

Analysis of plant DNA methylation profiles using R

Catoni, Marco; Zabet, Nicolae Radu

DOI:

[10.1007/978-1-0716-1134-0_21](https://doi.org/10.1007/978-1-0716-1134-0_21)

License:

None: All rights reserved

Document Version

Peer reviewed version

Citation for published version (Harvard):

Catoni, M & Zabet, NR 2021, 'Analysis of plant DNA methylation profiles using R', *Methods in molecular biology*, vol. 2250, pp. 219-238. https://doi.org/10.1007/978-1-0716-1134-0_21

[Link to publication on Research at Birmingham portal](#)

Publisher Rights Statement:

Catoni M., Zabet N.R. (2021) Analysis of Plant DNA Methylation Profiles Using R. In: Cho J. (eds) Plant Transposable Elements. Methods in Molecular Biology, vol 2250. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1134-0_21

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

1 **Analysis of plant DNA methylation profiles using R**

2 Marco Catoni¹ and Nicolae Radu Zabet²

3

4 **Affiliation:**

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

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

7

8 **Correspondence:**

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

10 **Radu Zabet:** nzabet@essex.ac.uk

11

12 **Running head:**

13 Analysis of plant DNA methylation profiles using R

14

15 **Abstract**

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

27

28 **Keywords:** Plant Epigenetics, Differentially Methylated Regions, cytosine methylation

29

30 Introduction

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

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

45 The analysis of the data produced by these strategies can be divided in two parts. The first part
46 includes well-established protocols of alignment to a reference genome, followed by calling
47 methylation levels at each cytosine position by comparing the number of covering reads supporting
48 the presence of methylation (read as cytosines) and the absence of methylation (read as thymine)
49 [5–7].

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

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

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

75

76 **Materials**

77 ***Cytosine methylation report***

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

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

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

86 Here is an example:

87	3	417	+	13	3	CG	CGT
88	3	418	-	6	0	CG	CGC
89	3	421	-	2	5	CHH	CAA
90	3	427	-	6	1	CHH	CAA
91	3	428	-	2	5	CHH	CCA
92	3	429	+	1	19	CHH	CCT
93	3	430	+	11	9	CHG	CTG
94	3	432	-	5	4	CHG	CAG
95	3	433	+	15	3	CG	CGT
96	3	434	-	9	2	CG	CGC

97
98 It is important to note that the protocol requires this seven-column text file and not a file specifically
99 generated by Bismark. This means that other tools can be used to perform the methylation call, such
100 as BS-Seeker [14, 15] or BSMAP [7], as long as the output of those tools is then converted to a text
101 file with the seven columns described above.

102 ***DMRcaller***

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

109 ***Tools to export DMRs from R***

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

115 **IGV**

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

118 **Workflow**

119 **Loading files**

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

- 123 - **context** – the context of the cytosine (CG, CHG or CHH).
- 124 - **readM** – the number of methylated reads (corresponding to the ‘count methylated’ field in
125 the CX_report file).
- 126 - **readN** – the total number of reads (the sum of ‘count methylated’ and ‘count unmethylated’
127 fields in the CX_report file).
- 128 - **trinucleotide context** - the specific context of the cytosine (as the corresponding field
129 reported in the CX_report file).

130 **Calculate conversion rate**

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

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

143 Load the cytosine report:

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

145

146 Extract the chloroplast methylation data:

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

148

149 Calculate conversion:

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

151

152 **Correction for conversion rate**

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

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

157 $m^* = \lfloor \max(0, m - n(1 - c)) \rfloor$

158

159 m^* = adjusted number of methylated reads per cytosine position.

160 m = original number of methylated reads per cytosine position.

161 n = total number of reads per cytosine position.

162 c = the conversion rate.

163

164 In R this can be implemented as:

165

```
166 CX_report_adjusted <- CX_report
```

```
167 CX_report_adjusted$readsM <- round(CX_report$readsM - CX_report$readsN * (1-
```

```
168 conversion))
```

```
169 CX_report_adjusted$readsM[CX_report_adjusted$readsM < 0 ] <- 0
```

170

171 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:

172 $n^* = \lfloor nc \rfloor$

173

174 n^* = adjusted number of total reads per cytosine position

175 n = original number of reads per cytosine position

176 c = estimated conversion rate

177

178 In R is implemented as:

179

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

181

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

183

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

185

186

187 **Generate bedGraph for genome browser visualization**

188 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),

195 First, the CX_report should be loaded in R:

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

197

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

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

200

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

202

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

204

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

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

207

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

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

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

213

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

218 **Computing the methylation frequency**

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

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

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

```
228 CX_report_cov <- CX_report[which(CX_report$readsN > 10)]
```

229

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

232 - Methylation percentage frequency for CG methylation

```
233 CX_report_CG <- CX_report_cov[CX_report_cov$context=="CG"]
```

```
234 CG_freq <- hist(100* CX_report_CG$readsM / CX_report_CG$readsN,
```

```
235 breaks=seq(0,100,by=10), plot=FALSE)
```

236

237 - Methylation percentage frequency for CHG methylation

```
238 CX_report_CHG <- CX_report_cov[CX_report_cov$context=="CHG"]
```

```
239 CHG_freq <- hist(100* CX_report_CHG$readsM / CX_report_CHG$readsN,
```

```
240 breaks=seq(0,100,by=10), plot=FALSE)
```

241

242 - Methylation percentage frequency for CHH methylation

```
243 CX_report_CHH <- CX_report_cov[CX_report_cov$context=="CHH"]
```

```
244 CHH_freq <- hist(100* CX_report_CHH$readsM / CX_report_CHH$readsN,
```

```
245 breaks=seq(0,100,by=10), plot=FALSE)
```

246

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

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

```
249 "#D55E00", "#CC79A7")
```

```

250 bar_counts <- rbind(100*CG_freq$counts/sum(CG_freq$counts),
251                    100*CHG_freq$counts/sum(CHG_freq$counts),
252                    100*CHH_freq$counts/sum(CHH_freq$counts))
253 rownames(bar_counts) <- c("CG", "CHG", "CHH")
254 colnames(bar_counts) <- paste0(CG_freq$breaks[1:(length(CG_freq$breaks)-1)], "-",
255                               CG_freq$breaks[2:length(CG_freq$breaks)])
256 barplot(bar_counts, xlab="% of methylation", beside = TRUE,
257         ylab=paste0("% of Cs"), las=1, ylim=c(0,100), col=cbbPalette[c(7,6,4)])
258 legend("topright", rownames(bar_counts), fill=cbbPalette[c(7,6,4)], bty="n")
259

```

260 Figure 2 shows that majority of CHH sites display low or lack of methylation (< 10%), while, for CG
261 sites , there is a large proportion of sites (approximately 40%) that display high level of methylation
262 (>80%). Finally, majority of CHG sites are unmethylated but there is a small proportion of sites
263 displaying intermediary and high level of methylation.

264

265 **Coverage calculation and spatial correlation**

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

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

```

271 wt <- DMRcaller::readBismark("GSM2384978_wt_processed.txt.gz")
272 met1 <- DMRcaller::readBismark("GSM2384979_met1-1_processed.txt.gz")
273
274

```

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

```

278
279 DMRcaller::plotMethylationDataCoverage(wt, met1, breaks = c(1,5,10,15),
280 conditionsNames=c("WT","met1-1"), context = c("CG", "CHG", "CHH"), labels=LETTERS)
281

```

282 This step allows to evaluate the sequencing depth and setup strategies for the downstream analysis.
283 In particular, for this dataset, we found that approximately 30-40% of the cytosines have at least 5-
284 10 reads (Figure 3), which means that calling differentially methylated cytosines might have missed
285 some true sites. Increasing the sequencing depth can partially solve this problem, but even highly
286 sequenced libraries will not lead to all cytosines having at least 10 reads (see **Note 3**). There are
287 several ways to computationally address this issue, and most of them assume merging several
288 cytosines and pooling together the reads in those regions. We will discuss several options in the
289 following sections below.

290 **Calculate DNA methylation in features.**

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

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

```

297 data(methylationDataList)
298 data(GEs)

```


299

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

303

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

307

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

310

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

314

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

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

317

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

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

321

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

324

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

331

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

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

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

343

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

345

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

347

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

```

349
350
351 We can now use the function to generate the plot:
352 par(cex=0.9)
353 par(mar=c(4, 4, 3, 1)+0.1)
354 DMRcaller::plotLocalMethylationProfile(methylationDataList[["WT"]],
355                                       methylationDataList[["met1-3"]],
356                                       chr3Reg,DMRsCGList,
357                                       conditionsNames = c("WT", "met1-3"),
358                                       GEs>windowSize = 300,main="CG methylation")
359

```

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

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

```

368 DMGenes <- genes[overlapsAny(genes, DMRs)]
369

```

370 **Call Differentially Methylated Regions (DMRs)**

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

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

379
380 An example of how to compute the DMRs in CG context with *noise_filter* method is:

```

381 DMRsNoiseFilterCG <- DMRcaller::computeDMRs(methylationDataList[["WT"]],
382                                             methylationDataList[["met1-3"]],
383                                             context = "CG", method = "noise_filter",
384                                             windowSize = 100, pValueThreshold = 0.01,
385                                             minCytosinesCount = 4, minProportionDifference = 0.4,
386                                             minGap = 200, minReadsPerCytosine = 4,
387                                             cores = 1)
388

```

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

```

390 DMRsBinsCHH <- DMRcaller::computeDMRs(methylationDataList[["WT"]],
391                                       methylationDataList[["met1-3"]],
392                                       context = "CHH", method = "bins", binSize = 100,
393                                       pValueThreshold = 0.01, minCytosinesCount = 4,
394                                       minProportionDifference = 0.1, minGap = 200,
395                                       minReadsPerCytosine = 4, cores = 1)
396

```

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

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

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

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

```
428 CX_CTRL_rep1 <- DMRcaller::readBismark("CX_CTRL_rep1.txt")  
429 CX_CTRL_rep2 <- DMRcaller::readBismark("CX_CTRL_rep2.txt")  
430 CX_TEST_rep1 <- DMRcaller::readBismark("CX_TEST_rep1.txt")  
431 CX_TEST_rep2 <- DMRcaller::readBismark("CX_TEST_rep2.txt")  
432
```

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

```
434 CX_all_data <- DMRcaller::joinReplicates(CX_CTRL_rep1, CX_CTRL_rep2)  
435 CX_all_data <- DMRcaller::joinReplicates(CX_all_data, CX_TEST_rep1)  
436 CX_all_data <- DMRcaller::joinReplicates(CX_all_data, CX_TEST_rep2)  
437  
438
```

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

```
440 condition_labels <- c("CTR", "CTR", "TEST", "TEST")  
441
```

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

```
444 DMRs_CG <- DMRcaller::computeDMRsReplicates(CX_all_data, condition =  
445 condition_labels, context = "CG", method = "bins")  
446
```

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

```
449 write.table(as.data.frame(DMRs_CG), file="DMRs_CG.txt", sep="\t", quote=F)  
450 rtracklayer::export(DMRs_CG, "DMRs_CG.gff3")  
451
```

452 **Call Differentially Methylated Cytosines (DMCs)**

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

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

```
461  
462 DMCs <- DMRcaller::computeDMRs(methylationDataList[["WT"]],  
463                               methylationDataList[["met1-3"]],  
464                               regions = chr_local, context = "CG",  
465                               method = "neighbourhood", test = "score",  
466                               pValueThreshold = 0.01, minCytosinesCount = 1,  
467                               minProportionDifference = 0.4, minGap = 0,  
468                               minSize = 1, minReadsPerCytosine = 4)  
469
```

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

474

475 **Plot DMRs on chromosomes**

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

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

```
485  
486 chr1 <- GRanges(seqnames = "1", ranges = IRanges(1, 30427671))  
487
```

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

```
490  
491 chr1_wt <- DMRcaller::computeMethylationProfile(wt, chr1,windowSize = 200000,  
492 context = "CG")  
493  
494 chr1_met11 <- DMRcaller::computeMethylationProfile(met11, chr1,windowSize = 200000,  
495 context = "CG")  
496
```

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

```
499  
500  
501 plot((start(chr1_wt) + end(chr1_wt))/2, 100*chr1_wt$Proportion, type="l", lty=1,  
502      lwd=2, col=cbbPalette[1], main="CG methylation on Chr 1", xlab="", xaxt="n",  
503      ylab="methylation percentage", ylim=c(0,100))  
504 lines((start(chr1_met11) + end(chr1_met11))/2, 100*chr1_met11$Proportion, lty=1,  
505      lwd=2, col=cbbPalette[6])  
506 legend("topright", c("WT", "met1-1"), col=cbbPalette[c(1,6)], bty="n", lty=1,  
507      lwd=2)  
508  
509
```

510 Notes

- 511 1. CX_report files are generated in Bismark by running the *bismark_methylation_extractor*
512 command and specifying the `--cytosine_report` and `--CX` options. For a detailed
513 description of the use of Bismark please refer to the user manual [6].
- 514 2. The corrected CX_report files can be directly downloaded from
515 ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM2384nnn/GSM2384978/suppl/GSM2384978_wt_processed.txt.gz and from
516 ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM2384nnn/GSM2384979/suppl/GSM2384979_met1-1_processed.txt.gz
517
- 518 3. One might ask why we need more reads covering each cytosine. There are two answers
519 to this: firstly, we need more reads to perform reliable statistical test to detect
520 differential methylation and secondly, the more reads we have the more robust we can
521 call the actual methylation level when this has intermediary values (e.g., we need at
522 least 4 reads to call a cytosine being methylated in 75% of the cases).
523
- 524 4. There are different methods to call DMRs and it appears that the method used should be
525 selected depending on the methylation context, coverage and tissues used to generate
526 the data. The DMRcaller tool implements three methods to call DMRs and the
527 performances of each of them has been previously discussed [10].
- 528 5. If a binary methylation change is expected (i.e. regions pass from being highly
529 methylated to a complete unmethylated status) as often happens for methylation in CG
530 context in plants, a higher value of this parameter helps to reduce noise generated by
531 random changes. By contrary, limited variations in methylation (more common for CHH
532 context) require a lower value of this parameter to allow detection of small changes.
- 533 6. This setting applied to the `minGap` parameter can be useful in case of multiple sample
534 comparisons, due to the fact that the number of DMRs found in each comparison is
535 directly informative of the portion of genome with methylation difference. Therefore, if
536 `minGap` is set to 0, the DMR lists would not be required to be normalised by their length
537 when compared across samples.

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

552 **Acknowledgment**

553 We thank Ms. Jessica Scivier (University of Birmingham, UK) for critical reading and proofreading of
554 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, Downen 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 KY, 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-](https://doi.org/10.1016/S0168-9452(00)00411-8)
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

630 **Figure captions**

631 **Fig. 1**

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

640 **Fig. 2**

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

645

646 **Fig. 3**

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

651

652 **Fig. 4**

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

662

663 **Fig.5**

664 Low resolution methylation profile along chromosome 1 of *Arabidopsis thaliana* wild type and *met1-*
665 *1* mutant [26], obtained by merging cytosine methylation in CG context in windows of 200 kb size.
666 Highest methylation levels are located at centromeres and pericentromeric regions.

667

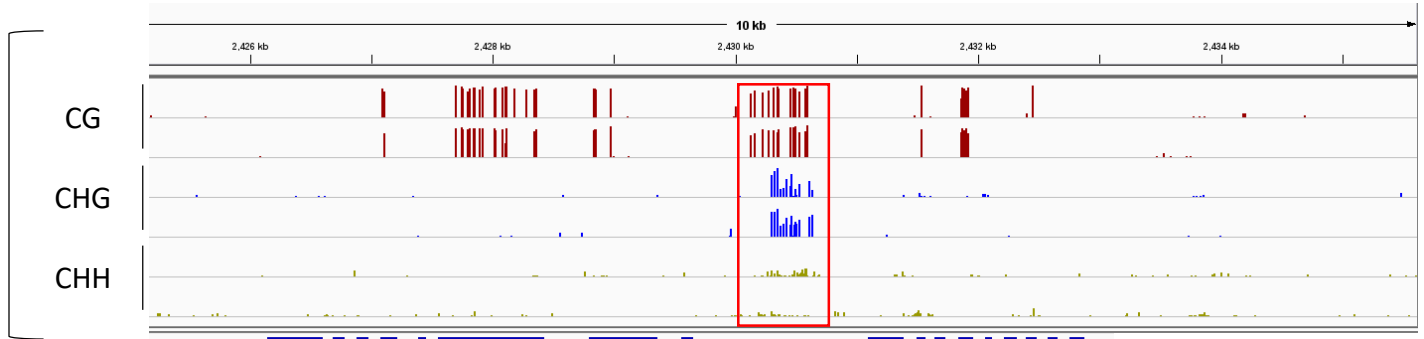
668

669

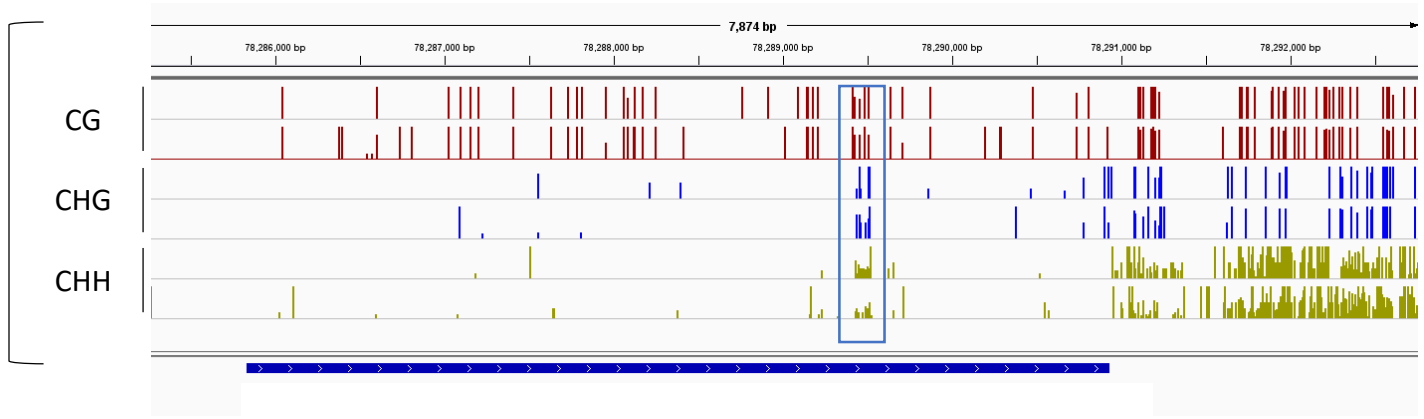
670

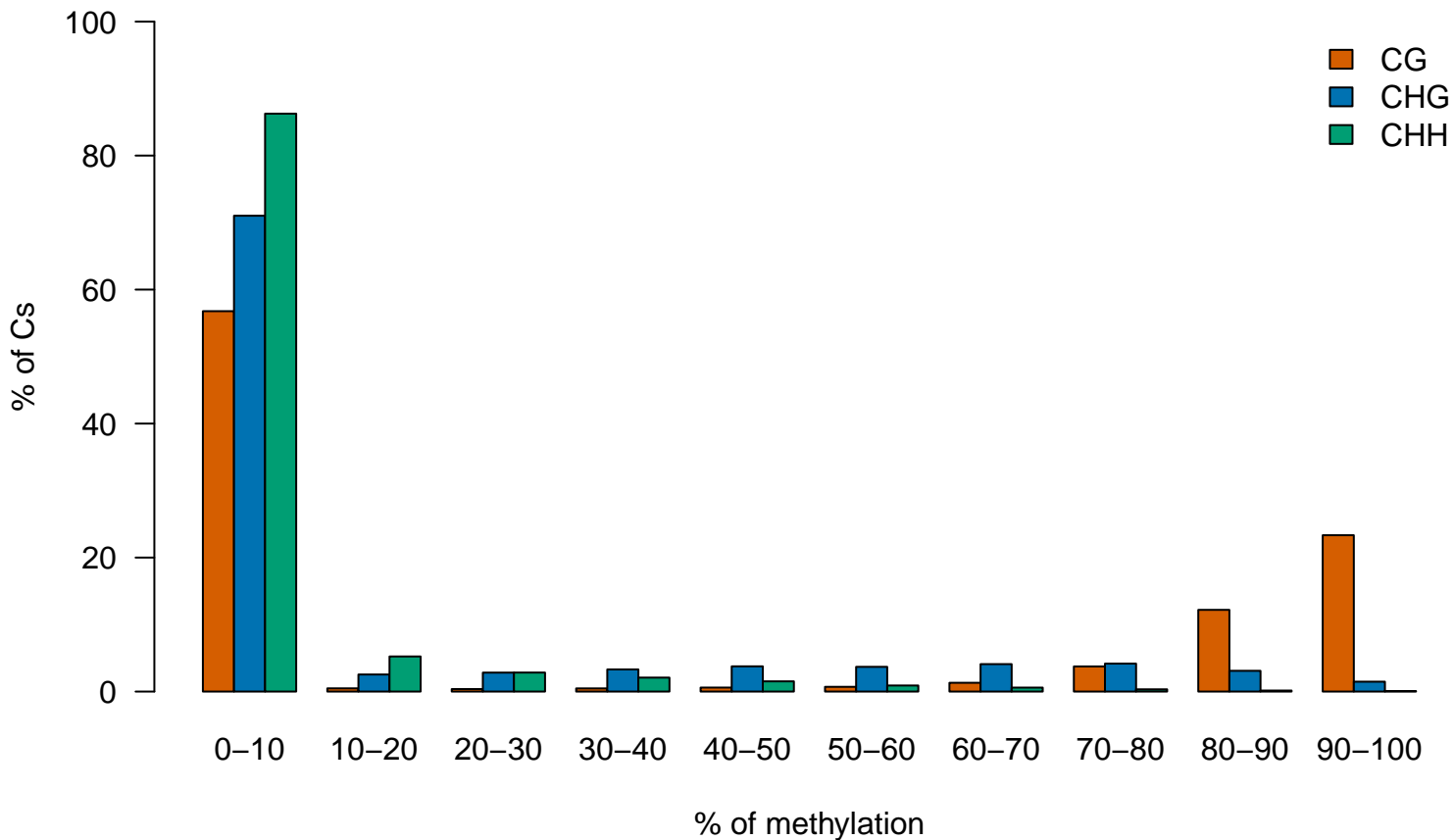
671

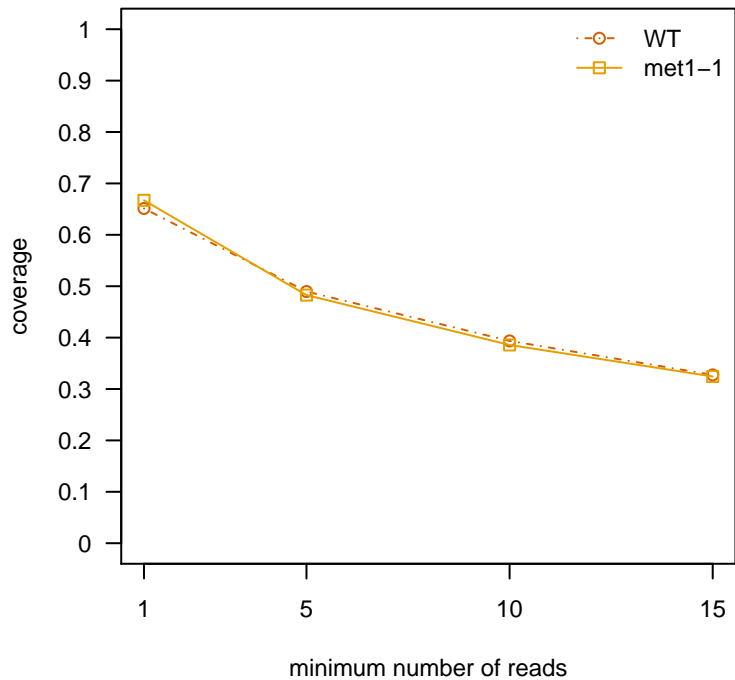
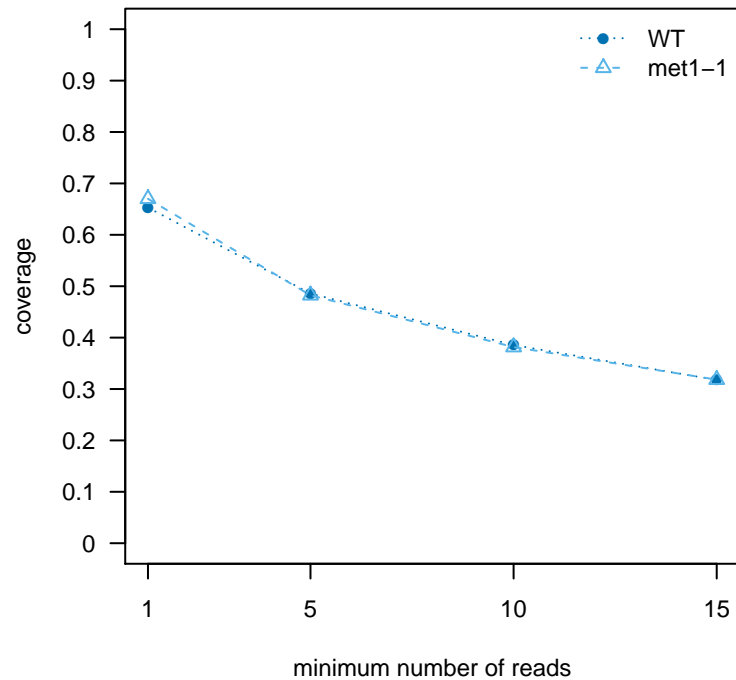
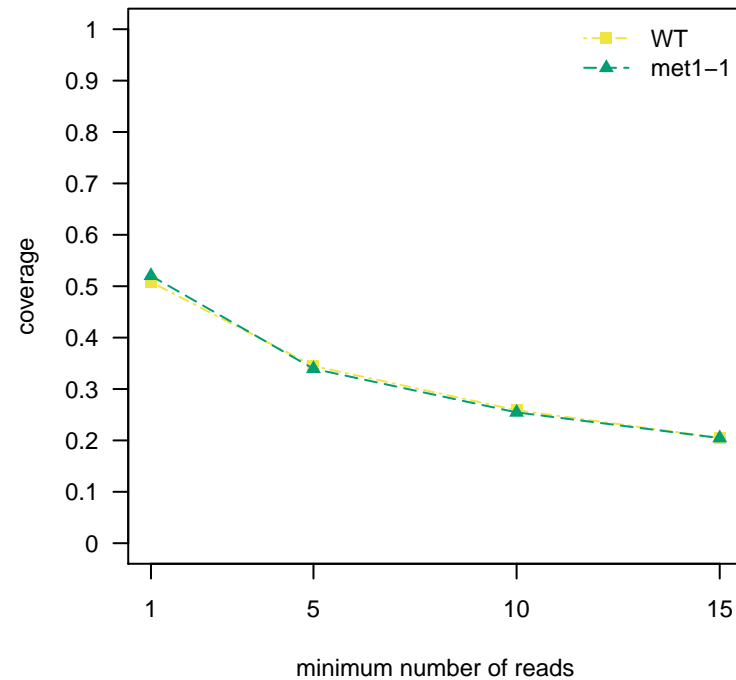
A. thaliana



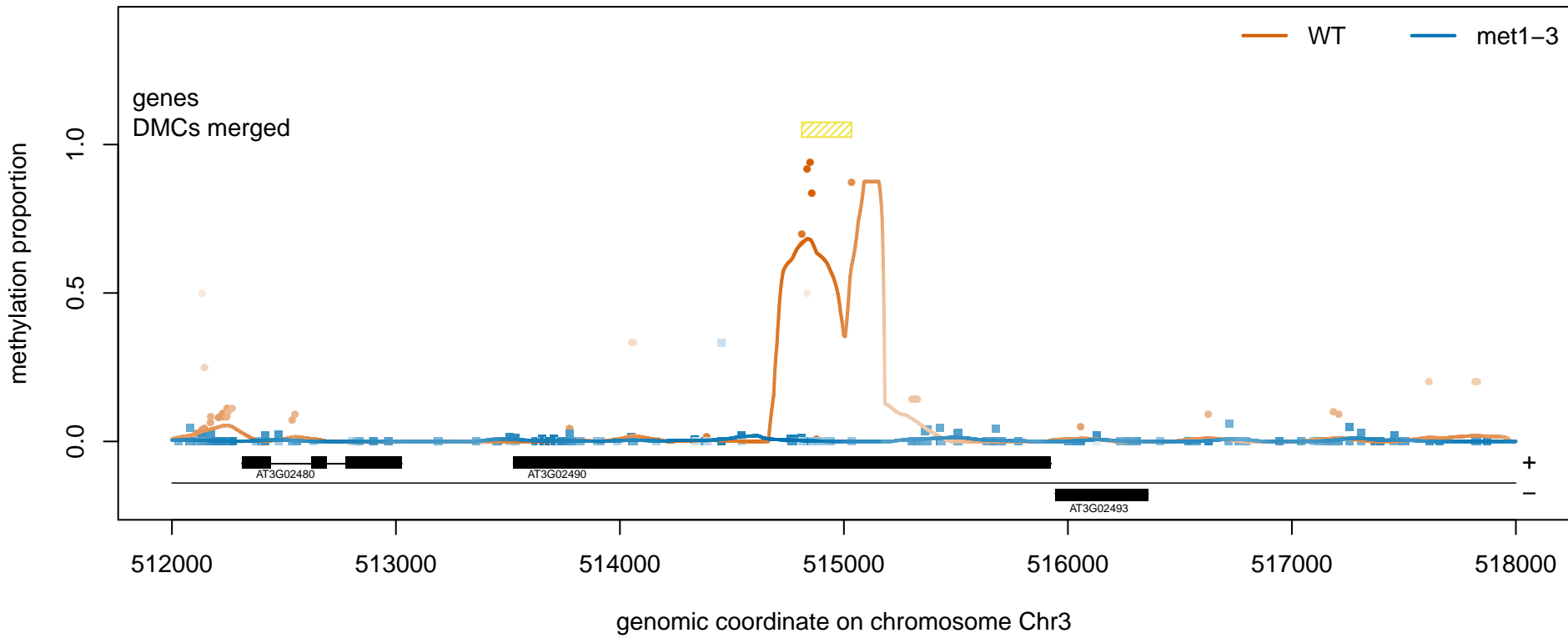
S. melongena





A Coverage in CG context**B Coverage in CHG context****C Coverage in CHH context**

CG methylation



CG methylation on Chr 1

