Chapter 6

Selecting Hypomethylated Genomic Regions Using MRE-Seq

Elisabeth Wischnitzki, Kornel Burg, Maria Berenyi, and Eva Maria Sehr

Abstract

Here, we describe a method capable of filtering the hypomethylated part of plant genomes, the so-called hypomethylome. The principle of the method is based on the filtration and sequence analysis of small DNA fragments generated by methylation-sensitive four-cutter restriction endonucleases, possessing $^{(5me)}$CpG motifs in their recognition sites. The majority of these fragments represent genes and their flanking regions containing also regulatory elements—the gene space of the genome. Besides the enrichment of the gene space, another advantage of the method is the simultaneous depletion of repetitive elements due to their methylated nature and its easy application on complex and large plant genomes. Additionally to the wet lab procedure, we describe how to analyze the data using bioinformatics methods and how to apply the method to comparative studies.

Key words Plant, Hypomethylome, Gene space, DNA methylation, Reduced representation libraries, MRE-seq, Non-model organisms, De novo assembly, Reference-based assembly, Comparative analysis

1 Introduction

Epigenetic modifications like DNA methylation influence gene expression without changing the underlying DNA sequence. The methylation of cytosines in the DNA is a reversible process common in plants. However, this methylation does not occur randomly in the genome but appears in pattern of methylated stretches in the genome. It has been observed that the gene space (genes and their flanking regions) is showing low methylation levels (hypomethylated) while cytosine methylation is found predominantly in repetitive elements (e.g. transposable elements). Not only the methylation of the gene body influences the expression but especially methylation pattern in the promoter regions has been associated with differential expression indicating a direct role of DNA methylation in the regulation of gene expression [1, 2]. Thus, investigating a plant's methylome (the methylated part of the genome) or hypomethylome (the non-methylated part of the genome) is an

Reinhard Hehl (ed.), *Plant Synthetic Promoters: Methods and Protocols*, Methods in Molecular Biology, vol. 1482,
DOI 10.1007/978-1-4939-6396-6_6, © Springer Science+Business Media New York 2016

increasingly popular strategy for understanding the genetic and environmental interactions behind biological processes. Methylation-sensitive restriction enzyme-based genome digestion used for creating reduced representation libraries allows the enrichment of gene space-related sequences by selecting for the hypomethylome [3–9]. The obtained libraries represent not only exons but also potential regulatory regions where regulatory sequences like transcription factor binding sites may reside. More precisely, the identified regions contain additionally introns and gene-flanking regions both up- and downstream. A combination of these libraries with next-generation sequencing (NGS) for their characterization is called MRE-seq (Methylation-sensitive Restriction Enzyme followed by sequencing) [10, 11].

Here, we detail a modified MRE-seq technique designed for the isolation and characterization of a plant's hypomethylome. The method is based on the restriction digestion of total genomic DNA with methylation-sensitive frequent cutter restriction endonucleases resulting in short DNA fragments of hypomethylated regions. PCR-based size selection, next-generation sequencing, and bioinformatics analysis of these short genomic fragments provide a comprehensive sequence representation and characterization of the hypomethylome [8, 9]. In this chapter, we describe both the filtration method and the bioinformatics procedure to analyze the data and give recommendations for performing a comparative analysis between samples.

Our method provides an easy tool to produce reduced representation libraries enriched for gene space omitting repetitive elements from small amounts of genomic DNA samples and opens the way for comparative analysis of genetic and epigenetic variation among genotypes or tissues even in a larger set of samples.

2 Materials

2.1 Plant Material

1. Leaves of the selected plant species were used for the analysis. We recommend fresh material but any material that will yield high-molecular-weight genomic DNA can be used (*see* **Note 1**).
2. The experiments presented are based on rice, *Oryza sativa* ssp. *indica* variety SHZ-2A (seeds are kindly provided by R. Mauleon, IRRI International Rice Research Institute, Los Banos, Philippines) and Norway spruce, *Picea abies* (L.) H. Karst (twigs kindly provided by S. Schüler, Department of Forest Genetics, Austrian Research Centre for Forests, Vienna, Austria).

2.2 Buffers, Enzymes, Adapters, and PCR Primer

1. PCR grade water was prepared in the laboratory by UV irradiating Millipore Synergy generated ion exchanged water in Stratagene UV Stratalinker 2400 for 20 min and then sterile filtered.

2. CTAB lysis buffer: 140 mM Sorbitol, 220 mM Tris–HCl pH 8.0, 22 mM EDTA, 880 mM NaCl, 1 % Sarcosyl and 0.8 % CTAB.
3. Phenol-chloroform-isoamyl alcohol: 25:24:1 saturated with 10 mM Tris–HCl, pH 8.0, 1 mM EDTA.
4. Chloroform–isoamyl alcohol: 24:1 mixture of chloroform and isoamyl alcohol.
5. Isopropyl alcohol.
6. 70 % EtOH, ethanol diluted by distilled water.
7. Absolute EtOH, 100 % ethanol.
8. Liquid N_2.
9. 10 % SDS, Sodium Dodecyl Sulfate dissolved in distilled water.
10. 3.0 M NaOAc pH 5.2, 3 M sodium acetate, pH adjusted with acetic acid.
11. 5 mM Tris–HCl pH 8.0, Tris–HCl dissolved in distilled water, pH adjusted with NaOH.
12. 10 mM Tris–HCl pH 8.0.
13. (10×) restriction enzyme buffer.
14. PCR buffer (10×).
15. $MgCl_2$ (25 mM), Magnesium chloride dissolved in distilled water.
16. dNTP (20 μM), Mixture of the four deoxy nucleotide triphosphates.
17. BSA, Bovine Serum Albumin.
18. 20 mM EGTA.
19. Based on our experience the four-cutter restriction endonucleases HpaII, AciI, and Bsh1236I with recognition sites containing a CpG motif as methylation site provide the best results for the isolation of the hypomethylome of plant genomes [9] (*see* Table 1).
20. T4 DNA ligase (5 U/μL) and buffer.

Table 1
Restriction enzymes

Restriction enzyme	Cut site	End	Buffer	°C	Source
HpaII	$C^{5me}CGG$	5′GC	Neb1	37	New England Biolabs
AciI	$C^{5me}CGC$	5′GC	Neb3	37	New England Biolabs
Bsh1236I	$CG^{5me}CG$	CG blunt	Neb4	37	Thermo Fischer Scientific
PmeI	GTTTAAAC	A blunt	Neb4	37	New England Biolabs

Table 2
Adapter and PCR primer

	Designation	Sequence of oligo[a]	Feature
Adapter A	PmeI_CGWA	5′ GCACGACTGTTTAAA	
Adapter B	PmeI_CGB	5p′ CGTTTAAACAGTCGT	5′ Phosphorylation
Adapter B blunt	PmeI_CGBlunt	5p′ TTTAAACAGTCGT	5′ Phosphorylation
PCR primer	PmeI_CG17	5′ CACGACTGTTTAAACGG	

[a] The adapter sequences are designed to not restore the original restriction site (*see* **Note 5**)

21. HotStar Taq DNA Polymerase.
22. RNase A: 100 mg/mL diluted in water.
23. Proteinase K: 20 mg/mL purchased as liquid.
24. Bal3126. Adapters and PCR primer are listed in Table 2.

2.3 Equipment

1. Retsch mill: Retsch MM301, 25 mL jar, 15 mm steel ball.
2. High speed centrifuge, e.g. Sorvall RC6, SS34 rotor.
3. Eppendorf centrifuge, e.g. Eppendorf Centrifuge 5415D.
4. Vortex mixer, e.g. Scientific Industries Vortex Genie.
5. Thermocycler MJ research.
6. Agilent Bioanalyzer.
7. Glass rod; 1–1.5 mm diameter glass capillary pipet with melted, closed tips.
8. −20 °C freezer.
9. Waterbath.
10. Micropipettes 2, 20, 200, 1000 μL.

2.4 Disposables

1. 35 mL Polyallomer Nalgene Conical Oak Ridge centrifuge tubes.
2. 500 μL, 1 mL, and 2 mL Eppendorf tubes.

2.5 Recommended Analysis Software

The recommended tools (*see* Table 3 and Subheadings 3.6–3.9) reflect our experience and are not exclusive. They can be exchanged with appropriate alternatives depending on the data, the analysis, or new technological developments. Most of the tools are command-line based and require a Unix system. In general the analysis will be much faster, if the separate steps can be run in parallel. For the de-novo assembly, we recommend a compute cluster with a high amount of RAM.

Table 3
Recommended analysis software

Software	References	Analysis
CutAdapt	[16]	Removal of adapter sequences (*see* Subheading 3.6)
Trimmomatic	[17]	Removal of low-quality and short sequences (*see* Subheading 3.6)
Bowtie2	[18]	Removal of other sequences (*see* Subheading 3.6) Reference-based assembly (*see* Subheading 3.7) De novo assembly (*see* Subheading 3.7) De novo assembly (*see* Subheading 3.7) Mixed approach (*see* Subheading 3.7) Comparative sequence analysis (*see* Subheading 3.9)
samtools	[21]	Reference-based assembly (*see* Subheading 3.7) De novo assembly (*see* Subheading 3.7) Mixed approach (*see* Subheading 3.7) Comparative sequence analysis (*see* Subheading 3.9)
Bedtools	[22]	Reference-based assembly (*see* Subheading 3.7) De novo assembly (*see* Subheading 3.7) Mixed approach (*see* Subheading 3.7) Annotation (*see* Subheading 3.8) Comparative sequence analysis (*see* Subheading 3.9)
Trinity	[23, 24]	De novo assembly (*see* Subheading 3.7) Mixed approach (*see* Subheading 3.7)
FLASH	[36]	De novo assembly (*see* Subheading 3.7)
Blast	[25, 26]	Mixed approach (*see* Subheading 3.7) Annotation (*see* Subheading 3.8) Comparative sequence analysis (*see* Subheading 3.9)
InterproScan	[27, 28]	Annotation (*see* Subheading 3.8)
Blast2GO	[29, 30]	Annotation (*see* Subheading 3.8)
cd-hit	[31, 32]	Comparative sequence analysis (*see* Subheading 3.9)
IGV	[33, 34]	Comparative sequence analysis (*see* Subheading 3.9)

3 Methods

The method consists of the following different procedures which are further detailed in the Subheadings: 3.1. Preparation of high-molecular-weight genomic DNA, 3.2. Adapter preparation, 3.3. Enzymatic digestion of genomic DNA and ligation of adapters, 3.4. Amplification of adapter ligated DNA, 3.5. Sequencing, 3.6. Data processing and quality control of the raw reads, 3.7. Identification of hypomethylated regions detailing the different assembly strategies, 3.8. Different annotation methods. An additional Subheading 3.9 is focusing on the application of comparative analysis between different samples.

3.1 Preparation of High-Molecular-Weight Genomic DNA

Genomic DNA from *Oryza sativa* and *Picea abies* was prepared with a modified protocol from Janice Keller and Ian Bancroft [12].

1. Grind plant material (0.5 g) to fine powder in liquid N_2 with steel balls in a Retsch mill (Retsch MM301, 25 mL jar, 15 mm steel ball).
2. Melt the frozen powder in 7 mL of 65 °C CTAB lysis buffer and 1 mL of 10% SDS (35 mL Polyallomer Nalgene Conical Oak Ridge centrifuge tubes).
3. Incubate samples at 65 °C in a waterbath for 30 min with occasional vigorous vortex shaking.
4. After incubation extract the samples twice with 9 mL of chloroform-isoamyl alcohol (24:1) and centrifuge at 8000 × *g* for 10 min (Sorvall RC6, SS34 rotor).
5. Transfer the upper aqueous phase into a new tube (same as above).
6. Precipitate with 0.8 volumes of isopropyl alcohol.
7. Incubate for 10 min at room temperature.
8. Centrifuge the samples at 15,000 × *g* for 20 min.
9. Wash pellets with 70% EtOH, dry and dissolve in 600 μL 5 mM Tris–HCl pH 8.0 containing 300 ng of RNase A.
10. Incubate samples at 37 °C for 1 h.
11. Add 30 μg of Proteinase K and incubate at 37 °C for an additional hour.
12. Extract the samples twice with equal volumes of phenol–chloroform–isoamyl alcohol and twice with equal volumes of chloroform–isoamyl alcohol (24:1) in 2 mL Eppendorf tubes.
13. Precipitate the extracted samples by adding 0.1 volumes of 3 M NaOAc pH 5.2 and 2 volumes of absolute EtOH.
14. Roll out the high-molecular-weight genomic DNA with a glass rod.
15. Wash the samples with 70% EtOH and let it dry.
16. Dissolve the DNA overnight in 100 μL of 5 mM Tris–HCl pH 8.0.

3.2 Adapter Preparation

1. Dissolve the lyophilized adapters and primers at 100 μM concentration in sterile PCR grade water.
2. Dilute an aliquot of both A and B (or B blunt) adapters (*see* Table 2) to 10 μM concentration.
3. Mix in a 1:1 ratio.
4. Anneal in thermocycler by heating the mix to 95 °C for 5 min and subsequently cooling it stepwise by 5 °C/5 min to 25 °C.
5. Store the annealed adapters at −20 °C aliquoted in 500 μL Eppendorf tubes (*see* **Note 2**).

6. The oligo pairs PmeI_CGWA and PmeI_CGB oligos are used for HpaII and AciI enzyme site adapters, while for Bsh1236I enzyme site PmeI_CGWA and PmeI_CGBlunt oligos are used (*see* Table 2).

3.3 Digestion and Ligation of Genomic DNA

1. Use a single reaction for the restriction digestion and adapter ligation of the genomic DNA. The 50 μL reaction mix contains 300 ng purified genomic DNA (*see* **Note 3**), 5 μL (10×) restriction enzyme buffer (*see* Table 1), 4 μL (10 μM) annealed adapter, 40 units of restriction enzyme, 10 units of T4 ligase, and fill up with water to 50 μL.
2. Incubate the digestion-ligation reaction overnight at 37 °C.
3. After heat inactivation at 65 °C for 20 min in the thermocycler, dilute the samples 1:1 with 50 μL PCR grade water and extract subsequently by equal volumes of phenol–chloroform–isoamyl alcohol, then by chloroform–isoamyl alcohol (24:1).
4. Precipitate the extracted DNA with 2.5 volumes of absolute EtOH in the presence of 0.3 M NaOAc pH 5.2, wash with 70% EtOH. Dry and dissolve in 100 μL 5 mM Tris–HCl pH 8.0.

3.4 Amplification of the Adapter-Ligated Genomic Fragments

1. Amplification of the adapter-ligated DNA (*see* **Note 4**): For Illumina sequencing, attain fragments by PCR amplification of the restriction-digested and adapter-ligated genomic DNA samples. Use 5 μL of digested and adapter-ligated DNA (about 15 ng) and 4 μL of 10 μM amplification primer PmeI_CG17 in a 50 μL PCR reaction containing 5 μL PCR buffer (10×), 1 μL $MgCl_2$ (25 mM), 1 μL dNTP (20 μM), 0.5 μL HotStar Polymerase (2.5 units) and add PCR grade water to 50 μL. Initialize the PCR at 95 °C for 15 min followed by 25–30 cycles of 95 °C 30 s/55 °C 40 s/72 °C 50 s and finish by 72 °C for 5 min. The exact number of cycles has to be evaluated experimentally (*see* **Note 6**). The selected three restriction endonucleases are performing equally well for the filtration (*see* ref. 9 and Fig. 1). Note that the sizes of the predominant fragments (bands) are characteristic both for the restriction enzyme (*see* Fig. 1) and for the analyzed plant genome (not shown) and are reproducible (*see* Fig. 2).
2. Removal of the adapter sequences: To increase the length of the usable sequence information, the majority of the adapter sequence should be removed. This can be achieved by PmeI digestion, because its rare cut site GTTTAAAC is present in the adapter sequence. Digest 1 μg of the PCR amplificates with PmeI (NEB) in NEB4 buffer, in two steps under the presence of 100 ng/μL BSA. Perform the first digestion in a 50 μL reaction volume, containing 100 U PmeI enzyme on 37 °C for 2 h followed by a subsequent volume increase to 100 μL in 1× NEB4 buffer including

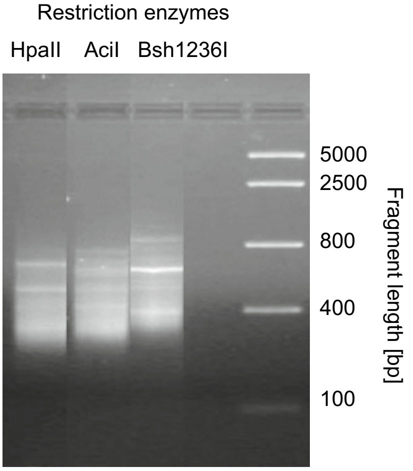

Fig. 1 The results of the size-selective amplification of *Oryza sativa*-digested genomic DNA show characteristic predominant fragment sizes for the different restriction enzymes (0.8 μM PmeI_CG17 primer concentration; 27 PCR cycles)

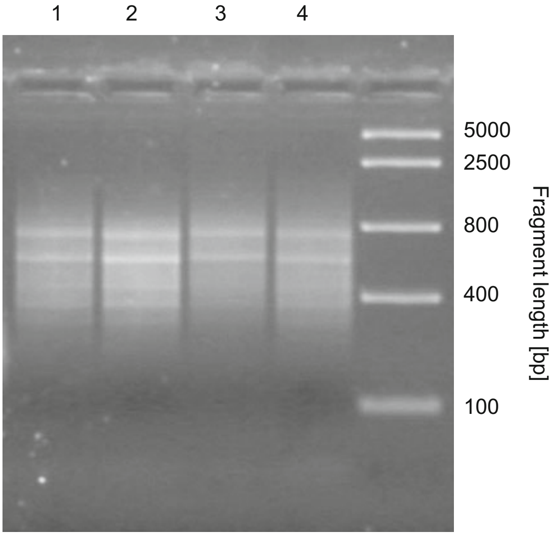

Fig. 2 Filtrated fragments of genomic DNA show reproducible predominant fragment sizes. The results for four different *Picea abies* samples representing different phenotypes are shown

additional 50 U PmeI and incubate for additional 2 h. Finally, stop the reaction at 65 °C for 20 min. Purify the samples (100 μL) with 100 μL phenol–chloroform–isoamyl alcohol (25:24:1), then twice with the same volume of chloroform–isoamyl alcohol. Pipet the upper phase into a new 1.5 mL Eppendorf tube, and precipitate the reaction by adding 0.1 volumes of 3.0 M NaOAc pH 5.2 and

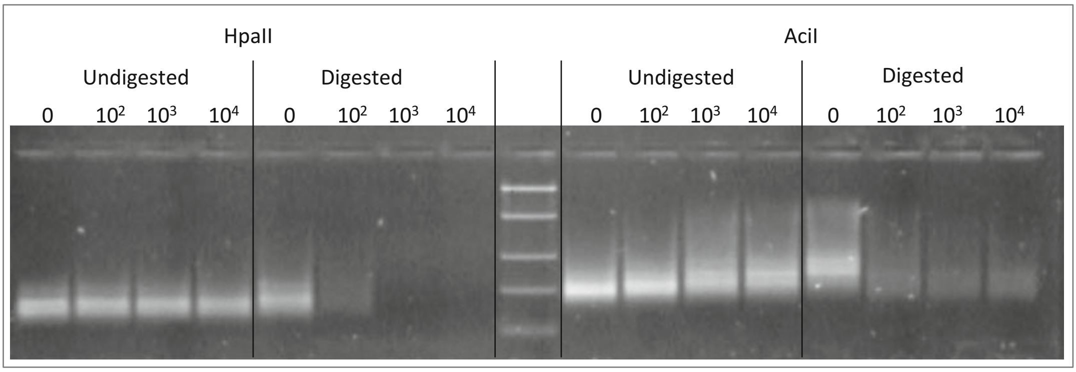

Fig. 3 Test for the efficiency of adapter removal. Samples after PmeI digestion show strongly reduced or no amplification with the PmeI_CG17 primer due to the removed sequence (Digested). In comparison in undigested samples the amplification is still possible for all dilutions (Undigested). All samples were diluted up to 10^4 fold and subsequently amplified with the PmeI_CG17 primer

2.5 volumes of absolute EtOH for 1 h at −20 °C. Centrifuge the precipitated DNA at 4 °C with 16,000 × *g* in an Eppendorf centrifuge (full speed) for 20 min then wash with 70% EtOH. Centrifuge for 10 min at 15,000 × *g*. Discard the supernatant and let the pellet dry at room temperature. Dissolve in 5 mM Tris–HCl pH 8.0 to reach a concentration of about 200 ng/ μL. Store at −20 °C. Test the effectivity of the adapter removal (*see* **Note** 7 and Fig. 3).

3.5 Sequencing

1. During our analysis we tried different sequencing technologies (Illumina HiSeq2000, MiSeq and Ion Torrent) and discovered that the method is suitable for different sequencing technologies as the obtained results were comparable [9]. However, the data created with the Ion Torrent technology are similar to those created by the Illumina technology, but show a much lower coverage. Therefore, the decision for a specific technology is more a matter of availability and personal preferences. However, we recommend to base the decision on the size of the expected fragments (for the expected read length) and on the desired coverage (for the technology). We have been able to enrich for the gene space and deplete repetitive sequences in *Picea abies* with a rather low genome wide coverage of 0.1×. This represents a theoretical coverage of the gene space of ~4× as only about 2.4% of the spruce genome is described as gene and gene-like sequences [13]. Similar results were also achieved in *Crocus sativus* [9]. For more information about the recommended coverage *see* **Note 8**.
2. The sequencing technology of Illumina requires the fragments to be sufficiently different in the first few bases to be able to distinguish the so-called sequence clusters within one lane. If the

sequences are identical, the technology cannot distinguish between actually different fragments and the sequencing run will yield no results. Due to the amplification adapter and the identical genomic cut site, this is the case for the isolated fragments, even if the amplification adaptor is removed. However, there are different methods to approach this technical issue. (1) Adding PhiX fragments as recommended by Illumina for amplicon sequencing did not yield satisfying results in our experiments. A test showed that we had to add 30% of PhiX in order to receive satisfying data. This presents a practical loss of information and can be circumvented by applying one of the next proposed solutions. (2) Dilute the samples with random genomic fragments properly sized to Illumina sequencing instead of PhiX. This way an additional unfiltered genomic reference dataset is obtained (*see* **Note 9**). (3) Removing the amplification adaptor and the genomic cut site with an exonuclease will result in random sequence ends of the fragments suitable for Illumina sequencing. This method can be performed additionally or instead of the removal of the amplification adapter in the protocol. In our hands, it was successfully applied to samples from the saffron crocus [9]. For the removal of a few base pairs at each end of the fragments, the exonuclease Bal31 was used for the digestion (https://www.neb.com/products/m0213-nuclease-bal-31), which has already been applied in a number of studies for the controlled length reduction of linear double-stranded DNA, including studies focusing on telomere truncation [14, 15]. Bal31-driven shortening of the fragments, however, needs optimization for the analyzed samples. Set a 50 μL reaction containing 2 μg of PCR amplified fragments, 1× Bal31 reaction buffer and 1 U of Bal 31 enzyme. Incubate at 30 °C and take samples of 5 μL at 15, 30, 60, 120, etc. seconds, up to 6 min. Mix the removed samples immediately with 5 μL of 40 mM EGTA pre-warmed to 65 °C and incubate for 10 min to stop the reaction. To visualize the progression of the shortening, the samples are diluted 1:5 with 10 mM Tris–HCl pH 8.0 and loaded to an Agilent Bioanalyzer. A size shift of about 30–40 bps is expected for a proper removal of the uniform parts at the fragment ends. Using the results of the time-course experiment, do estimate the necessary digestion time to reach this goal. The size reduction is fragment length dependent, since longer fragments are stronger affected. Therefore, we highly recommend separate optimization for each studied set of samples/species. To prepare fragments for the sequencing set up the same reaction and use the identified time for optimal shortening.

3.6 Data Processing

All sequence reads should be cleaned following the recommended procedure. Especially if the method is used for comparative studies, all samples should be treated in the same way in order to guarantee high-quality data and comparability of the datasets. Some de novo assembly tools include preprocessing but we recommend filtering the reads before further analysis steps. Based on our experience this yields the best results and presents the best basis for comparative analyses.

1. Removal of adapter sequences: In case the digestion of the adapter sequences was not complete, there might be still parts of it present in the dataset. These smaller parts of the adapter sequence should be removed from the read sequences as they are artificially added during the procedure and do not reflect the plant genome. Any available tools that can handle not only a complete adapter sequence but also its substrings can be used for this process, e.g. CutAdapt [16]. This step is still recommended even if the adapter removal step was replaced by the exonuclease digestion (*see* Subheading 3.5).
2. Removal of low-quality regions and short sequences: Low-quality regions with low base call accuracy can affect the mapping and de novo assembly of the fragments by introducing wrong information. Also very short reads should be removed as they may not be mapped uniquely to the reference or may affect the assembly. Therefore, the removal of those sequences is a necessary step to ensure the quality of the results. Any available preprocessing tool is able to perform this procedure, e.g. Trimmomatic [17]. The threshold might depend on the data and the analysis. We recommend applying a Q-value threshold of 30, representing a base call accuracy of 99.9%, and a minimal length threshold of 50 bp. The latter could be set lower if other sequencing methods are used producing shorter reads.
3. Removal of other sequences (optional, *see* **Notes 10** and **11**): Depending on the aims of the analysis it is useful to remove certain sequences prior to further analysis steps, e.g. repetitive elements, ribosomal, chloroplast, or mitochondrial sequences. The filtering can be performed with different tools. We use bowtie2 [18] for filtering the processed reads. We recommend treating each read of a pair separately and applying option: "--un". This will result in a file containing all reads that did not create a sufficient alignment. As reference general databases can be used like, e.g. REdat for repetitive elements [19] or customized datasets depending on your data and research question. If a reference of the chloroplast or the mitochondrion is lacking for the studied species, we recommend using the sequences of close relatives and adjust the threshold to allow more variation, if necessary (e.g. option "--local").

3.7 Identifying Hypomethylated Regions

The procedure for the identification of hypomethylated regions depends on the studied species. More precisely the deciding factor

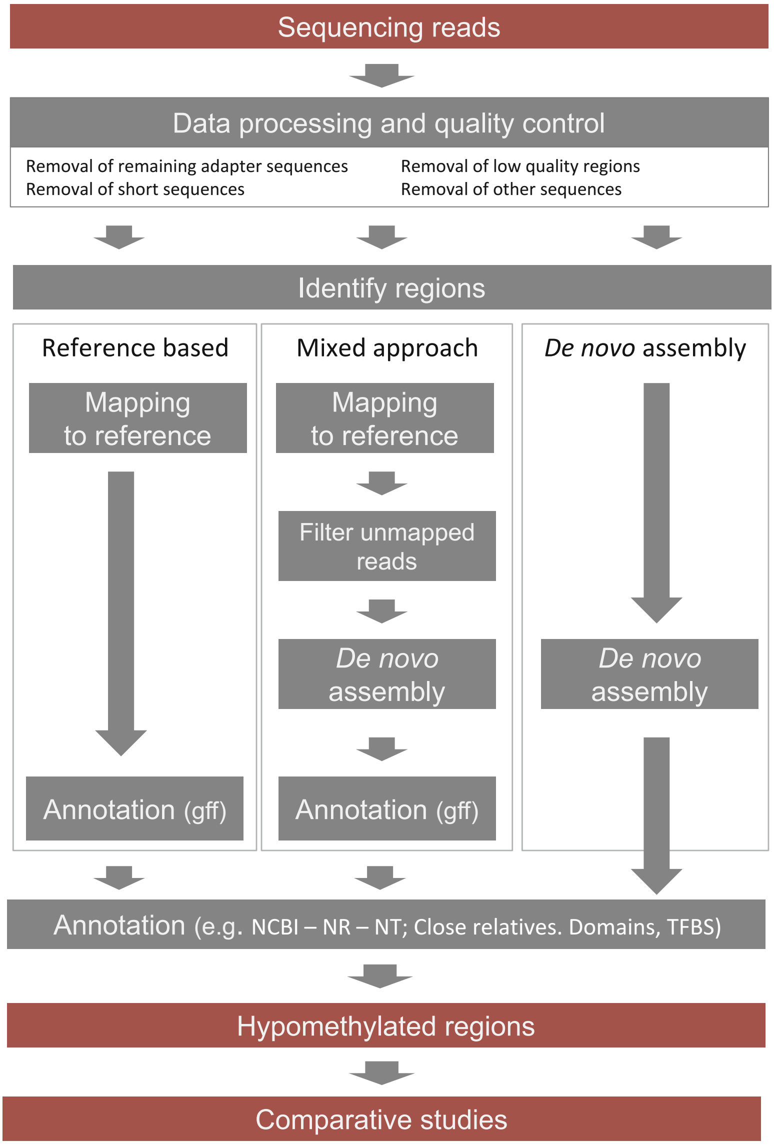

Fig. 4 Bioinformatics workflow for the identification and analysis of hypomethylated regions

is whether a reference sequence is available or not. If a reference sequence is available, a reference-based assembly can be performed. If no reference is present, a de novo assembly approach has to be applied. A combination of reference-based and de novo assembly is a favorable option to yield the best results for the analyzed dataset. Those different analysis procedures are detailed in the following sections (*see* Fig. 4).

1. Reference-based assembly: The reference-guided assembly is performed by assigning all high-quality reads to the reference

sequences. There are different read mapping tools available that are suitable for this procedure. The choice should depend on the analysis and is personal preference. For a review *see* [20]. We recommend bowtie2 with default settings [18] as it allows a bit more variation in the alignment as other tools which is preferable for comparative studies.

Furthermore, we apply additional filter for retaining reliable results and to further reduce the possibility for false-positive sequences. We recommend selecting only regions which are hit by at least five reads (bowtie2 [18], samtools [21], bedtools [22], developed perl-scripts). This threshold will depend on the analysis and the datasets but will increase the reliability of the results.

2. De novo assembly: The de novo assembly for each dataset is performed using Trinity [23, 24] (*see* **Note 12**) with minimal contig length of 100 bp. This parameter can be adjusted if necessary based on the filtrated fragment length and obtained read length, e.g. if only 300 bp reads were obtained this can be set higher. For all technologies and datasets we tested, this value was sufficient. The analysis of various tools showed that Trinity yields the best results. For the purpose of our method—to enrich for the gene space—an assembly method focusing on the transcriptome is best suited. Based on our experience Trinity also performs well for fragments not derived from the gene body [9].

   The resulting contigs should be evaluated by mapping the high-quality reads used for the assembly to the assembled sequences using, e.g. bowtie2 [18] and only contigs consisting of at least five reads should be retained similar to the reference-based assembly.

3. Mixed approach: Whether a reference or de novo approach is applied depends on the organism. If a genome sequence is available apply the reference-based approach first. If a high fraction of reads could not be aligned to the reference genome, use the de novo approach. However, a mixed approach is also possible or even recommendable in some cases. Perform a reference-based assembly first and extract the reads that could not be aligned to the reference (similar to the data processing step "Removal of other sequences" using, e.g. bowtie2 option –un [18]) and perform a de novo assembly with this subset to identify specific sequences that were either not present in the reference or differ to much from the reference to be aligned properly. The de novo sequences can afterward be compared to the reference using blast [25, 26] to identify the potential locations in the reference that differ between your samples and the reference [9] (*see* **Note 13**). De novo-assembled contigs that could not be located in the genome should be subjected to separate annotation to determine their origin.

3.8 Annotation

1. In the case a reference-based assembly was performed, the coordinates of the identified regions should be compared with the available annotation (usually available in gff or gff3 format) using, e.g. bedtools [22]. An additional similarity search is also recommended (*see* **Note 14**).
2. The annotation of the de novo-assembled contigs can be performed by similarity searches using blast [25, 26] against the databases NR and NT from NCBI (http://www.ncbi.nlm.nih.gov/). If close relatives have been sequenced, we recommend a separate blast run against those as well (*see* **Note 14**).
3. Further annotation using, e.g. known transcription factor binding sites (*see* **Note 15**), InterproScan [27, 28] or Blast2GO [29, 30] can additionally provide useful information and is highly recommended.

3.9 Comparative Sequence Analysis

The strategies for comparative studies depend again on the studied species and the availability of a reference. We present some general suggestions but the exact downstream workflow will depend on the specific question to be answered (*see* **Note 16**). However, regardless which method is used it is important to treat all samples subjected to the comparative analysis the same way to ensure comparable results.

1. In case a reference has been published, all samples should be aligned separately to this reference and annotated as described for the reference based assembly. The coordinates of the identified regions can be compared between the samples using, e.g. bedtools [22] to identify regions that are unique to samples or conditions (e.g. intersectbed -v). This can also be applied to identify parts of larger hypomethylated regions that are differentially methylated as it might occur if a regulatory element is methylated in the promoter in one sample but not in the other (e.g. subtractBed).
2. For samples with no reference there are several possibilities. Either all reads are used to create a reference hypomethylome by de novo assembly which is further used as common reference for a reference-based assembly and the subsequent analysis, as described above. Or the samples are de novo-assembled separately and clustered using, e.g. cd-hit [31, 32] to identify unique sequences. For the identification of overlapping sequences, we recommend post-processing the clustering results and creating separate alignments for the exact identification of the differentially methylated parts. Also similarity searches using, e.g. blast [25, 26] are recommended to identify similar sequences. Here again further global alignments might be recommendable to study differences in more detail.

3. For a mixed assembly method as described above, we recommend to add the additional de novo-assembled contigs to the previously known reference. This extended reference will present a more complete basis for the identification of differentially methylated regions. We recommend performing a reference-based analysis with this extended reference as described above.
4. Visualization: If a common reference is present the read alignments can be prepared using, e.g. samtools [21] as bam and bai files and visualized together with the bed files indicating the locations of the identified regions (output of the analysis described in the previous steps) with, e.g. the Integrated Genome Viewer (IGV) [33, 34].

4 Notes

1. Select the DNA isolation method suitable for the chosen plant species, resulting in high-molecular-weight DNA with OD 260/280 nm>1.8 and 260/230 nm ratio>2 to warrant the proper digestibility by the restriction endonuclease. The described method was successfully applied in rice, Norway spruce, banana, sweet potato, saffron, pepper-bark tree and in the 1RS chromosome arm of rye. Store high-molecular-weight genomic DNA at 4 °C to reduce the fragmentation by frequent freezing-melting.
2. Aliquot the annealed adapters in amounts, which may be used up in one ligation reaction and store at −20 °C. Use a single tube for a single restriction-ligation reaction and always discard the rest of the adapters. The annealed adapters are stable at −20 °C for at least 6 months.
3. The amount of DNA used in the reaction is dependent on the genome size of the selected species. The larger the genome the more input DNA is needed. 300 ng of input DNA is equivalent to about 15,000 Norway spruce or about 6×10^5 rice genomes. It is possible to perform the method with less than 20 ng genomic DNA, but think of the genome size! Do not use more than 300 ng DNA per digestion ligation reaction because of the increasing possibility of partial digestion. Always use high concentration restriction enzymes. Think of inhibitory effects of, e.g. glycerine if too much volume of the restriction enzyme is used.
4. Optimization of the PCR: Short fragment amplification is not favored by use of a single adapter and primer, since it allows looping of the short fragments, thereby prohibiting their amplification. However, this handicap may be overcome by

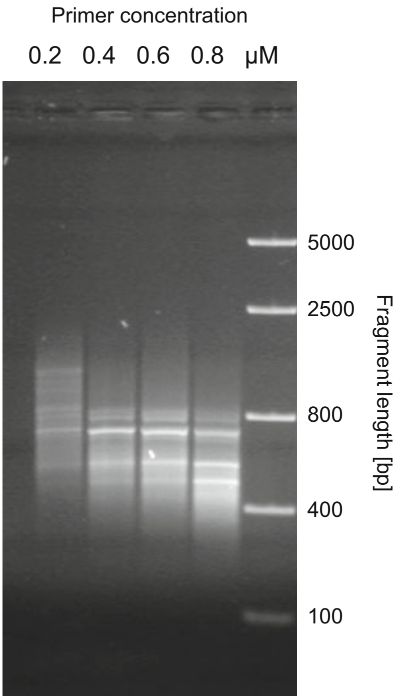

Fig. 5 Amplification of adapter-ligated genomic DNA of *Oryza sativa* produces fragments between 200 and 800 bp for PmeI_CG17 primer concentrations of 0.6–0.8 μM (digestion with 10 ng HpaII restriction enzyme, amplification is shown for different concentrations between 0.2 and 0.8 μM PmeI_CG17 amplification primer)

adjusting the amplification primer concentration [35]. Therefore, the optimization of primer concentration to obtain properly sized PCR products for the Illumina HiSeq sequencing is necessary. The results in Fig. 5 show that increasing the primer concentration in the PCR reaction favors the amplification of shorter amplicons. In the presented case 0.8 μM primer concentration resulted in approx. 200–800 bp fragments, fulfilling the size requirements of both, the hypomethylome filtration and the Illumina HiSeq sequencing.

5. We have observed that the amount of fragments derived from transposable elements can be further decreased by digesting the PCR-amplified fragments with the same restriction endonuclease. The previous methylated sites become unmethylated after PCR amplification and thus digestible. The designed adapters do not restore the original cut site and are therefore not affected. After digestion, the intact fragments—having adapter sequences on both ends—can be further amplified by PCR. Note that this procedure will not only reduce the amount of fragments derived from transposable elements but also affect methylated sites in, e.g. genes. However, this is only a minor fraction (unpublished data).

6. Do not over-amplify the PCR. Make reactions with different cycle number to evaluate the optimal amplification. The high abundant fragments should not be showing up obtrusively on the agarose gel. We recommend parallel amplifications to get enough material for the subsequent sequencing.
7. It is important to remove as much adapter sequences from the fragments as possible before sequencing (see also Subheading 3.5). Test the level of adapter removal with re-amplification of the digested samples. Dilute the digested samples with 5 mM Tris–HCl pH 8.0 to 100, 1000 and 10,000 fold and amplify as before. Compare to non-digested samples (*see* Fig. 3). Repeat the digestion if necessary.
8. Recommended coverage: Based on the results obtained in rice we performed an in silico simulation to estimate the minimal coverage necessary to identify the hypomethylome of the whole genome. Reads were randomly selected to represent different coverage thresholds and allocated to the genome sequence. The data show that with a genome coverage of 3× still 92 % of the regions were identified. This corresponds to a coverage of the gene-space of ~6× [9]. Therefore, we recommend a minimal coverage of 6–7× of the gene space.
9. The dilution with random genomic fragments instead of adding PhiX sequence provides the additional advantage of having an unfiltrated dataset as reference. This may be the preferable alternative if such a reference dataset is advantageous for the analysis. Furthermore, this option is well suited for sequencing on an Illumina HiSeq machine which provides sufficient coverage for the filtrated fragments even if 30–50 % of the output derived from the random fragments. However, the use of machines with a lesser sequence output (e.g. MiSeq) is not recommended.
10. Mitochondrial and chloroplast DNA are most likely still present in the data. Depending on the questions you want to answer, it is advantageous to remove those sequences before further analysis. However, depending on the used enzyme, sequences from genomic chloroplast and mitochondrial regions can still be present even after removal of the mitochondrial and chloroplast sequences. This can be due to alternative methylation mechanisms or DNA modification which causes the enzyme to cut despite the methylation. This issue has been discussed in [9].
11. Removal of PhiX-sequence from Illumina datasets: We noticed that in most sequencing datasets a limited amount of PhiX-DNA is present even without adding it specifically. This varies between 0.1 and 5 % depending on the dataset but is on average about 1 %. We recommend removing read sequences derived from the PhiX-genome from the datasets before further analysis (see Subheading 3.6).

12. Computational power necessary for a de novo assembly: Depending on the amount of reads integrated into the analysis the RAM demand of Trinity or any other de novo assembler, might be rather high. According to the manual Trinity needs as a basic recommendation approximately ~1G of RAM per ~1 M pairs of Illumina reads. Complex datasets might require even more (https://github.com/trinityrnaseq/trinityrnaseq/wiki/Trinity-Computing-Requirements). Keep in mind that the assembly process will take some time. To reduce time and RAM demand, we recommend using the "--normalize_reads" parameter. Furthermore, paired reads can be tested prior to assembly whether they overlap and could be combined into one longer sequence using, e.g. FLASH [36].

13. Comparing de novo-assembled contigs to genomes using blast: Be aware that similar hits can occur especially in polyploid genomes or genomes with large-scale duplication events in the past. In the rice genome about 10% of the de novo-assembled contigs produced multiple occurrences with identical hit-statistics [9]. Also large gene families with small sequence divergence within the family can lead to multiple similar blast hits. For a more detailed analysis, we recommend looking at the read alignment for those regions and try to discriminate whether different haplotypes, gene copies, or different family members are present.

14. Additional annotation of identified regions: Running a similarity search against NR and NT is always recommended even if an annotation is available for the used reference. A lot of genes are annotated as "hypothetical protein" or with similar rather uninformative descriptions. This can provide at least a bit more information about the gene function. Also the available information might have changed since the annotation of the reference and new data is available. We recommend an additional search with an e-value threshold of 1e-5 to obtain additional information about the identified regions.

15. Identifying known regulatory elements in sequence data produces a high amount of false-positive data, depending on the analyzed elements. The majority of transcription factor binding sites are very short and occur everywhere in the genome simply by chance. Therefore, these data should be additionally confirmed or otherwise treated with care. However, if those elements are identified in differentially methylated regions this might give important hints to differences in the regulation of the affected gene.

16. Comparative studies: For the detection of small sequence differences between samples (e.g. SNPs, InDels, SSRs, etc.) that may hint to differential methylation of alleles between the samples or that may cause differential methylation, we recommend

performing a de novo assembly for each dataset and compare the interesting sequences separately. This approach may identify differences, which may have been excluded in a pure reference based analysis.

Acknowledgements

This work was supported by the AIT Austrian Institute of Technology GmbH.

References

1. Rabinowicz PD, Citek R, Budiman MA et al (2005) Differential methylation of genes and repeats in land plants. Genome Res 15:1431–1440. doi:10.1101/gr.4100405
2. Wang J, Marowsky NC, Fan C (2014) Divergence of gene body DNA methylation and evolution of plant duplicate genes. PLoS One 9, e110357. doi:10.1371/journal.pone.0110357
3. Springer NM, Xu X, Barbazuk WB (2004) Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. Plant Physiol 136:3023–3033. doi:10.1104/pp.104.043323
4. Palmer LE, Rabinowicz PD, O'Shaughnessy AL et al (2003) Maize genome sequencing by methylation filtration. Science 302:2115–2117. doi:10.1126/science.1091265
5. Rabinowicz PD, Palmer LE, May BP et al (2003) Genes and transposons are differentially methylated in plants, but not in mammals. Genome Res 13:2658–2664. doi:10.1101/gr.1784803
6. Raleigh EA, Murray NE, Revel H et al (1988) McrA and McrB restriction phenotypes of some E. coli strains and implications for gene cloning. Nucleic Acids Res 16:1563–1575
7. Whitelaw CA, Barbazuk WB, Pertea G et al (2003) Enrichment of gene-coding sequences in maize by genome filtration. Science 302:2118–2120. doi:10.1126/science.1090047
8. Berenyi M, Mauleon RP, Kopecky D et al (2009) Isolation of plant gene space-related sequence elements by high C+G patch (HCGP) filtration: model study on rice. Plant Mol Biol Report 27:79–85. doi:10.1007/s11105-008-0063-2
9. Wischnitzki E, Sehr EM, Hansel-Hohl K et al (2015) How to isolate a plant's hypomethylome in one shot. Biomed Res Int 2015:570568. doi:10.1155/2015/570568
10. Li D, Zhang B, Xing X, Wang T (2015) Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. Methods 72:29–40. doi:10.1016/j.ymeth.2014.10.032
11. Zhang B, Zhou Y, Lin N et al (2013) Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome Res 23:1522–1540. doi:10.1101/gr.156539.113
12. Keller J, Bancroft I. 3_CTAB_DNA_extraction. https://www.arabidopsis.org/download_files/Protocols/compleat_guide/3_CTAB_DNA_extraction.pdf
13. Nystedt B, Street NR, Wetterbom A et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584. doi:10.1038/nature12211
14. Ballal RD, Saha T, Fan S et al (2009) BRCA1 localization to the telomere and its loss from the telomere in response to DNA damage. J Biol Chem 284:36083–36098. doi:10.1074/jbc.M109.025825
15. Dlaska M, Anderl C, Eisterer W, Bechter OE (2008) Detection of circular telomeric DNA without 2D gel electrophoresis. DNA Cell Biol 27:489–496. doi:10.1089/dna.2008.0741
16. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10. doi:10.14806/ej.17.1.200
17. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi:10.1093/bioinformatics/btu170
18. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi:10.1038/nmeth.1923
19. Nussbaumer T, Martis MM, Roessner SK et al (2013) MIPS PlantsDB: a database framework

for comparative plant genome research. Nucleic Acids Res 41:D1144–1151. doi:10.1093/nar/gks1153

20. Reinert K, Langmead B, Weese D, Evers DJ (2015) Alignment of next-generation sequencing reads. Annu Rev Genomics Hum Genet 16:133–151. doi:10.1146/annurev-genom-090413-025358
21. Li H, Handsaker B, Wysoker A et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352
22. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi:10.1093/bioinformatics/btq033
23. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. doi:10.1038/nbt.1883
24. Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512. doi:10.1038/nprot.2013.084
25. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi:10.1016/S0022-2836(05)80360-2
26. Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi:10.1186/1471-2105-10-421
27. Jones P, Binns D, Chang H-Y et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. doi:10.1093/bioinformatics/btu031
28. Mitchell A, Chang H-Y, Daugherty L et al (2014) The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. doi:10.1093/nar/gku1243
29. Conesa A, Götz S, García-Gómez JM et al (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676. doi:10.1093/bioinformatics/bti610
30. Conesa A, Götz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008:1–12. doi:10.1155/2008/619832
31. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. doi:10.1093/bioinformatics/btl158
32. Fu L, Niu B, Zhu Z et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. doi:10.1093/bioinformatics/bts565
33. Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26. doi:10.1038/nbt.1754
34. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192. doi:10.1093/bib/bbs017
35. Shagin DA, Lukyanov KA, Vagner LL, Matz MV (1999) Regulation of average length of complex PCR product. Nucleic Acids Res 27, e23
36. Magoc T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963. doi:10.1093/bioinformatics/btr507