# A Blind and Independent Benchmark Study for Detecting Differentially Methylated Regions in Plants - Supplementary Information -

Clemens Kreutz S. Nilay Can Ralf Schulze Bruening Rabea Meyberg
Zsuzsanna Mérai Noe Fernandez-Pozo Stefan A. Rensing

## 1 Aethionema arabicum sample preparation and data generation

*Ae. arabicum* (Brassicaceae), represents one of the angiosperm species from this study. Dry seed samples from two different ecotypes (Turkey and Cyprus) of Ae. arabicum were processed (Haudry *et al.*, 2013; Mohammadin *et al.*, 2018). Seeds were derived from maternal plants grown at 20° C or 25° C. Genomic DNA was extracted using the protocol from (Qin *et al.*, 2009) with some modifications. Briefly, 30 mg dry seed was ground in liquid nitrogen. After adding 700 $\mu$l of the extraction buffer (100 mM Tris, pH 8.0, 10 mM EDTA, 2% SDS, 1.5% PVP, 3 U Proteinase K), samples were incubated at 60° C for 60 minutes, then supplemented with NaCl and CTAB to the final concentration of 1.4M and 0.2%, respectively, and further incubated at 65° C for 10 minutes. Equal volume of chloroform:isoamylalcohol (24:1, v/v) was added and mixed gently, incubated at 0° C for 30 min, and centrifuged at 4° C, 13,000g for 10 minutes. The supernatant was treated with RNase at 37° C for 30 minutes and centrifuged again at 4° C, 13000g for 10 min. DNA was precipitated with 0.55 volume of isopropanol, washed with 70% ethanol and dried at ambient temperature. To remove remnants of the gel-like seed mucilage, additional purification step was applied using *DNA Clean & Concentration columns* (Zymo Research). The DNA concentration was quantified using *Qubit Fluorometer* (Thermo Fischer Scientific) and the DNA quality was checked by gel electrophoresis. Library preparation was performed from 80ng gDNA using the *Pico Methyl-SeqTM Kit* (Zymo Research). Each library was further size-selected using *AMPure XP beads* (Beckman Coulter) keeping the fragment range between 100-700 bp. Sequencing was performed in the Vienna Biocenter Core Faculty (www.vbcf.ac.at). 100 bp fragments were obtained in single-end mode with Illumina *HiSeq 2500* according to standard protocols of the manufacturer for cluster generation and sequencing. After trimming the bisulfite treated reads with *Trimmomatic* 0.36 (Bolger *et al.*, 2014), data were mapped with *Bismark* v0.17.0 (Krueger and Andrews, 2011) with multicore 3, -N 1 and score min L,-1 -1 parameters using the v2.5 reference genome. Sequence data from *Aethionema arabicum* can be found in *CoGe* database (https://genomevolution.org/coge/) under the following id: v2.5, id33968. SAM alignment files were sorted and merged with *Samtools* v1.3.1 (Li *et al.*, 2009).

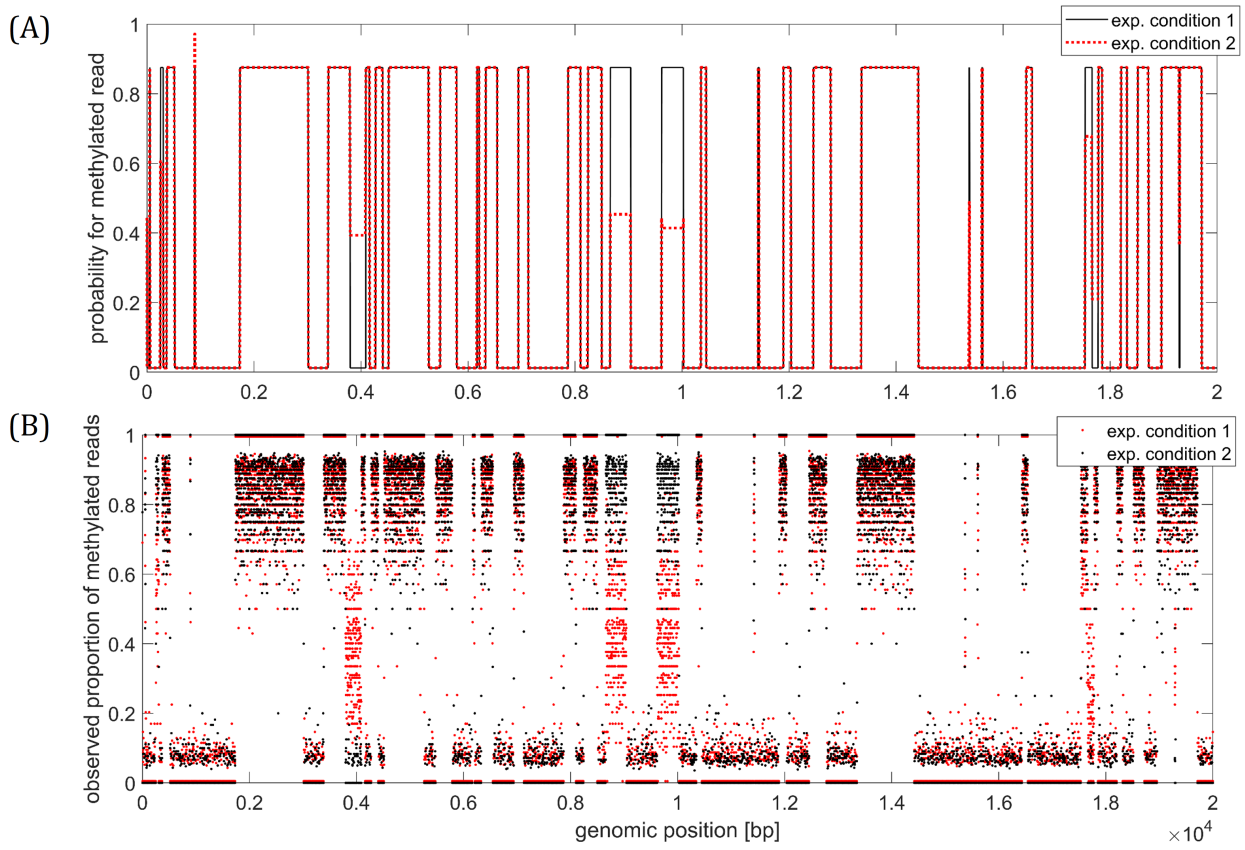



Figure 1: **Illustration of simulating methylation differences.** Panel (A) shows an example for simulated methylation differences using data set Phypa-G-CG as template. In the reference sample (black line), the probability of methylated reads methylated region is 0.8834, in unmethylated regions 0.0171. For simulation of methylation differences in a case sample, the changes for these probabilities are uniformly drawn for the subset of differentially methylated regions (red dotted line). Panel (B) shows methylation frequencies which were simulated accordingly.

## 2 Illustration of methylation differences

Figure 1 illustrates how methylation difference are simulated if a case sample (red dotted line) is compared with a control sample (black line). The probabilities for measuring a read indicating methylation is controlled by the simulation parameters $P_{\text{meth}}$ for methylated regions and $P_{\text{un}}$ for unmethylated regions. As a first step, regions with differential methylation in the case sample are drawn. Then the change in probability $\Delta$ is drawn from a uniform distribution in the range $[0, 0.5]$ as well as the direction of regulation. The red dashed line in panel (A) illustrates the resulting differences which results in measurements like depicted in panel (B).

# 3 Smoothing BS-seq data

For discriminating regions with low methylation levels from regions with higher methylation levels, a custom smoothing approach is applied which is summarized and illustrated in Figure 2. In order to discriminate low and high methylation levels, first two thresholds which are shown as horizontal dashed lines in panel (A) are calculated from the data. As upper threshold, we used the minimum of 80% and of the 80% percentile of the fraction of methylated reads. Similarly, the maximum of 20% and the 20% percentile is defined as lower threshold. These two thresholds classify each methylation site into low (blue dots), intermediate (gray dots), or high (red dots) methylation level. Positions with intermediate levels were then assigned to unmethylated or methylated according to the closest read which are beyond the thresholds and thereby already assigned as low or high level. This procedure subdivides the genomic range into measured low and high methylation levels, but in a rather non-smooth manner as shown in panel (B).

We then used the *LOWESS (Local regression using weighted linear least squares* implementation in Matlab with three different smoothing spans $10, 20, 50$ to obtain smooth averages of the local methylation levels. For this step, we did not account for position distances, i.e. LOWESS smoothing is always based on 10, 20, or 50 consecutive methylation sites. The three curves, plotted as black lines in panel (C), are then merged into a single curve by taking the minimum or maximum, depending on the methylation state from the previous step (gray shading in (C)). For low levels, the minimum over the three curves was taken, for high levels the maximum which results in the red curve shown in panel (C).

The smoothed curve is then used to define new thresholds by taking the 20% and 80% quantiles (panel (D)). Next, the data points are again assigned to low and high methylation levels like in step (A). Again, data in the intermediate range between the two thresholds were assigned with respect to being closer to a low or high methylation region (panel (E)). This finally leads to a smooth subdivision of the genomic range into presumably methylated (termed PM and highlighted by red shading) and presumably unmethylated (termed PU) regions. This assignment is then used to calculate data attributes for characterizing methylated and unmethylated regions as indicated in Figure 3. The two thresholds/quantiles plotted in Figure 2 in panels (D) and (E) are also utilized as two out of 16 data attributes. These 16 data attributes were considered during calibration of the simulation parameters to perform realistic simulations.

The approach is available as part of the implementation of our decision tree at the github repository https://github.com/kreutz-lab/DMR-DecisionTree.

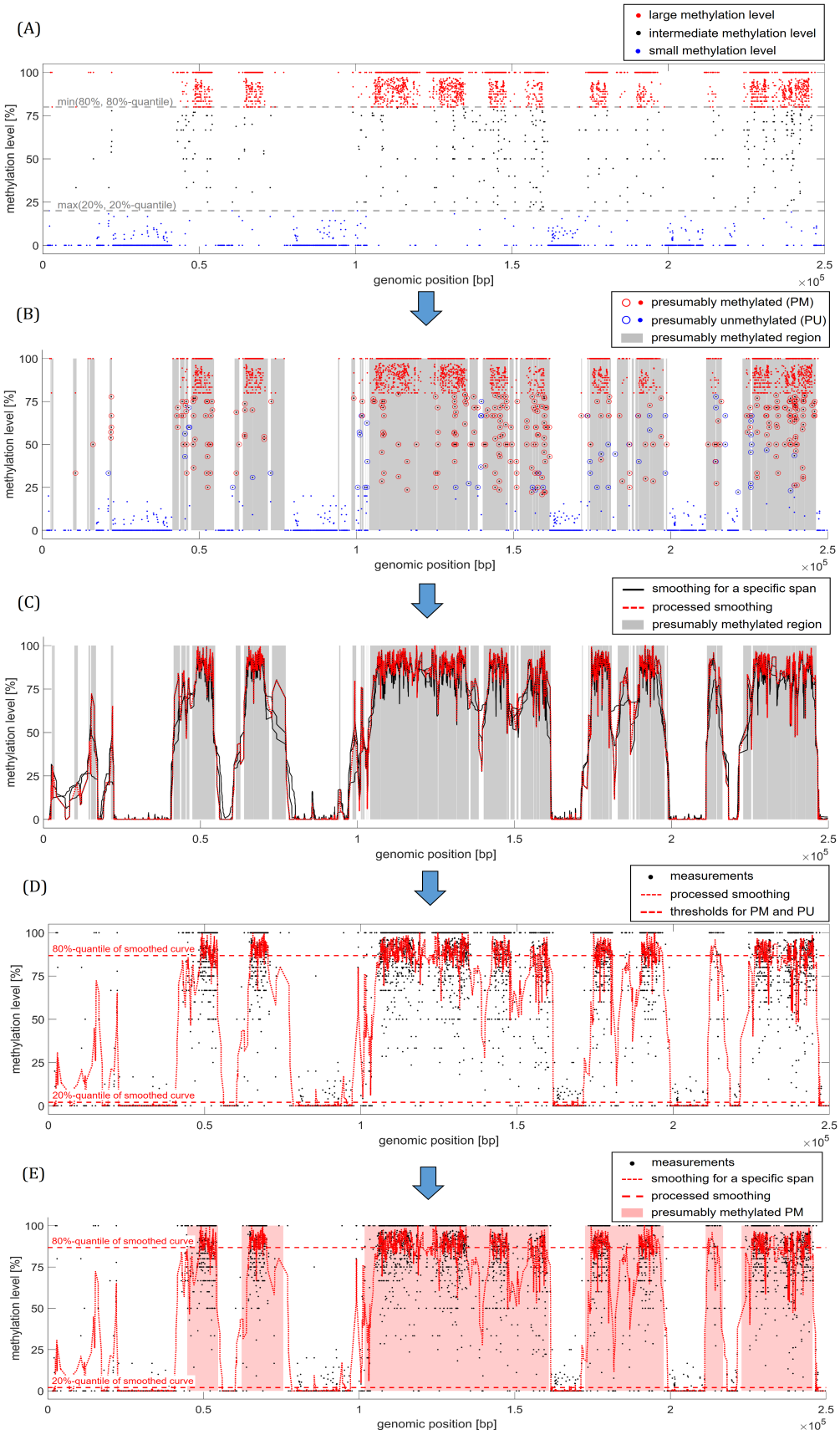



Figure 2: Illustration of the smoothing approach applied for discriminating regions which are presumably methylated (PM) from presumably unmethylated (PU) regions. First, two thresholds are used to define methylation sites with low and large methylation levels (A). Then, sites with intermediate levels are assigned according to the closest site with low or large levels (B). The gray background indicates methylated regions. The three smoothed curves (black lines) in panel (C) are then processed, i.e. the minimum or maximum of the three curves is chosen depending on methylation state. The 20% and 80% quantiles are then calculated to define new thresholds for methylated and unmethylated. Sites between both thresholds are again assigned according to the closest site above or below the threshold. Red shading indicates the finally obtained classification into presumably methylated.

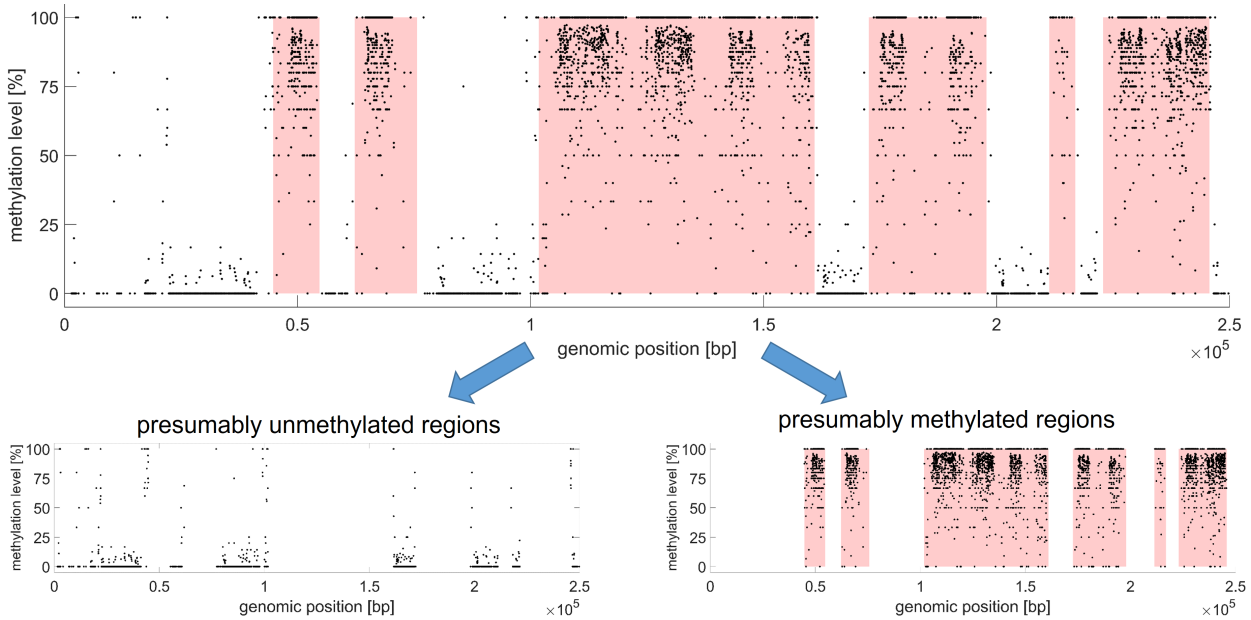
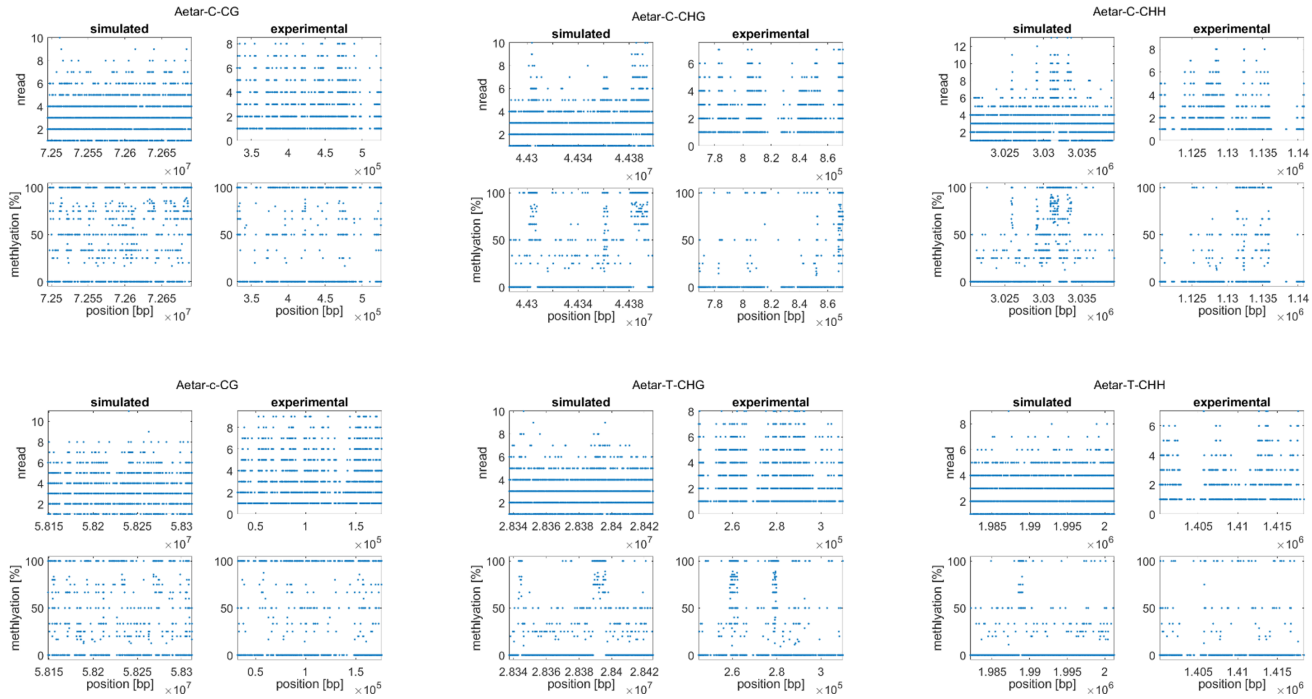



Figure 3: Subdividing the genome into presumably low and high methylation levels enables calculation of data attributes for both cases. As written in the main text, these attributes comprise means and standard deviations of the number of reads and of the measured methylation levels. Moreover, the mean and standard deviation of the lengths of these regions is evaluated as data attributes.

# 4 Characteristics of the experimental data, stochasticity and replicates

Figure 4 and Figure 5 show representative examples of the experimental and simulated data. The data from *Ae. arabicum* has the smallest coverage. *Ae. arabicum* has a much smaller coverage than *A. thaliana*. Thus, the number of reads and the methylation levels exhibit discretization. The experimental number of reads measured from *A. thaliana* has a rather large variance and seems to show spatial correlations in particular for *Arath-CHH* (lower row, right panel). Such correlations can only be partly imitated by the *WGBSSuite* by distinguishing presumably methylated and unmethylated regions. For some other data sets, the local coverage seems to be correlated with the local methylation levels, e.g. for *P. patens* (Figure 5, panel B).

## (A) Ae. arabicum




## (B) A. thaliana

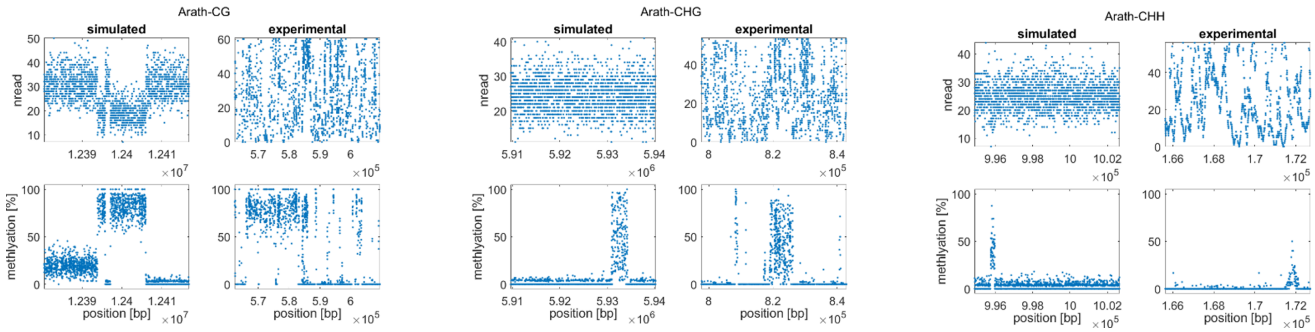



Figure 4: Representative comparison of experimental and simulated data for *Ae. arabicum* in panel (A) and *A. thaliana* in panel (B). Regions of 2000 consecutive read positions were randomly drawn for each experimental and each simulated data set.

These plots nicely illustrate that our BS-seq data cover a broad range of different data characteristics. Moreover, it might also indicate that it is challenging, to properly choose the 13 simulation parameters of the *WGBSSuite*. In order to generate the data as realistically as possible, we considered several attributes of the observed number of reads, the corresponding methylation levels and the distances. As explained in the main part of the paper, 16 attributes were calculated for characterizing the experimental data in presumably methylated (PM) and presumably unmethylated (PU) regions:

1. Mean (M) and standard deviation (SD) of the number of reads $n$.
2. Mean and standard deviation of the methylation levels $Meth$

3. Mean and standard deviation of the $\log_{10}$ distances of read positions $dPos$
4. Mean and standard deviation of the $\log_{10}$ lengths of presumably methylated (PM) regions according to our smoothing approach
5. Mean and standard deviation of the $\log_{10}$ lengths of presumably unmethylated (PU) regions
6. Smoothing threshold (ST) for the number of reads
7. Smoothing threshold for the methylation level

These attributes were used to optimize the configuration parameters of the *WGBSSuite*. Figure 6 illustrates similarities of the experimental data sets at the level of these attributes. The heatmap has been generated by hierarchical clustering based on euclidean distance using the `amap` R-package (Lucas, 2018).

The largest difference occurs between the samples generated for *CHH* context and the other two contexts. The *CHH* motif occurs more frequently than *CHG* and *CG* in the DNA. This has an impact on several data attributes that are related to the density of the read positions. The decrease in the methylation levels that is observed for *CHH* might be related to strand-specific methylation.

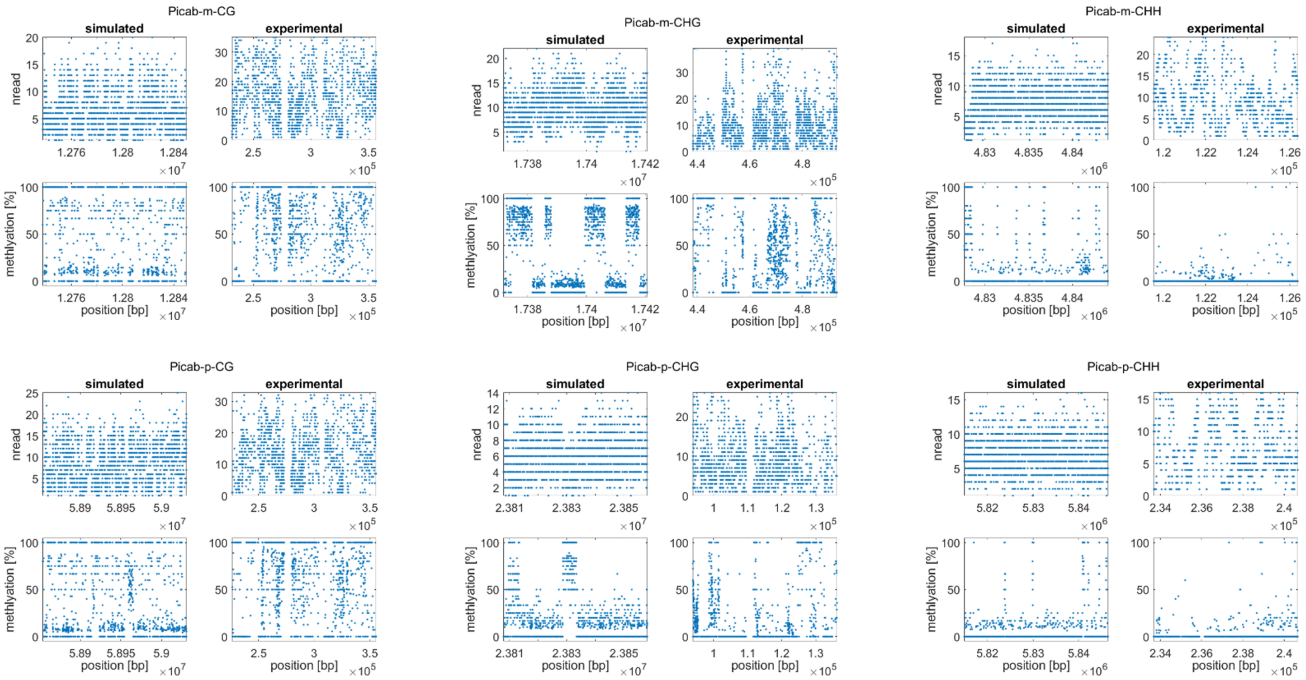
For *Ae. arabicum*, *P. abies* and *P. patens* we had data from two biological conditions for each DNA context. Those two samples always cluster together which indicates that in our data sets the differences between organisms are larger than difference between biological conditions.

Figure 7 shows the relationship between the means and standard deviations of the number of reads in panel (A) and of the methylation levels in panel (B) for simulated data (gray) and the experimental data (black). For this depiction, 20 consecutive positions were randomly drawn 100 times for each of the 21 data templates. Our simulation data covers the whole range that occurs for the experimental data. Panel (A) also indicates that the simulated data has a slightly increased standard deviation for the local coverage.

Our study is based on the 21 experimental data sets summarized in Tables 1 in the main text. These data sets enables characterization of the variability of BS-seq data over the evaluated biological conditions. Stochasticity of the simulated data is introduced by randomly drawing the positions and lengths of CpG islands, the positions of the DNA motifs (CG, CHG or CHH), as well as the lengths and locations of methylated regions and random selection of the subset of methylated sites.

Because the data does not comprise replicates of the same biological conditions, it is not possible to evaluate the variability over real biological replicates. Therefore, the stochasticity between replicates of our simulated reads exclusively originates from the Poisson distribution. Moreover, random subset of these reads are drawn as being methylated according to the binomial distribution. Both distributions are commonly assumed for sequencing data and are implemented in the *WGBSSuite* (Rackham *et al.*, 2015) that has been developed for simulating BS-Seq data.

## (A) P. abies




## (B) P. patens

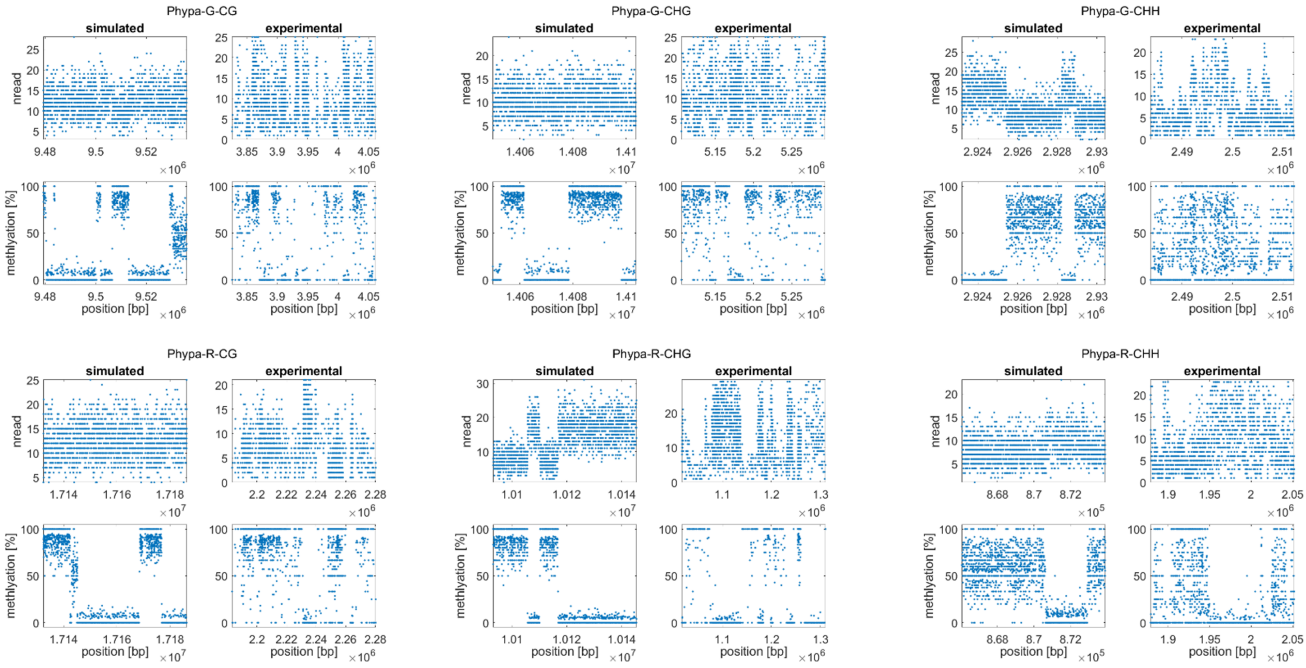



Figure 5: Representative comparison of experimental and simulated data for *P. abies* in panel (A) and *P. patens* in panel (B). For plotting purpose, regions of 2000 consecutive read positions were randomly drawn for each experimental and each simulated data set. For *P. abies* and *CHH* context, there are only short and weakly methylated regions. For the *CG* context, methylation is very heterogeneous and it is hardly possible to distinguish methylated and unmethylated regions by eye. For *P. patens*, all three DNA contexts display regions with increased and decreased methylation levels.

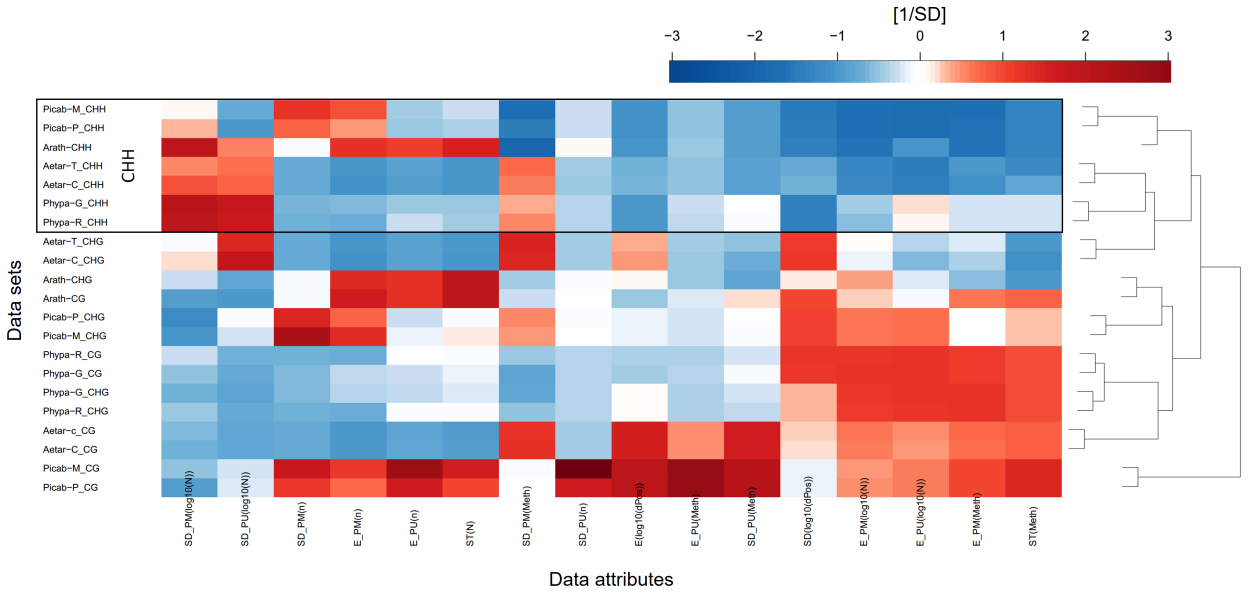



Figure 6: Clustering of the data attributes illustrates differences and similarities of the experimental data sets. All samples from the *CHH* context cluster in one branch that is highlighted by the black box. If two biological conditions were evaluated for the same organism, the two *CG* and *CHG* samples always cluster together. The data attributes have been standardized for this depiction. SD = standard deviation, E = expectation (arithmetic mean), n = number of reads, Meth = measured methylation levels, PM = presumably methylated region, PU = presumably unmethylated region, ST = smoothing threshold, N = number of read positions in the region.

(A) Mean-SD relationship for the number of reads

(B) Mean-SD relationship for methylation levels

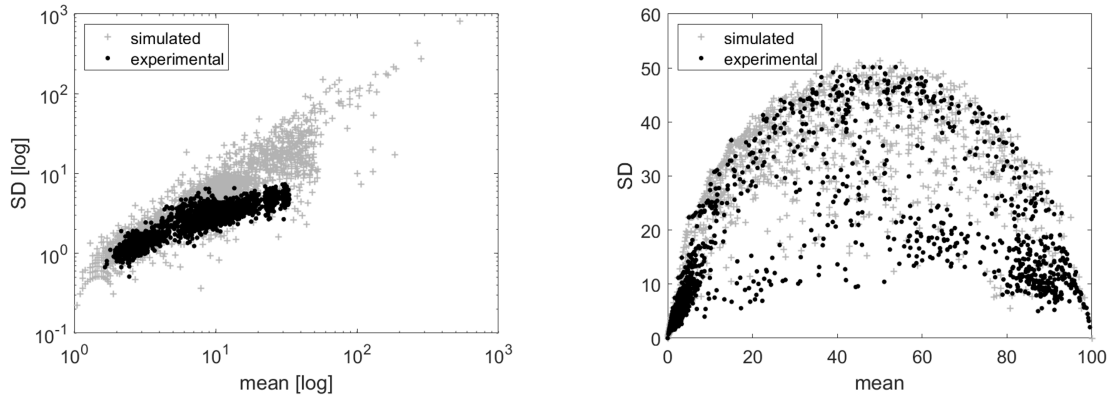



Figure 7: Relationship of means and standard deviations for randomly drawn consecutive data points. Panel (A) on the left shows the relation for the number of reads on the log-scale, the plots on the right depicts the outcome for methylation levels.

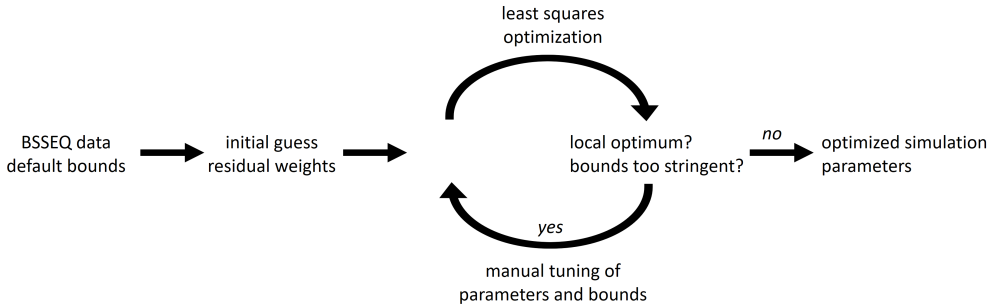



Figure 8: Illustration of the workflow applied for calibrating simulation parameters. An iterative cycle of deterministic least-squares optimization and manual tuning/perturbation of the parameters and the bounds was applied to prevent local optima and to avoid too stringent bounds. The procedure is repeated until the simulated data was as similar as possible to the experimental data.

# 5 Optimization of Simulation Parameters

Table 1: Overview about the simulation parameters available in the *WGBSSuite* for generating BS-seq data. 13 parameters $\theta_1, \ldots, \theta_{13}$ were estimated individually for each experimental data set and used for simulating realistic data.

| Parameter description | Symbol or value | Fitting options |
|---|---|---|
| Number of methylation sites | $1e6$ | Fixed value |
| Balance between up- and down-regulation | 0.5 | Fixed value |
| Magnitude of differential methylation | $U(0, 2)$ | Uniformly drawn for simulations |
| Distribution for observed methylation status | "binomial" | Fixed |
| Success probability for methylated sites | $\theta_1 := P_{\text{meth}}$ | Bounds: [0.6, 1] |
| Success probability for unmethylated sites | $\theta_2 := P_{\text{un}}$ | Bounds: [0, 0.3] |
| Error rate for methylated sites | $\theta_3 := E_{\text{meth}}$ | Bounds: [0.01, 4], fitted on log-scale |
| Error rate for unmethylated sites | $\theta_4 := E_{\text{un}}$ | Bounds: [0.03, 0.1], fitted on log-scale |
| Mean number of reads in methylated regions | $\theta_5 := M_{\text{meth}}$ | Bounds: [1, 30], fitted on log-scale |
| Mean number of reads in unmethylated regions | $\theta_6 := M_{\text{un}}$ | Bounds: [1, 30], fitted on log-scale |
| Rate of methylation sites in "CpG" islands | $\theta_7 := R_{\text{island}}$ | Bounds: [0.1, 2], fitted on log-scale |
| Rate of methylation sites in deserts | $\theta_8 := R_{\text{desert}}$ | Bounds: [0.001, 0.2], fitted on log-scale |
| Probability decay for methylation state | $\theta_9 := d$ | Bounds: $[10^{-0.2}, 10^{1}]$, fitted on log-scale |
| Transition probabilities desert $\rightarrow$ island | $\theta_{10} := T_{12}$ | Bounds: [0.05, 0.8], fitted on log-scale |
| Transition probabilities island $\rightarrow$ desert | $\theta_{11} := T_{21}$ | Bounds: [0.02, 0.4], fitted on log-scale |
| Transition probabilities unmethylated $\rightarrow$ methylated | $\theta_{12} := \Pi_{12}$ | Bounds: $[10^{-3}, 10^{-0.5}]$, fitted on log-scale |
| Transition probabilities methylated $\rightarrow$ unmethylated | $\theta_{13} := \Pi_{41}$ | Bounds: $[10^{-3}, 10^{-0.5}]$, fitted on log-scale |

As described in the main text, the simulation parameters were calibrated by optimizing an objective function assessing similarity of simulated and experimental data. Optimization of the simulation parameters faces several serious limitations and challenges. First, the objective function is stochastic because the simulated data sets are generated by randomly drawing the frequency and positions of methylation sites, the length and frequency of *CpG* islands and the total number reads and proportion of methylated reads. We reduced this issue by initializing the random number generator for each simulation in the same manner. At the same time, the data sets were chosen large enough to ensure that the chosen random seed does not generate a non-representative special case.

Another limitation is that the objective function $V(\vec{\theta})$ is non-continuous because *CpG* islands as well as regions which are methylated/unmethylated and differentially methylated are intervals which non-continuously change if the simulation parameters are altered. Moreover, because of the non-linear and non-continuous dependency of the objective function $V(\vec{\theta})$ on the parameters $\theta$, it is also not possible to calculate derivatives $\nabla V(\vec{\theta})$ of the objective function.

Several optimization algorithms were tested which do not require derivative information (e.g. patternsearch and genetic algorithms) but they showed weak performance. We therefore applied deterministic non-linear optimization with on a trust-region based nonlinear least squares optimization algorithm implemented in Matlab's optimization toolbox in function `lsqnonlin`. To prevent local optima, we alternatingly performed an automatic optimization and manual tuning. Moreover, the parameters bounds are extended adaptively if the initial bounds emerged as too stringent. The overall optimization workflow including the iterative cycle between automatic and manual optimization is summarized in Figure 8.

Manual inspection was performed by

- plotting experimental and simulated data along the genome,

- plotting histograms for methylation levels and for distances of consecutive methylation sites,
- evaluating the fraction of methylation sites in islands/deserts and the occupancy of the four methylation states used in the *WGBSSuite*,
- evaluating the effect of manual parameter changes on the objective function $V$

This inspection cannot be automatically utilized for changes of the simulation parameters of the *WGBSSuite* but provided information for manually reducing disagreement between measurements and simulation data. Table 2 shows the finally obtained optimized simulation parameters for the individual data sets.

Table 2: Overview about the estimated simulation parameters for each individual data sets. The partly uncommon column names indicate the argument names in the *WGBSSuite* as well as to the mathematical notation in this paper.

| Data set | ProbSuccinMeth $P_{\text{meth}}$ | ProbSuccinNonmeth $P_{\text{un}}$ | Err Meth $E_{\text{meth}}$ | Err Non-meth $E_{\text{un}}$ | Mean Meth $M_{\text{meth}}$ | Mean Nonmeth $M_{\text{un}}$ | DenseInIsland, cpgmatrix(2), $R_{\text{island}}$ | DenseInDesert, cpgmatrix(1), $R_{\text{desert}}$ | distvalue $d$ | StayInIsland, transCp-gLoc(3), $T_{22}$ | LeaveDesert, transCp-gLoc(1), $T_{12}$ | LeaveMeth, transPi(1) $\Pi_{12}$ | LeaveNonmeth, transPi(3), $\Pi_{34}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Phypa-G-CG | 0.88 | 0.012 | 0.049 | 0.054 | 9.4 | 11.6 | 0.62 | 0.021 | 0.27 | 0.072 | 0.14 | 0.0018 | 0.0011 |
| Phypa-R-CG | 0.87 | 0.011 | 0.085 | 0.051 | 10.9 | 11.9 | 0.64 | 0.024 | 0.43 | 0.061 | 0.17 | 0.0007 | 0.00059 |
| Phypa-R-CHG | 0.87 | 0.01 | 0.11 | 0.051 | 6.5 | 15.9 | 0.47 | 0.027 | 0.37 | 0.052 | 0.16 | 0.00071 | 0.00038 |
| Phypa-R-CHH | 0.6 | 0.026 | 0.026 | 0.053 | 6.8 | 8.9 | 0.96 | 0.19 | 0.34 | 0.094 | 0.28 | 0.00071 | 0.00029 |
| Phypa-G-CHH | 0.66 | 0.0021 | 0.23 | 0.067 | 7.5 | 13.8 | 0.85 | 0.2 | 0.75 | 0.14 | 0.28 | 0.00032 | 0.00018 |
| Phypa-G-CHG | 0.9 | 0.011 | 0.11 | 0.06 | 9.8 | 8.9 | 0.53 | 0.024 | 0.42 | 0.063 | 0.17 | 0.0011 | 0.00086 |
| Aetar-C-CG | 0.93 | 0.042 | 0.035 | 0.051 | 2.8 | 1.7 | 0.34 | 0.0065 | 0.33 | 0.054 | 0.11 | 0.015 | 0.0094 |
| Aetar-c-CG | 0.94 | 0.043 | 0.052 | 0.055 | 1.9 | 2.3 | 0.47 | 0.0021 | 0.24 | 0.034 | 0.35 | 0.026 | 0.011 |
| Aetar-C-CHH | 0.85 | 0.019 | 0.052 | 0.049 | 5.3 | 1.6 | 0.55 | 0.062 | 0.69 | 0.18 | 0.19 | 0.0053 | 0.00037 |
| Aetar-T-CHH | 0.86 | 0.024 | 0.064 | 0.054 | 1.3 | 1.5 | 0.28 | 0.085 | 0.61 | 0.037 | 0.13 | 0.0021 | 0.00018 |
| Aetar-C-CHG | 0.95 | 0.028 | 0.048 | 0.058 | 3.3 | 1.2 | 0.54 | 0.015 | 0.18 | 0.16 | 0.16 | 0.0033 | 0.00064 |
| Aetar-T-CHG | 0.93 | 0.032 | 0.049 | 0.059 | 2.8 | 1.7 | 0.69 | 0.018 | 0.32 | 0.39 | 0.15 | 0.0039 | 0.00055 |
| Picab-P-CG | 0.94 | 0.036 | 0.043 | 0.057 | 3.0 | 10.7 | 0.45 | 0.011 | 1.0 | 0.054 | 0.036 | 0.00054 | 0.00049 |
| Picab-M-CG | 0.95 | 0.036 | 0.033 | 0.054 | 3.0 | 9.0 | 0.19 | 0.017 | 0.47 | 0.13 | 0.048 | 0.0066 | 0.011 |
| Picab-M-CHG | 0.8 | 0.033 | 0.18 | 0.061 | 7.0 | 10.5 | 1.0 | 0.035 | 0.55 | 0.8 | 0.042 | 0.0025 | 0.0011 |
| Picab-P-CHG | 0.83 | 0.05 | 0.25 | 0.044 | 3.5 | 5.4 | 1.8 | 0.027 | 0.39 | 0.77 | 0.11 | 0.0039 | 0.001 |
| Picab-M-CHH | 0.6 | 0.006 | 0.066 | 0.059 | 2.6 | 6.7 | 0.62 | 0.11 | 0.45 | 0.14 | 0.1 | 0.032 | 0.0027 |
| Picab-P-CHH | 0.87 | 0.015 | 0.034 | 0.077 | 2.5 | 5.9 | 0.37 | 0.053 | 1.2 | 0.056 | 0.31 | 0.016 | 0.00031 |
| Arath-CG | 0.82 | 0.012 | 0.31 | 0.079 | 17.8 | 30.0 | 1.4 | 0.04 | 0.43 | 0.61 | 0.19 | 0.00082 | 0.00033 |
| Arath-CHG | 0.44 | 0.0063 | 1.3 | 0.055 | 23.3 | 23.3 | 1.8 | 0.048 | 0.2 | 0.49 | 0.21 | 0.0013 | 0.00047 |
| Arath-CHH | 0.21 | 0.022 | 0.26 | 0.057 | 18.0 | 24.2 | 0.25 | 0.26 | 0.49 | 0.6 | 0.23 | 0.0013 | 0.00018 |

# 6 Application of DMR algorithms and configuration options

As explained in the main part of this article, the analysis of the simulated data sets was performed in a blind manner. This means that the analyses in the Marburg lab were performed without any knowledge about how the data was generated. Like in real application setting, the algorithms had to be configured as good as possible by manual inspection and/or by using default configuration as guiding instruction. In this section, we summarize the configuration options which were chosen by the Marburg lab. All scripts used for the analyses are provided online at https://github.com/kreutz-lab/DMR-DecisionTree.

There are three configuration parameters which were set to the same values for all algorithms:

- The default p-value for all tools was set to 0.05.
- The minimum number of CpGs per DMR was set to 10.
- The minimum methylation difference was set to 10%

The R package *MethylKit* v1.8.1 from Bioconductor Release 3.8 has been used to apply *MethylKit*. The core commands are

```
MethylRawList_filtered <- filterByCoverage(MethylRawList, lo.count=10, hi.perc=99.9)
MethylRawList_methDiffDSS <- calculateDiffMethDSS(MethylRawList_filtered, mc.cores=2)
MethylRawList_methDiff_subsetDSS <- getMethylDiff(MethylRawList_methDiffDSS,
difference=10,qvalue=0.05)
```

*MOABS* takes inputs files from MethylKit's fitted files. First, a custom Python script is run to get proper input files for *MOABS*:

```
python toMoabsMcompInput.py inputFolder outputFolder   # produces input files for moabs
```

Then the *MOABS* executable is run from command line:

```
moabs-v1.3.2.src.x86_64_Linux.data/bin/mcomp -p 2 -d 10
   -r wt_r1.CG.bed,wt_r2.CG.bed,wt_r3.CG.bed -r ko_r1.CG.bed,ko_r2.CG.bed,ko_r3.CG.bed
   -m wt.bed ko.bed -c comp.wt.vs.ko.txt  # filenames have to be adapted for CHG, CHH
```

Then, the columns of interest are selected via

```
cut dmr_M1_wt.CG.bed_vs_ko.CG.bed.txt -f1,2,3|tail -n +2
   > moabs_dmrs.bed  # filenames have to be adapted for CHG, CHH
```

*Metilene* is run from command line via

```
metilene_linux64 -a group1 -b group2  input_file > output_file
```

For *CHH* and *CHG* contexts, the option `-G 200` is used. The following *metilene* Perl script generates DMR files and some visualisation like quantile-quantile plots:

```
perl metilene_output.pl -q output_file -o output_dir
```

Control and sample groups were analysed by *Defiant* via

```
defiant -L control,case -b -c 0 -i control1.txt,control2.txt,control3.txt case1.txt,case2.txt,case3.txt
```

The option put`-c 0` was chosen to set the coverage parameter to zero because *Defiant* generated empty output for *Ae. arabicum* because the coverage was too small.

The core R commands for *DMRcate* are

```
obj_bsseq <- makeBSseqData(samples, sampNames)
DSSres <- DMLtest(obj_bsseq, group1=sampNames[1:3], group2=sampNames[4:6], smooth=F)
wgbsannot <- cpg.annotate(datatype="sequencing", DSSres, analysis.type="differential")
wgbs.DMRs <- dmrcate(wgbsannot, lambda = 1000, C = 50, pcutoff = 0.05, mc.cores = 1)
```

Again, Bioconductor release 3.8 was applied.

The core R commands for *Methylsig* are

```
meth_cpg <- methylSigReadData(filelist, sample.ids = sample_ids,
     assembly = "sim_cpg_g129", treatment = rep(c(1,0),
     each = length(sample_ids)/2), minCount=10, context = "CpG", destranded = F)
meth_cpg_tile <- methylSigTile(meth_cpg, win.size = 200)
meth_cpg_dmr <- methylSigCalc(meth_cpg_tile, groups = c(1,0), min.per.group=3,
    local.disp = T, winsize.disp= 200, local.meth = T, winsize.meth = 200)
meth_cpg_dmr_filtered <- meth_cpg_dmr[meth_cpg_dmr[,"qvalue"] < 0.05 \& abs(meth_cpg_dmr[,"meth.diff"] > 10)]
```

The core R commands for *Bsmooth* are

```
dataBSmooth <- BSmooth(m_dataBS, mc.cores=1, verbose=T)
dataBS_cov <- getCoverage(dataBSmooth ==10 )
BS_cov_filtered <- which(rowSums(dataBS_cov[,m_dataBS\$Type == "test"]>= 2)>=2 \&
    rowSums(dataBS_cov[,m_dataBS\$Type == "control"]>=2)>=2)
dataBSmooth <- dataBSmooth[BS_cov_filtered,]
dataBSmooth_tstat <- BSmooth.tstat(dataBSmooth,
    group1 = c("percentC_sample1", "percentC_sample2","percentC_sample3"),
    group2 = c("percentC_sample4", "percentC_sample5","percentC_sample6"),
    estimate.var="same", local.correct = T, verbose = T)
```

# 7 Computation times

| tool | Version | All | | P.patens | | P.abies | | A. thaliana | | Ae. arabicum | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | time, min | RAM, min | time, min | RAM, min | time, min | RAM, min | time, min | RAM, min | time, min | RAM, min |
| **Defiant** | v1.1.3 | **0,14** | 55,3 | 0,13 | 71,4 | 0,17 | 6,98 | 0,13 | 71,4 | 0,14 | 71,5 |
| **MethylKit** | v1.8.1 | **20,7** | 1098,5 | 33,8 | 1258,5 | 24,7 | 764,1 | 19,8 | 1389,5 | 4,60 | 982,0 |
| **Metilene** | v0.2-7b | **70,9** | 1335,5 | 83,8 | 288,8 | 0,14 | 71,3 | 55,6 | 3404,0 | 144,0 | 1578,0 |
| **MethylSig** | v0.5.2 | **102,4** | 1009,3 | 118,0 | 467,0 | 197,5 | 1392,2 | 55,2 | 1260,2 | 39,0 | 917,7 |
| **MOABS** | v1.3.2.src.x86_64_Linux | **251,8** | 4943,4 | 124,9 | 3527,8 | 288,0 | 2378,8 | 414,2 | 6933,1 | 180,0 | 6934,1 |
| **DMRCate** | BioC Release 3.8 | **329,9** | 1969,8 | 293,6 | 1412,4 | 278,3 | 2820,2 | 463,3 | 1336,1 | 284,5 | 2310,6 |
| **Bsmooth** | BioC Release 3.8 | **2246,9** | 1592,5 | 730,7 | 1012,4 | 333,0 | 3148,3 | 7504,0 | 293,0 | 419,9 | 1916,4 |

| tool | Version | All | | P.patens | | P.abies | | A. thaliana | | Ae. arabicum | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | time, max | RAM, max | time, max | RAM, max | time, max | RAM, max | time, max | RAM, max | time, max | RAM, max |
| **Defiant** | v1.1.3 | **12,0** | 866,2 | 1,38 | 71,5 | 46,3 | 3250,0 | 0,15 | 71,7 | 0,19 | 71,7 |
| **MethylKit** | v1.8.1 | **65,7** | 2022,4 | 93,5 | 1303,7 | 60,5 | 1359,1 | 55,5 | 1634,9 | 53,5 | 3792,1 |
| **Metilene** | v0.2-7b | **959,1** | 1885,3 | 85,6 | 1050,0 | 66,2 | 178,8 | 89,5 | 4288,0 | 3595,0 | 2024,6 |
| **DMRCate** | BioC Release 3.8 | **3816,7** | 3726,6 | 313,0 | 3374,4 | 307,1 | 4136,6 | 548,0 | 2780,5 | 14098,6 | 4615,1 |
| **MethylSig** | v0.5.2 | **5252,0** | 1353,9 | 465,9 | 1768,0 | 551,9 | 1396,1 | 162,7 | 1262,7 | 19827,4 | 988,6 |
| **MOABS** | v1.3.2.src.x86_64_Linux | **5572,6** | 6977,6 | 130,2 | 6935,1 | 2016,0 | 6936,9 | 19292,4 | 7101,7 | 852,0 | 6936,7 |
| **Bsmooth** | BioC Release 3.8 | **6103,3** | 1843,7 | 941,4 | 1111,0 | 964,6 | 3158,8 | 12225,5 | 410,7 | 10282,0 | 2694,4 |

| tool | Version | All | | P.patens | | P.abies | | A. thaliana | | Ae. arabicum | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | time, mean | RAM, mean | time, mean | RAM, mean | time, mean | RAM, mean | time, mean | RAM, mean | time, mean | RAM, mean |
| **Defiant** | v1.1.3 | **4,1** | 327,4 | 0,6 | 71,4 | 15,5 | 1109,5 | 0,1 | 71,5 | 0,1 | 57,2 |
| **MethylKit** | v1.8.1 | **41,6** | 1360,8 | 77,1 | 1333,2 | 43,4 | 1080,6 | 41,4 | 1511,9 | 4,6 | 1517,5 |
| **Metilene** | v0.2-7b | **51,9** | 2090,5 | 87,2 | 745,3 | 33,0 | 2107,8 | 69,9 | 3882,7 | 17,4 | 1626,1 |
| **DMRCate** | BioC Release 3.8 | **777,7** | 2374,5 | 307,4 | 2133,7 | 295,8 | 3260,4 | 500,0 | 1844,3 | 2007,7 | 2259,7 |
| **MethylSig** | v0.5.2 | **2203,7** | 1231,9 | 356,2 | 1508,8 | 375,3 | 1394,3 | 125,9 | 1261,7 | 7957,2 | 762,8 |
| **MOABS** | v1.3.2.src.x86_64_Linux | **2204,8** | 5968,8 | 138,4 | 4672,4 | 1536,0 | 5416,8 | 6740,9 | 2371,5 | 404,0 | 5548,6 |
| **Bsmooth** | BioC Release 3.8 | **3580,1** | 1633,3 | 953,5 | 1078,1 | 615,4 | 3152,6 | 7385,5 | 365,2 | 5366,0 | 1937,4 |

Table 3: Minimal, average and maximum of the runtimes (in seconds) and RAM (in kB) load for the different tools for analyzing the simulation data with 1e6 genomic positions. *Defiant* is clearly the fastest approach and also requires least memory. *Metilene* and *MethylKit* are faster than the other tools.

The DMR algorithms were applied on a linux computer with Intel(R) Xeon(R) CPU E5-2609-0 2.40GHz and 62.9 GB RAM. The required computation times and memory usages are summarized in Table 3. *Metilene* that had on average the best performance is also faster than most other tools on average. Only *Defiant* is clearly faster and *MethylKit* exhibits a minor benefit in terms of computation times.

# 8 Result of multivariate analysis of performance

Figure 9 shows power curves for each algorithm and the 21 analysed data settings. These curves indicate the dependency of the F1-score on the magnitude of the "signal", i.e. the underlying difference of the methylation levels. In real experimental studies, the magnitudes of the methylation differences depend on the similarity of the investigated biological conditions. The performance of DMR approaches in general depends on this similarity (or difference). Thus, it depends on the biological setting at which point these power curves have to be evaluated to assess the performances for data from new applications. In addition, the power of the approaches also depends on the number of replicates that were available as well as on the sequencing depth. We assessed the performance of the DMR approaches for a signal-to-noise ratio for three replicates and the sequencing depth given by our data templates. More replicates and/or an increased number of reads in general improve the signal-to-noise ratio. As a rule of thumb, a larger sample size always increases the power of statistical analyses. Thus, in our context a large number of replicates and/or an increased sequencing depth enables identification of smaller methylation differences.

Nevertheless, the rankings of the approaches is mostly concordant over a large or even the whole range of methylation differences. Therefore, the ranking of the DMR approaches hold for a broad range of signal-to-noise levels. Metilene is most frequently superior, but has a rather weak performance for Aetar-T-CHH, Aetar-C-CHH, Aetar-c-CG and Aetar-C-CG.

The F1-scores averaged over the differential methylation levels $\Delta$ were analysed by a multivariate linear model to estimate the average impact of the choice of the DMR algorithm and for assessing significance. The sample size for this analysis is given by the number of simulated and analyzed data sets that could be increased. Increasing the number of data sets would decrease the size of the standard errors and would lead to further significant outcomes if there is are real, possibly small, performance advantages. This aspect should be considered for interprations of the calculated p-values.

Table 4: Outcomes of the multivariate regression analysis. The intercept denotes the performance of *metilene* for the Phypa-R-CG data set which has been defined as reference performance and yields an F1-score equals to 0.76. The other estimates denote changes of the F1-score relative to this number if the algorithm or data context changes. For this parameterization, it can be tested whether switching the algorithm cause significant performance changes relative to the reference analysis. SE denotes standard errors. Except for DMRcate, the estimated decreases of the performances of the alternative approaches are significant.

| | estimate | SE | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 0.75 | 0.072 | 10 | 4.6e-19*** |
| DMRcate | -0.052 | 0.054 | -0.95 | 0.34 |
| Defiant | -0.12 | 0.054 | -2.2 | 0.026* |
| BSmooth | -0.16 | 0.054 | -3 | 0.0037** |
| MethylKit | -0.27 | 0.057 | -4.8 | 3.9e-6*** |
| MethylSig | -0.35 | 0.057 | -6.1 | 9.8e-9*** |
| MOABS | -0.22 | 0.054 | -4 | 0.00011*** |
| MethylScore | -0.29 | 0.054 | -5.4 | 3.6e-7*** |
| Phypa-G-CHG | 0.006 | 0.088 | 0.068 | 0.95 |
| Phypa-G-CHH | 0.014 | 0.088 | 0.16 | 0.87 |
| Phypa-G-CG | 0.071 | 0.088 | 0.8 | 0.42 |
| Phypa-R-CHG | -0.1 | 0.088 | -1.2 | 0.25 |
| Phypa-R-CHH | 0.18 | 0.088 | 2 | 0.044* |
| Picab-P-CG | -0.2 | 0.088 | -2.3 | 0.021* |
| Picab-P-CHG | -0.078 | 0.088 | -0.89 | 0.38 |
| Picab-P-CHH | 0.042 | 0.088 | 0.48 | 0.63 |
| Picab-M-CG | -0.21 | 0.088 | -2.4 | 0.02* |
| Picab-M-CHG | 0.068 | 0.088 | 0.77 | 0.44 |
| Picab-M-CHH | 0.076 | 0.088 | 0.86 | 0.39 |
| Aetar-C-CG | -0.35 | 0.095 | -3.7 | 0.00032*** |
| Aetar-C-CHG | -0.36 | 0.095 | -3.7 | 0.00027*** |
| Aetar-C-CHH | -0.032 | 0.088 | -0.36 | 0.72 |
| Aetar-c-CG | -0.28 | 0.088 | -3.2 | 0.0016** |
| Aetar-T-CHG | -0.28 | 0.091 | -3.1 | 0.0026** |
| Aetar-T-CHH | -0.14 | 0.091 | -1.5 | 0.13 |
| Arath-CG | 0.2 | 0.088 | 2.3 | 0.025* |
| Arath-CHG | 0.028 | 0.088 | 0.32 | 0.75 |
| Arath-CHH | 0.012 | 0.088 | 0.14 | 0.89 |

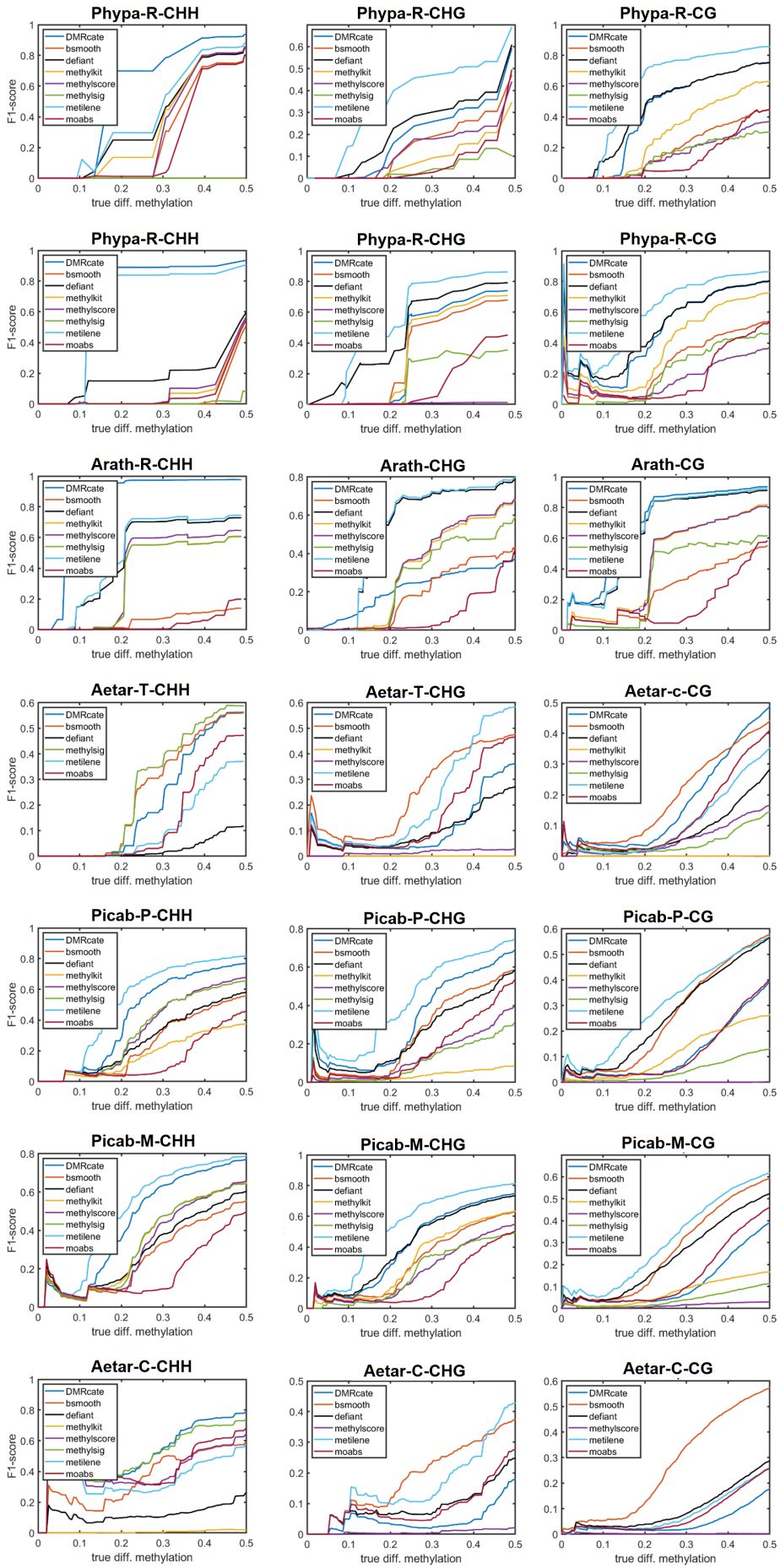



Figure 9: Power curves of the compared approaches for the individual data sets. Metilene as superior approach on average exhibits beneficial performance for most, but not all data sets. This indicates the requirement of a data-based selection guideline.

# References


Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15), 2114–2120.

Haudry, A., Platts, A. E., Vello, E., Hoen, D. R., Leclercq, M., Williamson, R. J., Forczek, E., Joly-Lopez, Z., Steffen, J. G., Hazzouri, K. M., *et al.* (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, **45**(8), 891.

Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, **27**(11), 1571–1572.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Lucas, A. (2018). *amap: Another Multidimensional Analysis Package*. R package version 0.8-16.

Mohammadin, S., Wang, W., Liu, T., Moazzeni, H., Ertugrul, K., Uysal, T., Christodoulou, C. S., Edger, P. P., Pires, J. C., Wright, S. I., *et al.* (2018). Genome-wide nucleotide diversity and associations with geography, ploidy level and glucosinolate profiles in Aethionema arabicum (Brassicaceae). *Plant Systematics and Evolution*, **304**(5), 619–630.

Qin, C., Shasha, W., Zhirui, D., Liping, Y., Bin, H., and Xiaolei, K. (2009). Optimization of DNA extraction from seeds of Sorghum sudanense (piper) stapf. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca*, **37**(1), 256–260.

Rackham, O. J. L., Dellaportas, P., Petretto, E., and Bottolo, L. (2015). WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools. *Bioinformatics*, **31**(14), 2371–2373.