 11:203. https://doi.org/10.1186/1471-2105-11-203

569 6. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq
570 applications. Bioinformatics 27:1571–1572. https://doi.org/10.1093/bioinformatics/btr167

571 7. Xi Y, Li W (2009) BSMAP: whole genome bisulfite sequence MAPping program. BMC
572 Bioinformatics 10:232. https://doi.org/10.1186/1471-2105-10-232

573 8. Zabet NR, Catoni M, Prischi F, Paszkowski J (2017) Cytosine methylation at CpCpG sites triggers
574 accumulation of non-CpG methylation in gene bodies. Nucleic Acids Res 45:3777–3784.
575 https://doi.org/10.1093/nar/gkw1330

576 9. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns
577 in plants and animals. Nat Rev Genet 11:204–220. https://doi.org/10.1038/nrg2719

578 10. Catoni M, Tsang JM, Greco AP, Zabet NR (2018) DMRcaller: a versatile R/Bioconductor package
579 for detection and visualization of differentially methylated regions in CpG and non-CpG
580 contexts. Nucleic Acids Res 46:e114. https://doi.org/10.1093/nar/gky602

581 11. Lister R, Pelizzola M, Dowen RH, et al (2009) Human DNA methylomes at base resolution show
582 widespread epigenomic differences. Nature 462:315–322.
583 https://doi.org/10.1038/nature08514

584 12. Eckhardt F, Lewin J, Cortese R, et al (2006) DNA methylation profiling of human chromosomes
585 6, 20 and 22. Nat Genet 38:1378–1385. https://doi.org/10.1038/ng1909

586 13. R Core Team (2019) R: A language and environment for statistical computing. R Foundation for
587 Statistical Computing. https://www.R-project.org/

588 14. Guo W, Fiziev P, Yan W, et al (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite
589 sequencing data. BMC Genomics 14:774. https://doi.org/10.1186/1471-2164-14-774

590 15. Huang KYY, Huang Y-J, Chen P-Y (2018) BS-Seeker3: ultrafast pipeline for bisulfite sequencing.
591 BMC Bioinformatics 19:111. https://doi.org/10.1186/s12859-018-2120-7

592 16. Lawrence M, Huber W, Pagès H, et al (2013) Software for Computing and Annotating Genomic
593 Ranges. PLOS Comput Biol 9:e1003118. https://doi.org/10.1371/journal.pcbi.1003118

594 17. Lawrence M, Gentleman R, Carey V (2009) rtracklayer: an R package for interfacing with
595 genome browsers. Bioinformatics 25:1841–1842.
596 https://doi.org/10.1093/bioinformatics/btp328

597 18. Akalin A, Franke V, Vlahoviček K, et al (2015) genomation: a toolkit to summarize, annotate and
598 visualize genomic intervals. Bioinformatics 31:1127–1129.
599 https://doi.org/10.1093/bioinformatics/btu775

600 19. Robinson JT, Thorvaldsdóttir H, Winckler W, et al (2011) Integrative genomics viewer. Nat
601 Biotechnol 29:24–26. https://doi.org/10.1038/nbt.1754

602 20. Feng S, Cokus SJ, Zhang X, et al (2010) Conservation and divergence of methylation patterning
603 in plants and animals. Proc Natl Acad Sci 107:8689–8694.
604 https://doi.org/10.1073/pnas.1002720107

605 21. Fojtová M, Kovařík A, Matyášek R (2001) Cytosine methylation of plastid genome in higher
606 plants. Fact or artefact? Plant Sci 160:585–593. https://doi.org/10.1016/S0168-9452(00)00411-
607 8

608 22. Li X, Wang X, He K, et al (2008) High-Resolution Mapping of Epigenetic Modifications of the
609 Rice Genome Uncovers Interplay between DNA Methylation, Histone Methylation, and Gene
610 Expression. Plant Cell 20:259–276. https://doi.org/10.1105/tpc.107.056879

611 23. Catoni M, Lucioli A, Doblas-Ibáñez P, et al (2013) From immunity to susceptibility: virus
612 resistance induced in tomato by a silenced transgene is lost as TGS overcomes PTGS. Plant J
613 75:941–953. https://doi.org/10.1111/tpj.12253

614 24. Song Q-X, Lu X, Li Q-T, et al (2013) Genome-Wide Analysis of DNA Methylation in Soybean. Mol
615 Plant 6:1961–1974. https://doi.org/10.1093/mp/sst123

616 25. Cerruti E, Gisbert C, Drost H-G, et al (2019) Epigenetic bases of grafting-induced vigour in
617 eggplant. bioRxiv 831719. https://doi.org/10.1101/831719

618 26. Catoni M, Griffiths J, Becker C, et al (2017) DNA sequence properties that predict susceptibility
619 to epiallelic switching. EMBO J 36:617–628. https://doi.org/10.15252/embj.201695602

620 27. Rigal M, Kevei Z, Pélissier T, Mathieu O (2012) DNA methylation in an intron of the IBM1
621 histone demethylase gene stabilizes chromatin modification patterns. EMBO J 31:2981–2993.
622 https://doi.org/10.1038/emboj.2012.141

623 28. Gouil Q, Baulcombe DC (2016) DNA Methylation Signatures of the Plant
624 Chromomethyltransferases. PLOS Genet 12:e1006526.
625 https://doi.org/10.1371/journal.pgen.1006526

626 29. Becker C, Hagmann J, Müller J, et al (2011) Spontaneous epigenetic variation in the Arabidopsis
627 thaliana methylome. Nature 480:245–249. https://doi.org/10.1038/nature10555

628

629

Figure captions

Fig. 1

Example of visualisation of epigenetic profiles with IGV. Methylation in the three main cytosine contexts is displayed for WGBS analysis of two replicates of wild type *A. thaliana* seedlings [26] and *Solanum melongena* leaf tissue [25], plotted at the *IBM1* gene locus (respectively AT3G07610 and SMEL_008g308130). The Arabidopsis *IBM1* gene contains a regulatory DNA sequence under epigenetic regulation (marked with a red rectangle) which must be methylated to allow proper splicing of the large *IBM1* intron [27]. By comparison of the *IBM1* loci in the two plants, a DNA region with similar methylation profile is evident in *S. melongena* (marked with a blue rectangle), suggesting that *IBM1* has similar epigenetic regulation in the two species.

Fig. 2

Distribution of the percentage of cytosine methylation in each sequence context in wild type *Arabidopsis thaliana* seedling (GSM2384978). The y axis indicates the frequency observed for the methylated cytosines that display the percentage of methylation indicated on the x axis. Fractions were calculated within bins of 10%, as indicated on the x axis.

Fig. 3

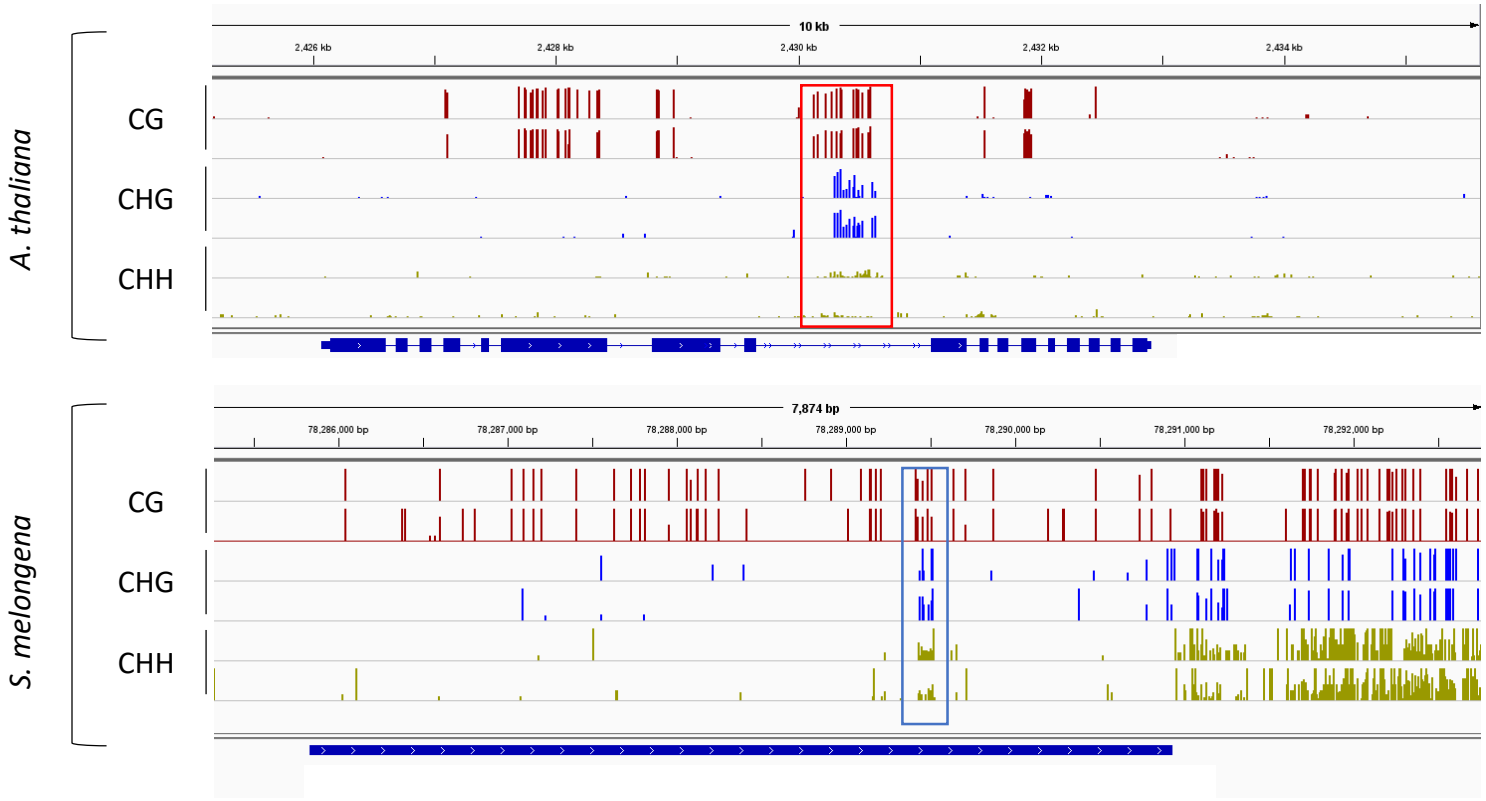
Methylation coverage calculated at the proportion of cytosine positions in the genome having at least a read depth of 1, 5, 10 and 15 reads respectively (indicated in the x axes). The data are taken from *Arabidopsis thaliana* wild type and *met1-1* mutant [26], and are displayed separately for the three main cytosine contexts.

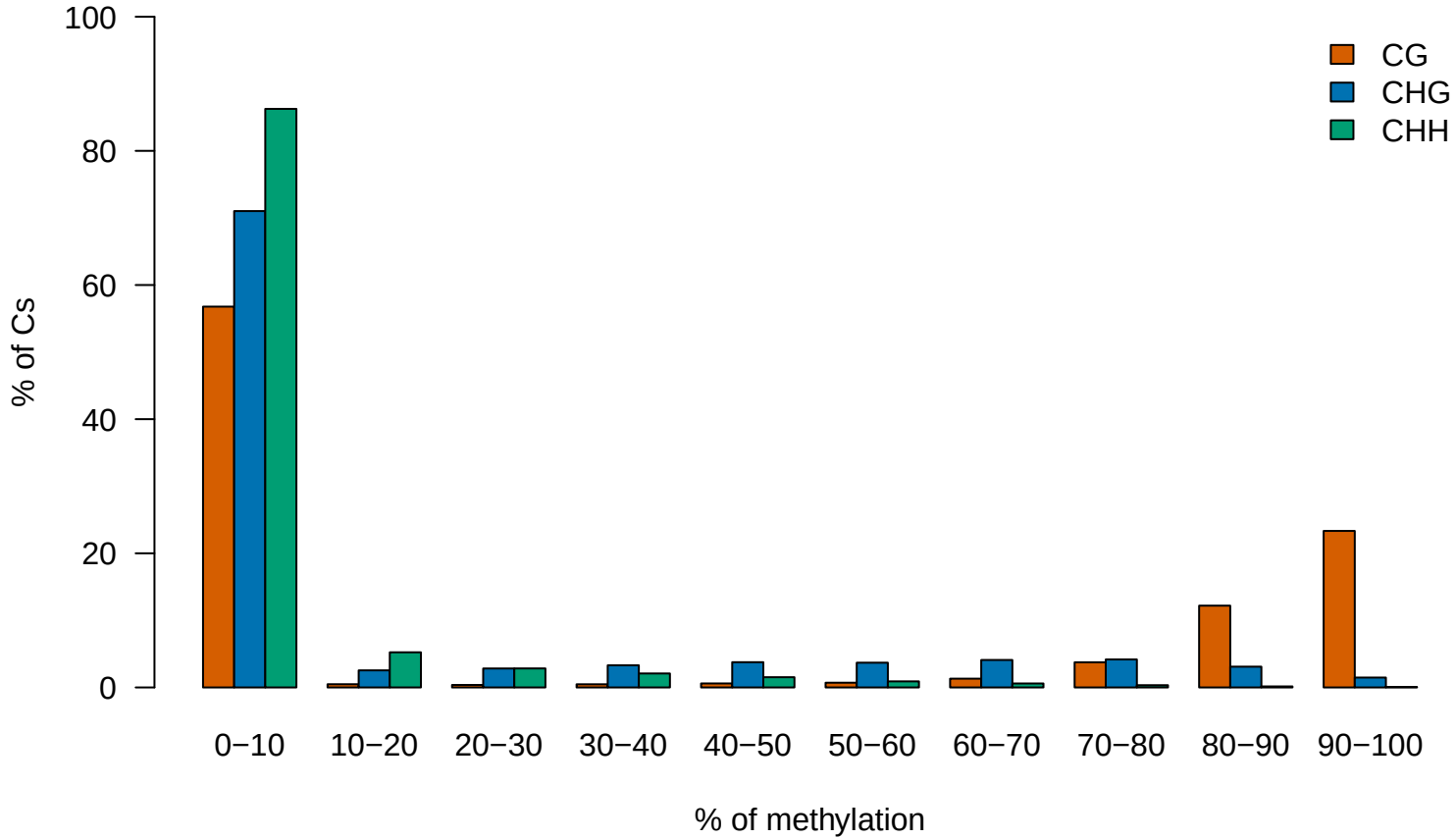
Fig. 4

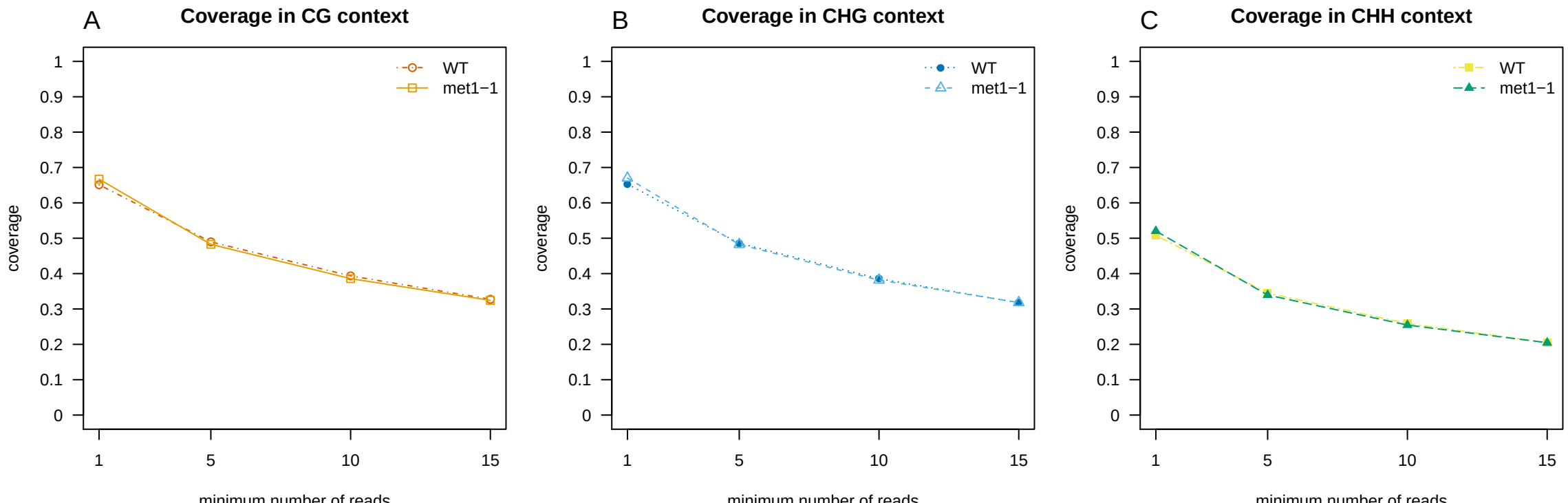
Local methylation profile plotted with DMRcaller, displaying the methylation at a Differentially Methylated Gene (DMG) located at chromosome 3. Each point on the graph represent methylation proportion of individual cytosines, in *Arabidopsis thaliana* wild type (red) or *met1* mutant (blue). The intensity of the dot colours is proportional to the read coverage of that particular cytosine (darker colours indicate higher coverage). The solid lines represent the smoothed profiles, and the intensity of the line colour is proportional to the coverage in the smoothed region. The list of annotated features used for the analysis (in this case gene exons) is displayed in the lower part of the graph as black boxes, separated in forward (+) or reverse (-) orientation. The differentially methylated region inside the gene sequence is represented by a yellow box on top of the graph.

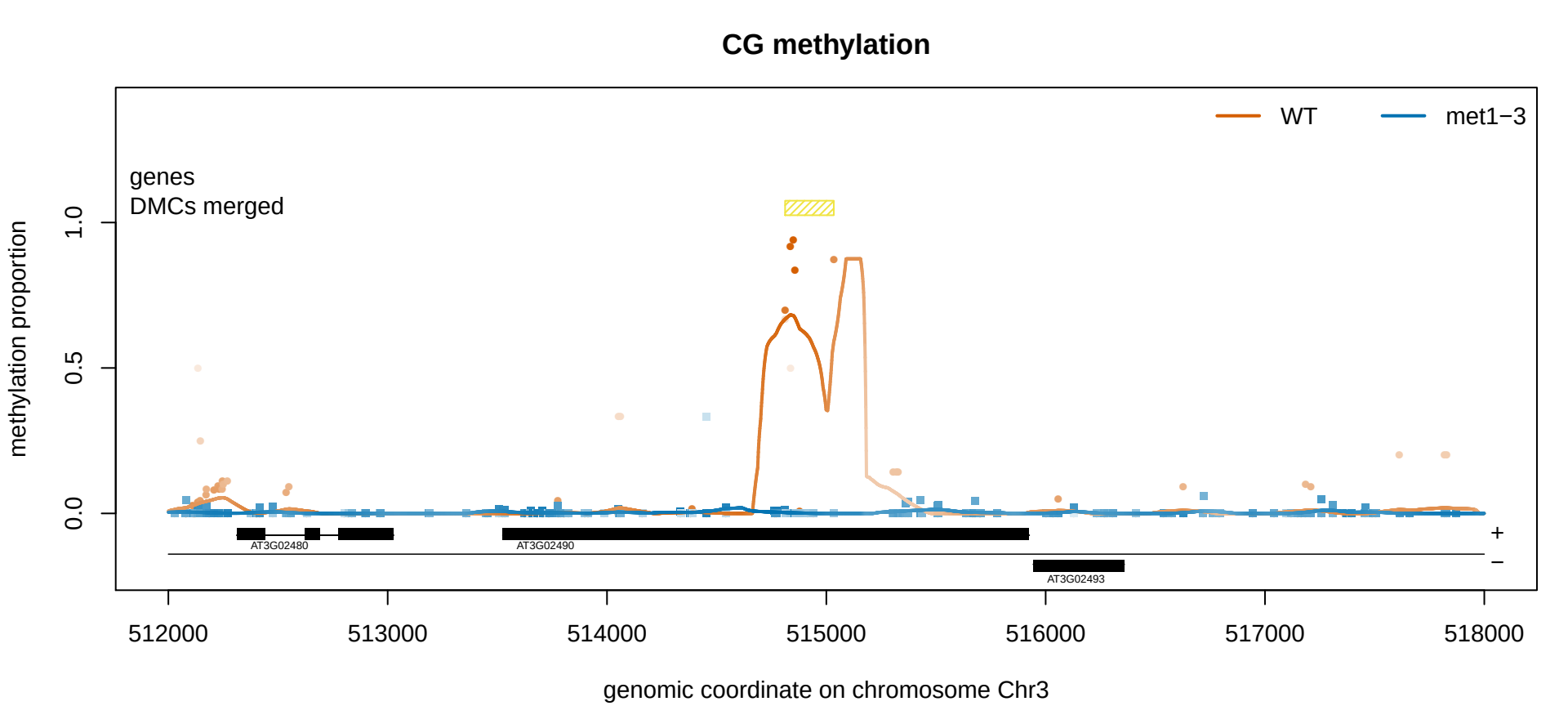
Fig.5

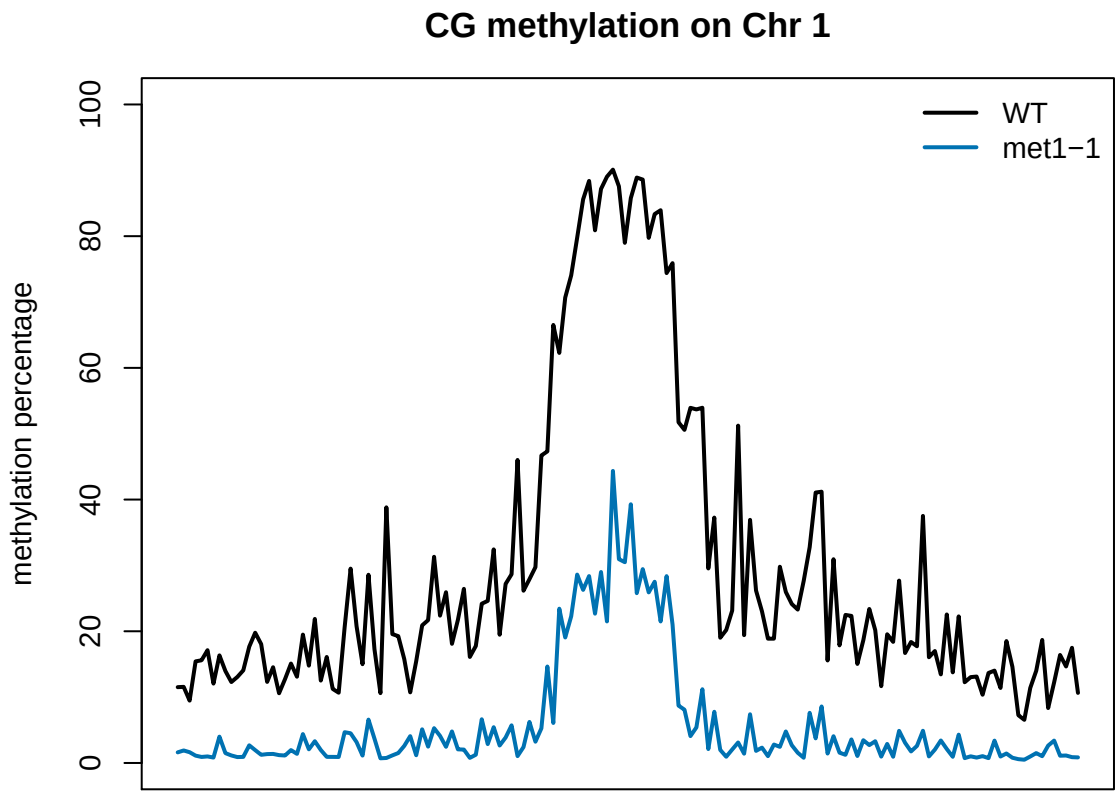
Low resolution methylation profile along chromosome 1 of *Arabidopsis thaliana* wild type and *met1-1* mutant [26], obtained by merging cytosine methylation in CG context in windows of 200 kb size. Highest methylation levels are located at centromeres and pericentromeric regions.


A. thaliana
CG
CHG
CHH
10 kb
2,426 kb
2,428 kb
2,430 kb
2,432 kb
2,434 kb
S. melongena
CG
CHG
CHH
7,874 bp
78,286,000 bp
78,287,000 bp
78,288,000 bp
78,289,000 bp
78,290,000 bp
78,291,000 bp
78,292,000 bp


100
80
60
40
20
0
% of Cs
CG
CHG
CHH
0−10
10−20
20−30
30−40
40−50
50−60
60−70
70−80
80−90
90−100
% of methylation


A
Coverage in CG context
WT
met1−1
coverage
minimum number of reads
B
Coverage in CHG context
WT
met1−1
coverage
minimum number of reads
C
Coverage in CHH context
WT
met1−1
coverage
minimum number of reads


CG methylation
WT
met1−3
genes
DMCs merged
methylation proportion
1.0
0.5
0.0
AT3G02480
AT3G02490
AT3G02493
+
−
512000
513000
514000
515000
516000
517000
518000
genomic coordinate on chromosome Chr3


CG methylation on Chr 1
methylation percentage
0
20
40
60
80
100
WT
met1−1