SUPPLEMENTAL INFORMATION

Data and text mining

ViewBS: a powerful toolkit for visualization of high-throughput bisulfite sequencing data

Xiaosan Huang¹, Shaoling Zhang¹, Kongqing Li², Jyothi Thimmapuram³ and Shaojun Xie³,*

¹College of Horticulture, State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China, 210095, ²Department of Rural Development, Nanjing Agricultural University, Nanjing, China, 210095, ³Bioinformatics Core, Purdue University, West Lafayette, Indiana 47907

*To whom correspondence should be addressed.

Supplemental methods

Linux system information for testing ViewBS

The test of ViewBS was performed on a DigitalOcean Droplet virtual server (512 MB Memory / 20 GB Disk). The version of Linux was CentOS Linux release 7.1.1503 (Core). Model name of the CPU was Intel(R) Xeon(R) CPU E5-2630L v2 @ 2.40GHz. The CPU was a one-core processor.

Processing of bisulfite reads

To evaluate the performance of ViewBS (Table S1), we used publicly available data deposited in NCBI GEO database [1]. The GEO accession number is GSE51304. The BS-seq data were for five samples: wild type (WT), cmt2-3 (cmt2), drm1 drm2 cmt2 triple mutant (drm12cmt2), cmt2 cmt3 (cmt3) double mutant, drm1 drm2 cmt2 cmt3 (drm12cmt23) quadruple mutants. Reads in SRA format were converted to fastq format. SolexaQA was used to remove low quality bases [2]. Quality filtered reads were then aligned to Arabidopsis reference genome (TAIR10) using Bismark [3]. The module bismark_methylation_extractor in Bismark was then used to extract context-dependent (CpG/CHG/CHH) methylation sites. The module coverage2cytosine (part of the Bismark package) was run to generate genome-wide cytosine report output based on the coverage information generated by bismark_methylation_extractor. DMRs were determined as described previously [4–6].

Processing of RNA-seq data

The corresponding RNA-seq data were downloaded from NCBI GEO database. Reads for each sample were aligned to TAIR10 reference genome using TopHat [7]. The bam files generated were then used as input in Cuffdiff [8] analysis. The FPKM values of WT sample were extracted from cuffdiff results. Only genes that were expressed (test status: OK) were extracted and ranked into five groups based on gene expression levels: lowest to highest (rank 1 to rank 5).

Preparation of input data for ViewBS

Before using as input in ViewBS, genome-wide cytosine methylation report should be compressed by bgzip and then indexed by tabix which are now part of the HTSlib project (https://github.com/samtools/htslib) [9]. Bgzip compresses the data file in the BGZF format, which is the concatenation of a series of gzip blocks with each block holding at most 216 bytes of uncompressed data. Before being compressed, the data file needs to be sorted first by sequence name and then by leftmost coordinate, which can be done with the standard UNIX command sort.

For BS-seq that is not processed by Bismark but by other tools like BRAT [10,11] BS-Seeker2 [12], ViewBS provides supports to convert DNA methylation data in other formats to the format of genome-wide cytosine methylation report. Support for other tools will be developed upon requests from the users.

Supplementary data for evaluation of ViewBS

Evaluation of tools for profiling of genome-wide DNA methylation
First, we generated the genome-wide profiling of BS-seq (Figure S1). The tool MethCoverage was used to profile the read coverage of BS-seq (Figure S3). The tool BisNonConvRate was used to estimate the non-conversion rate based on the non-methylated chloroplast genome. The non-conversion rates were 0.053, 0.048, 0.04, 0.046 and 0.075 for cmt2-3, drm12cmt2, drm12cmt23, cmt23 and WT, respectively (Figure S4). To profile global methylation level, the tool MethGlobalLev was used (Figure S5). In WT, the weighted methylation levels were 24.5%, 7.9% and 2.9% for CG, CHG and CHH context, respectively. Whereas in drm12cmt23, the weighted methylation levels were 21.9%, 0.4%, 0.5% for CG, CHG and CHH context, respectively. This is consistent with the functions of DRM1, DRM2, CMT2 and CMT3 in Arabidopsis [1]. The tool MethLevDist was used to profile the distribution of DNA methylation level of cytosines (Figure S6). The tool MethGeno was used to profile the DNA methylation patterns across each chromosome in Arabidopsis (Figure S7).

Evaluation of tools for visualizing DNA methylation patterns in selected regions

To evaluate the performance of MethOverRegion, transposable elements on chromosome 1 of TAIR10 reference genome were selected. CHG methylation patterns across the TE were plotted (Figure S2 B). Using this plot, we can see the effect of mutations on CHG methylation across TE. Users can generate similar results in other selected regions of interest, for example methylation in generic regions. This tool can also be used if the users want to compare DNA methylation patterns for regions with different features. For example, we generated DNA methylation patterns for five groups of genes with expression levels from lowest to highest (Figure S8). This analysis can be used to study the correlation between DNA methylation levels and gene expression levels. DNA methylation levels for one region (chr5:19497000-19499600) were generated by another tool named MethOneRegion (Figure S2 E). This tool can help users visualize DNA methylation level for specific regions.

To evaluate the performance of MethHeatmap, we first identified hypo-CHG-DMRs between WT and drm12cmt23 that showed significantly lower methylation levels in drm12cmt23 than in WT. For these hypo-CHG-DMRs identified, we extracted methylation level in other three tissues. One violin-boxplot and one heat map were generated by this tool (Figure S2 C and D). From the violin-boxplot, we can clearly see that DNA methylation levels in all four mutants were lower than that of WT. Mutants cmt23 and drm12cmt2 showed even lower methylation level than other two mutants. From the heat map, we can see more detailed information. DNA methylation level in these regions was lowest in drm12cmt23. In drm12cmt2, we can see half of the regions also showed lower DNA methylation level than in WT (Figure S2 C and D). These results are consistent with the functions of genes in the published results [1].

Runtime and memory usage for processing data

In order to reduce memory requirement, each time ViewBS will use tabix method to access the genome-wide cytosine methylation report if it needs to extract methylation information for a region. This will reduce the time and memory usage compared to loading the genome-wide cytosine methylation report into memory.

Here we listed the maximum memory consumptions and time used for each tool for testing ViewBS (Table S2). The command for GlobalMethLev consumed 870 seconds and the maximum memory used was 70,024 kb. The command for BisNonConvRate consumed 681 seconds and the maximum memory used was 69,252 kb. The command for MethCoverage consumed 704 seconds and the maximum memory used was 330,748 kb. The command for MethLevDist consumed 818 seconds and the maximum memory used was 78,924 kb. The command for MethGlobalLev consumed 870 seconds and the maximum memory used was 70,024 kb. The command for MethGeno consumed 1,001 seconds and the maximum memory used was 71,680 kb. The command for MethOverRegion consumed 363 (273) seconds and the maximum memory used was 73,876 (73,912) kb. The command for MethHeatmap consumed 384 seconds and the maximum memory used was 94,648 kb. The command for MethOverRegion consumed 2 seconds and the maximum memory used is 71,056 kb. All the tests for ViewBS were performed on a DigitalOcean Droplet virtual server with 512 MB Memory and 20 GB disk). Taken together, ViewBS is efficient in time and memory usage.

All the test data and command lines can be freely obtained at https://gitlab.com/BS-seq/ViewBS_testdata.
Legends for supplementary tables and figures

Table S1 Overview of currently available tools in ViewBS

Table S2 Performance of ViewBS on testing the BS-seq data

Figure S1 Overview of the workflow of ViewBS for the functions of genome-wide profiles of BS-seq data. Genome-wide cytosine methylation report for different samples (A) can be used in ViewBS along with genome. Functions like GlobalMethLev, BisNonConvRate, MethCoverage and MethLevDist (B-F) can be selected to generate genome-wide profiles.

Figure S2 Overview of the workflow of ViewBS for the functions of visualizing DNA methylation patterns in functionally important regions. Genome-wide cytosine methylation report for different samples can be used in ViewBS along with selected regions (A), for example genes, DMRs, etc. Then figures, meta-plots, heat maps, violin-boxplots can be generated (B-E). (B) Methylation patterns across TE for CHG context. (C and D) Heat map and violin-box plot for randomly selected hypo-CHG-DMRs between drm12/cmt23 and WT. (E) CHG methylation levels of a CHG-DMR, chr5:19497000-19499600.

Figure S3 Read coverage of BS-seq data for each sample

Figure S4 Non-conversion rate of BS-seq data

Figure S5 Global (bulk) methylation level of BS-seq data

Figure S6 Distribution of methylation level of BS-seq data

Figure S7 Methylation levels across the chromosomes for BS-seq data

Figure S8 Methylation levels across different groups of genes with different expression levels
Supplemental Tables and Figures

Table S1 Overview of currently available tools in ViewBS

<table>
<thead>
<tr>
<th>Type</th>
<th>Tool name</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome-wide profiling</td>
<td>GlobalMethLev</td>
<td>Generate global average methylation level</td>
</tr>
<tr>
<td></td>
<td>BisNonConvRate</td>
<td>Estimate non-conversion rate of BS-seq data.</td>
</tr>
<tr>
<td></td>
<td>MethCoverage</td>
<td>Generate statistics of read coverage</td>
</tr>
<tr>
<td></td>
<td>MethLevDist</td>
<td>Generate distribution of methylation levels</td>
</tr>
<tr>
<td></td>
<td>MethGeno</td>
<td>Plot methylation level across the chromosome</td>
</tr>
<tr>
<td>Visualization for selected regions</td>
<td>MethOverRegion</td>
<td>Generate meta-plot over selected regions</td>
</tr>
<tr>
<td></td>
<td>MethHeatmap</td>
<td>Generate heat map and violin-boxplot for selected regions</td>
</tr>
<tr>
<td></td>
<td>MethOneRegion</td>
<td>Visualize DNA methylation for a given regions</td>
</tr>
</tbody>
</table>

Table S2 Performance of ViewBS on testing the BS-seq data

<table>
<thead>
<tr>
<th>Type</th>
<th>Tool name</th>
<th>Time (seconds)</th>
<th>Memory (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome-wide profiling</td>
<td>GlobalMethLev</td>
<td>870</td>
<td>70,024</td>
</tr>
<tr>
<td></td>
<td>BisNonConvRate</td>
<td>681</td>
<td>69,252</td>
</tr>
<tr>
<td></td>
<td>MethCoverage</td>
<td>704</td>
<td>330,748</td>
</tr>
<tr>
<td></td>
<td>MethLevDist</td>
<td>818</td>
<td>78,924</td>
</tr>
<tr>
<td></td>
<td>MethGeno</td>
<td>1001</td>
<td>71,680</td>
</tr>
<tr>
<td>Visualization for selected regions</td>
<td>MethOverRegion 1</td>
<td>363/273</td>
<td>73,876/73,912</td>
</tr>
<tr>
<td></td>
<td>MethHeatmap</td>
<td>384</td>
<td>94,648</td>
</tr>
<tr>
<td></td>
<td>MethOneRegion</td>
<td>2</td>
<td>71,056</td>
</tr>
</tbody>
</table>

Note 1: There were two figures generated by MethOverRegion. The former one is for C and D.
Figure S1 Overview of the workflow of ViewBS for the functions of genome-wide profiles of BS-seq data. Genome-wide cytosine methylation report for different samples (A) can be sent to ViewBS along with genome. Functions GlobalMethLev, BisNonConvRate, MethCoverage and MethLevDist (B-F) can be used to generate genome-wide profiles.
Figure S2  Overview of the workflow of ViewBS for visualizing DNA methylation patterns in functionally important regions. Genome-wide cytosine methylation report for different samples can be used in ViewBS along with selected regions (A), such as genes, DMR, etc. Then figures meta-plots, heat maps, violin-boxplots can be generated (B–E). (B) Methylation patterns across TE for CHG context. (C and D) Heat map and violin-box plot for randomly selected hypo-CHG-DMRs between WT and drm12cmt23. (E) CHG methylation levels of a CHG-DMR, chr5:19497000-19499600.
Figure S3 Read coverage of BS-seq data for each sample
Figure S4 Non-conversion rate of BS-seq data
Figure S5 Global (bulk) methylation level of BS-seq data
Figure S6 Distribution of methylation level of BS-seq data
Figure S7 Methylation levels across the chromosomes for BS-seq data
Figure S8 Methylation levels across different groups of genes with different expression levels
References


