epiGBS: reference-free reduced representation bisulfite sequencing

Thomas P van Gurp[1], Niels C A M Wagemaker[2], Björn Wouters[1], Philippine Vergeer[3], Joop N J Ouborg[2] & Koen J F Verhoeven[1]

We describe epiGBS, a reduced representation bisulfite sequencing method for cost-effective exploration and comparative analysis of DNA methylation and genetic variation in hundreds of samples *de novo*. This method uses genotyping by sequencing of bisulfite-converted DNA followed by reliable *de novo* reference construction, mapping, variant calling, and distinction of single-nucleotide polymorphisms (SNPs) versus methylation variation (software is available at https://github.com/thomasvangurp/epiGBS). The output can be loaded directly into a genome browser for visualization and into RnBeads for analysis of differential methylation.

Epigenetic control by DNA methylation at cytosines (5_mC) is of paramount importance for cell regulation, differentiation and transposable element control[1]. Analysis of DNA methylation is enabled by bisulfite treatment, which converts unmethylated (but not methylated) cytosines to uracil. Subsequent sequencing of bisulfite-converted DNA allows for a quantitative estimation of DNA methylation. Various methods enable bisulfite sequencing of specific subsets of or all genomic DNA[2]. Reduced representation bisulfite sequencing (RRBS) focuses on a defined subset of genomic DNA combining restriction enzyme digestion with size selection. It was designed to target CpG islands[3], which are a common feature in mammalian, but not plant, genomes, making it non-ideal for non-vertebrate systems. Currently published RRBS methods limit the researcher in terms of enzyme choice[4] and multiplexing level[5] while requiring a reference genome for efficient mapping and variant calling[3]. Here we present epiGBS, a method that allows for straightforward, cost-effective and reference-free RRBS of highly multiplexed libraries in an accurate and versatile way. The method calls both DNA methylation polymorphisms and SNPs from the same bisulfite-converted samples while reconstructing the consensus sequence of the targeted genomic loci.

Common protocols for RRBS call for MspI digestion of genomic DNA followed by end repair, A-tailing, adaptor ligation and several purification and gel-extraction steps[3]. EpiGBS extends genotyping by sequencing (GBS)[6] with bisulfite treatment, which allows for a much simpler protocol and substantial reductions in per-sample costs. Like GBS, our protocol involves enzymatic digestion of individual samples followed by barcoded adaptor ligation and pooling of samples. Subsequently, SPRI (solid-phase reversible immobilization)-based size selection, nick translation[7], bisulfite treatment and PCR amplification are applied (**Fig. 1a**). We use nonphosphorylated adaptors to minimize adaptor dimerization. Nicks caused by the absence of 5′ phosphate groups in adaptors are repaired using nick translation, which sequentially replaces adaptor nucleotides from 5′→3′ using DNA polymerase I (**Fig. 1a**). By combining forward and reverse in-line barcoded adaptors compatible with Illumina sequencing primers, one can accomplish a 96-plex design with only 12 forward and 8 reverse adaptors containing 5-methylcytosines (**Supplementary Fig. 1**). Forward adaptors contain 4–6-base barcodes optimized for equal representation of per-cycle nucleotides during sequencing, thus minimizing phasing errors while maximizing signal intensity in Illumina sequencing[6], allowing for high-quality epiGBS libraries.

We created PstI-epiGBS libraries for several species that were pooled in a 96-plex sequencing library (**Supplementary Fig. 1e**). Average per-species inserts ranged from 134 to 193 bases (**Supplementary Fig. 1e**). Included were four *Arabidopsis thaliana* (*Arabidopsis*) samples derived from lineages with known DNA methylation from a previous whole-genome bisulfite sequencing (WGBS) study[8]. We carried out paired-end sequencing of these long fragments to facilitate efficient *de novo* clustering, reference reconstruction and BLAST-enabled functional classification.

Using the bisulfite-converted libraries only, we built a reference-free *de novo* bioinformatics pipeline by designing custom algorithms (https://github.com/thomasvangurp/epiGBS) aimed at clustering Watson and Crick reads derived from the same genomic location (same origin) (**Fig. 1b**). As bisulfite conversion targets only cytosines, unmodified guanines from the opposite strand provide the correct reference base after alignment (**Fig. 1b**). By aligning same-origin Watson and Crick reads, one can reconstruct the reference sequence *de novo* from bisulfite-converted reads (**Supplementary Fig. 2**). *De novo* reference reconstruction for inserts larger than 240 nucleotides (nt) is limited to the ~120 nt at the start and end of the locus (**Supplementary Fig. 3**).

To validate our clustering approach, we mapped *de novo*–obtained reference sequences for four *Arabidopsis* epiGBS samples to the *Arabidopsis* reference genome (TAIR10). On the basis of an *in silico* digest of TAIR10 with PstI, we expected 2,094 fragments of 11–300 nt, from which 1,862 loci (89%) could be mapped by *de novo*–obtained clusters. To assess the technical performance

[1]Department of Terrestrial Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, the Netherlands. [2]Department of Experimental Plant Ecology, Radboud University, Nijmegen, the Netherlands. [3]Plant Ecology and Nature Conservation Group, Wageningen University, Wageningen, the Netherlands. Correspondence should be addressed to T.P.v.G. (t.vangurp@nioo.knaw.nl) or K.J.F.V. (k.verhoeven@nioo.knaw.nl).

RECEIVED 8 JULY 2015; ACCEPTED 4 JANUARY 2016; PUBLISHED ONLINE 8 FEBRUARY 2016; DOI:10.1038/NMETH.3763

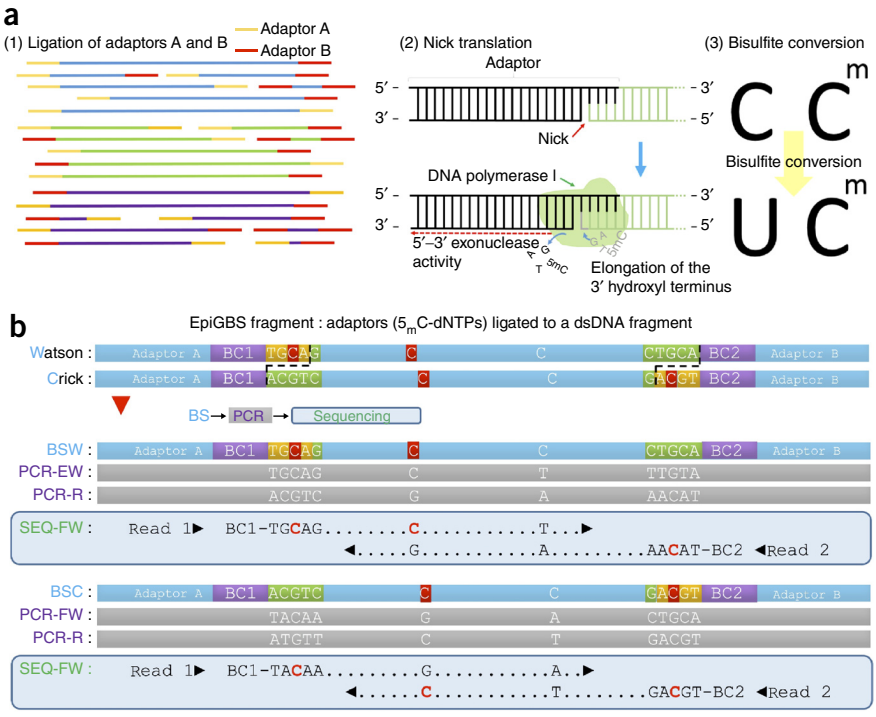

Figure 1 | Method design and results. (**a**) Genomic DNA is digested with the selected restriction enzyme for sample 1-N. After ligation of barcoded adaptors A and B (1), fragments are pooled, PCR purified and subjected to 0.8× SPRI size selection. Nick translation (2) is used to repair the nicks between the adaptor and the restriction fragment (**Supplementary Fig. 9**). Fragments are bisulfite converted (3) and PCR amplified to yield a sequencing library. (**b**) Unmethylated cytosines are bisulfite converted to uracil, whereas methylated cytosines remain intact. As only unmethylated PstI restriction sites result in digestion, the cytosines in enzyme-recognition sites (green) are converted. The cytosine in the adaptor is methylated (red) and thus remains unchanged. Eighteen cycles of PCR yield a sequencing library. Sequenced fragments contain a converted and an unconverted recognition site. The orientation of these sites is reversed for Watson and Crick reads; fragments with unconverted recognition sites on barcoded adaptor B are arbitrarily defined as Watson. BS, bisulfite sequencing; FW, forward; R, reverse.

of our *de novo* approach, we focused on a smaller subset of 1,691 genomic loci covered by at least ten Watson and ten Crick reads. An efficient *de novo* approach would produce the 1,691 clusters corresponding to the subset of genomic loci with sufficient mapped Watson and Crick reads. Clusters were mapped to 1,626 out of 1,691 loci (96%). Of these 1,626 loci, 5 were missed, 1,571 (97%) were covered once and only 50 (3%) were covered twice by the *de novo* generated clusters with a 0.1% rate of mismatch to the genomic reference sequence, confirming the sensitivity and accuracy of the *de novo* approach. The *de novo* clustering yielded 1,194 additional clusters larger than 300 nt that mapped back to the TAIR10 reference, as well as 16,498 clusters that mapped to microbial genomes or were unknown, which typically showed low coverage. In total, 96% of all mappable reads mapped to 3,056 *de novo*–obtained *Arabidopsis* clusters.

We use BWA-METH for mapping, as it provides accurate and reliable results and allows for the transfer of sample-identification information directly from FASTQ to BAM files[9]. After mapping, the resulting BAM files are split into Watson and Crick strand-mapping reads, after which variant calling is done with Freebayes[10]. Combining variant calls from both Watson and Crick strand allows one to distinguish genetic from epigenetic variation[11]. This is possible because methylation polymorphisms show as C/T (on the Watson strand) or A/G (on the Crick strand) but are nonvariable in the opposite strand, whereas genetic polymorphisms are variable on both strands. Genetic variations such as mutations in the enzyme-recognition site can lead to missing data for mutant samples. Likewise, mutant reads with structural variation can give rise to additional clusters to which only mutant reads will map, again leading to missing data. For a *de novo* analysis, samples should thus be genetically similar in order for epiGBS to find sufficient clusters with coverage for all samples. We estimated nonconversion rates on the basis of nonmethylated DNA of phage λ (Online Methods). Resulting methylation polymorphisms were converted into sample-specific methylation BED files allowing for subsequent analysis with RnBeads[12] or similar packages.

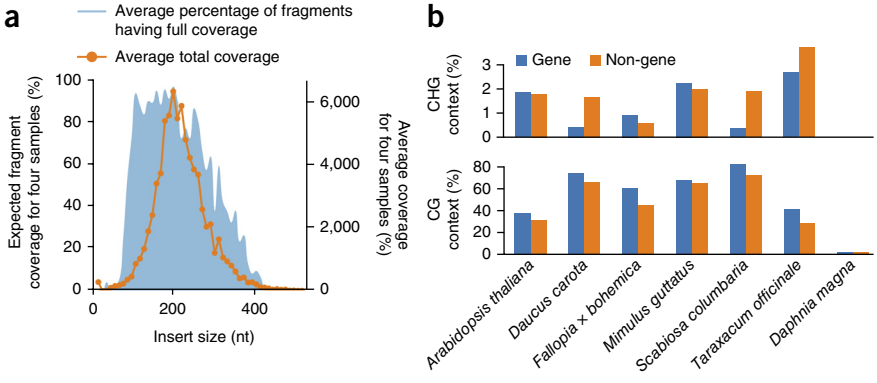

Figure 2 | Insert size and coverage in *Arabidopsis* and context-specific methylation distribution for seven species. (**a**) Insert size and coverage for *Arabidopsis thaliana*. The blue shading indicates the average percentage of sites covered per size bin of 10 nt for all four *Arabidopsis* samples sequenced, with coverage defined as present per individual per site if at least ten Watson and Crick reads are present. Approximately 90% of genomic loci with insert sizes between 170 and 220 were fully covered. The average coverage was maximal for fragments of 200 nt. (**b**) The percentage of symmetrically methylated cytosines for seven species in CG and CHG contexts for gene-related and non–gene-related clusters. Per species, we chose the four representative untreated individual specimens with the highest sequence coverage, except for *Mimulus guttatus*, *Daucus carota* and *Fallopia* × *bohemica*, for which we used only two specimens (Online Methods). Results for *Homo sapiens*, *Allium porrum* and phage λ are not shown because of insufficient coverage (Online Methods).

Comparisons between existing RRBS studies and WGBS[13,14], the current gold standard, have shown that MspI RRBS provides results that benchmark very well against WGBS data, demonstrating the accuracy and sensitivity of RRBS[2,15]. To validate epiGBS, we compared epiGBS methylation calls for four *Arabidopsis* samples to WGBS-based methylation information on previous-generation plants from an earlier study[8]. We compared methylation of cytosines with at least ten informative reads in WGBS data with positions covered by at least ten forward and ten reverse informative epiGBS reads (**Supplementary Fig. 4**). We obtained a Pearson R^2 of 0.95 for methylation levels at 12,389 cytosines in the CG context (**Supplementary Fig. 4b**). This is almost identical to the R^2 of 0.96 obtained for comparisons of intergenerational WGBS[8] values for CG methylation at the same genomic locations (**Supplementary Fig. 4a**). Correlations for cytosines in non-CG contexts were lower and showed a slight bias that is attributable to the CHG methylation sensitivity of PstI (**Supplementary Fig. 4b**). (This slight bias can be completely avoided if a methylation-insensitive restriction enzyme such as Csp6I is used.) Given sufficient coverage, this approach results in intergenerational cytosine methylation correlations for epiGBS data that are as strong as intergenerational WGBS-WGBS comparisons (**Supplementary Fig. 4d**). Also, the detection of differentially methylated positions (DMPs) on the basis of PstI-epiGBS data produced results nearly identical to those obtained via DMP calling using WGBS data (**Supplementary Fig. 5**).

In epiGBS, the number and type of targeted loci vary with genome size and restriction-site distribution; the average coverage is typically highest for fragments of around 200 nt (**Fig. 2a**). Like GBS, epiGBS is flexible with respect to the use of different restriction enzymes, and this can be exploited to bias sequencing toward (or away from) specific genomic features. Using PstI in *Arabidopsis*, we targeted 2,260 loci with coverage of >100, representing ~0.37% of genomic DNA. As PstI is sensitive to CHG methylation, repetitive DNA in *Arabidopsis* is largely avoided (**Supplementary Fig. 6**), which allows one to focus the sequencing effort on coding regions where most differentially methylated regions in *Arabidopsis* are located[8]. Investigating a biased subset of the genome also means that PstI-based epiGBS methylation characterization is not necessarily representative of genome-wide methylation patterns, but such biases can be overcome easily with the use of a different (methylation-insensitive) enzyme, as we demonstrated using Csp6I (**Supplementary Fig. 4c**).

Our bioinformatics pipeline produces methylation and SNP variant call files as well as methylation-level tabular files suitable for visualization in a genome browser such as IGV (**Supplementary Fig. 7**) or the recently published RnBeads pipeline[12], providing seamless integration and analysis of DMPs (Online Methods).

Our analysis confirms a prior observation in model species: cytosine methylation in the CG context is higher in genes than in non-gene clusters (**Fig. 2b**). We detected DMPs in all species. Strikingly, we detected more CHG DMPs in nonmodel plant species than in *Arabidopsis* (**Supplementary Fig. 8**). Our analysis of *Daphnia magna*, an important ecotoxicological model species, provides compelling evidence for the occurrence of DNA methylation in the CG context in this aquatic species[16] (**Fig. 2b**). Our method can thus be used to expand knowledge on the presence of DNA methylation in population or evolutionary studies of nonmodel species lacking a reference genome. Thus epiGBS provides a significant improvement over existing marker-based methylation screening tools such as MSAP[17], allowing for much more comprehensive studies on differences in DNA methylation in nonmodel species.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Raw reads have been deposited at the NCBI short read archive under accession PRJNA287755.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank F. Johannes (Department of Plant Science, Technical University of Munich, Munich, Germany) and C. Becker (Max Planck Institute for Developmental Biology, Tübingen, Germany) for providing *Arabidopsis* samples, and we thank G. Maes, J. van Houdt and E. Schijlen for sequencing several epiGBS libraries as we developed the protocol. *Daphnia magna* samples were provided by L. De Meester (University of Leuven, Leuven, Belgium). *Fallopia japonica* and *Fallopia* × *bohemica* samples were provided by C. Richards (University of South Florida, Tampa, Florida, USA). *Allium porrum* and *Daucus carota* seeds were provided by Nunhems Netherlands BV–Bayer CropScience. Finally, we thank C. Bock, C. Richards, R. Elshire, W. van der Putten and A. Biere for critical comments and helpful suggestions on the manuscript. This study was funded by the Netherlands Organization for Scientific Research (NWO-ALW grants 864.10.008 and 820.01.025 to K.J.F.V.).

AUTHOR CONTRIBUTIONS

T.P.v.G. conceived the method, performed lab work, developed software, analyzed data and wrote the paper. N.C.A.M.W. codeveloped the method, provided sample material, performed lab work and contributed to writing of the paper. B.W. codeveloped software and developed the RnBeads interface scripts. K.J.F.V. contributed to interpretation of the experiment and method and to writing of the paper. P.V. provided samples for the *Scabiosa* case study and contributed to interpretation of the experiment and method. J.N.J.O. contributed samples and to interpretation of the experiment and method.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Suzuki, M.M. & Bird, A. *Nat. Rev. Genet.* **9**, 465–476 (2008).
2. Beck, S. *Nat. Biotechnol.* **28**, 1026–1028 (2010).
3. Gu, H. *et al. Nat. Protoc.* **6**, 468–481 (2011).
4. Wang, J. *et al. BMC Genomics* **14**, 11 (2013).
5. Boyle, P. *et al. Genome Biol.* **13**, R92 (2012).
6. Elshire, R.J. *et al. PLoS ONE* **6**, e19379 (2011).
7. Rigby, P.W.J., Dieckmann, M., Rhodes, C. & Berg, P. *J. Mol. Biol.* **113**, 237–251 (1977).
8. Becker, C. *et al. Nature* **480**, 245–249 (2011).
9. Pedersen, B.S., Eyring, K., De, S., Yang, I.V. & Schwartz, D.A. Preprint at http://arxiv.org/abs/1401.1129 (2014).
10. Garrison, E. & Marth, G. Preprint at http://arxiv.org/abs/1207.3907 (2012).
11. Liu, Y., Siegmund, K.D., Laird, P.W. & Berman, B.P. *Genome Biol.* **13**, R61 (2012).
12. Assenov, Y. *et al. Nat. Methods* **11**, 1138–1140 (2014).
13. Lister, R. *et al. Cell* **133**, 523–536 (2008).
14. Cokus, S.J. *et al. Nature* **452**, 215–219 (2008).
15. Harris, R.A. *et al. Nat. Biotechnol.* **28**, 1097–1105 (2010).
16. Vandegehuchte, M.B. *et al. Environ. Int.* **35**, 700–706 (2009).
17. Schrey, A.W. *et al. Integr. Comp. Biol.* **53**, 340–350 (2013).

ONLINE METHODS

Description of samples. We collected DNA from individual specimens of several species to showcase the versatility of the epiGBS method. For several species, we used samples treated with 5-azacytidine, genistein or other stressors. In the per-species DNA methylation analysis presented in **Figure 2b** and **Supplementary Figure 8**, we included only untreated control samples. We have included details on the individual treatments per species to facilitate further analysis of the data presented here (**Supplementary Data 1**).

Scabiosa columbaria samples originated from an earlier experiment[18]. For this experiment, seeds were collected in a large French calcareous grassland and used to grow plants in the greenhouse of Radboud University Nijmegen. Plants were crossed, and F1 seeds were germinated on filter paper saturated with water or 50 µM 5-azacytidine[18]. Seedlings were transferred to either a soil mixture resembling calcareous soil (pH 7.5–8.0) or a soil mixture resembling acidic soil (pH 6.0–6.5). After 3 months, young leaves and plants were sampled for DNA extraction using the Macherey-Nagel Nucleospin Plant II kit for individual tubes. We used 400 ng of genomic DNA (gDNA) for the epiGBS analysis. Plants that were included in the epiGBS analysis all originated from the same maternal line.

Mimulus guttatus seeds were collected from a large Dutch population of plants in the vicinity of Arnhem, the Netherlands. The seeds were germinated and grown under controlled conditions in the greenhouse of Radboud University Nijmegen. Juvenile plants were divided into groups and subjected to one of four experimental conditions: control, drought, water-logged and submerged. After 4 weeks, young leaves were sampled for DNA extraction using the Macherey-Nagel Nucleospin Plant II kit for individual tubes. Eight samples (two control, two drought, two waterlogged and two submerged) were included in the epiGBS analysis (400 ng of gDNA from each sample).

Allium porrum and *Daucus carota* seeds were provided by Nunhems Netherlands BV–Bayer CropScience. Seeds of an inbred and hybrid line of each crop were germinated and grown under controlled conditions in the greenhouse of Radboud University Nijmegen. Seeds were germinated on filter paper saturated with water (control) or 40 µM genistein solution. After 4 weeks, young leaf tissue was sampled for DNA extraction using the Macherey-Nagel Nucleospin Plant II kit for individual tubes. We used 400 ng of gDNA per sample for the epiGBS analysis.

Daphnia magna samples were provided by Professor Luc De Meester (University of Leuven). Inbred families were obtained through clonal selfing within clones that were isolated from the LangeRodevijver fishpond near Leuven, Belgium. Outbred families were derived from resting eggs, collected from the dormant egg bank of the LangeRodevijver fishpond, that were hatched in the laboratory. More details can be found in ref. 19. Six samples (three outbred and three inbred individuals from the second clutch) were used in the epiGBS analysis (400 ng of gDNA of each sample).

Fallopia japonica and *Fallopia* × *bohemica* samples were provided by Professor Christina Richards (University of South Florida). Rhizomes were collected from field sites in Long Island, New York, USA, and subsequently propagated in the greenhouse to provide fresh leaf tissue. DNA was extracted from the third uppermost fully expanded fresh leaf using the Qiagen DNeasy plant mini kit. All samples were used to create a *de novo* reference, but only two samples (301 and 314) of *Fallopia* × *bohemica* (both from roadside habitats) were included in the epiGBS analysis (400 ng of gDNA of each sample).

Taraxacum officinale lines A68 and A34 are clonal apomictic dandelion lines[20] that have been maintained under greenhouse conditions for several generations; leaf material from several individuals in these lines from an ongoing experiment was collected over the course of several years. The leaf material was stored at −80 °C for 1–2 years. DNA isolation was performed for all samples simultaneously as described for the other samples. Plants were grown until flowering was completed under controlled conditions in a greenhouse (day/night temperatures of 21 °C/17 °C, 16-h photoperiod).

Arabidopsis thaliana seeds were provided by Claude Becker (Max Planck Institute for Developmental Biology) via Frank Johannes (Technical University of Munich). Plants from the previous generation of the same lineage were subjected to WGBS[8]. For each lineage, up to five individual plants were grown for 4 weeks under controlled conditions in a greenhouse (day/night temperatures of 21 °C/17 °C, 16-h photoperiod). A single plant per line was used for DNA isolation.

Human DNA (female) was bought from Promega (G1521).

Phage λ DNA was bought from Promega (cl857 Sam7; D1521). The phage λ cl857 Sam7 DNA was isolated from infected GM119, an *Escherichia coli* strain lacking both dam and dcm methylase activity (Promega).

Further sample details are available in **Supplementary Data 1**, in which per-sample genotypes, treatments, tissues, ages and barcodes used are specified for all species.

DNA extraction. We homogenized plant material by bead-beating frozen leaf tissue in a 2-mL Eppendorf tube with 2–3-mm stainless steel beads. No more than 100 mg of fresh tissue was used per sample. Samples with beads were taken from the freezer and stored in liquid nitrogen. For each batch of 12 samples, the tubes were put in a TissueLyser 24 adaptor set block that was partially submerged in liquid nitrogen to prevent thawing during sample placement. After 30 s of shaking at 30 Hz and 1,800 oscillations per minute, the tubes were submerged in liquid nitrogen, after which a second 30-s round of shaking was performed.

For most plant species mentioned, we isolated DNA using the Macherey-Nagel Nucleospin Plant II kit (details in **Supplementary Data 1**) for individual tubes. We followed the manufacturer's protocol with the following modifications. For cell lysis we used cell lysis buffer PL1 for 30 min instead of 10 min. After filtration through the pink nucleospin filter, we carefully pipetted the flow-through into a fresh 1.5-mL tube, avoiding the pellet that is often formed at this stage. We used an additional centrifugation step to avoid a small pellet; the clear supernatant was used in the following steps according to the manufacturer's protocol. As restriction enzymes are very sensitive to proteins and other contamination, we selected only samples with high purity (specifically, 260/280 and 260/230 ratios of at least 1.8 and 1.5, respectively). DNA concentration was determined using the Qubit 2.0 Fluorometric dsDNA HS assay kit (Q32851; Life Technologies).

Restriction digestion. Per individual specimen, 400 ng of gDNA was digested overnight (17 h) at 37 °C in a volume of 40 µL

containing 1× NEBuffer 3.1, 125 µg of BSA (NEB; B9000S) and 40 units of PstI (NEB; R0140S) in a 2-µl volume. After digestion, barcoded adaptors were ligated to the fragments.

Adaptor ligation. To minimize the possibility of misidentifying samples as a result of sequencing or adaptor synthesis error, we ensured that all pair-wise combinations of barcodes differed by a minimum of three mutational steps, and we modulated barcode lengths from 4 to 6 bp to maximize the nucleotide balance of the bases at each position in the overall set of sequencing reads (**Supplementary Fig. 1d**). Samples were pooled and processed per species after ligation. For the ligation, we combined 1,200 pg of both forward and reverse barcoded adaptors (**Supplementary Fig. 1**) in a 60-µL reaction containing 40 µL of gDNA digest, 1× T4 DNA ligase buffer and 4,000 units of T4 DNA ligase (NEB; M0202M/L) and ligated for 3 h at 22 °C and then overnight at 4 °C, with no inactivation afterward.

Cleanup and size selection. To assess the quality of libraries, we performed the pooling per species. After pooling, the total library volume was reduced by Qiaquick PCR cleanup (Qiagen; 28104) to 60 µL. The libraries were size selected by a 0.8× Agencourt AMPure XP (Beckman Coulter; A63880) purification favoring >200-bp DNA fragments and eluted in a total volume of 24 µL.

Nick translation. Due to the use of nonphosphorylated adaptors, epiGBS libraries contain nicks between the 3′ fragment overhang and the 5′ nonphosphorylated adaptor nucleotide. To prevent the loss of ssDNA adaptor strands (at the nicked position) during bisulfite treatment, we repaired nicks (**Supplementary Fig. 9**) via a 1-h nick translation reaction at 15 °C in a reaction volume of 25 µL containing 18 µL of the purified library, 2.5 µL of 10 mM 5-methylcytosine dNTP mix (Zymo Research; D1030), 1× NEBuffer 2 and 7.5 units of DNA polymerase I (NEB; M0209S).

Optional GBS PCR. At this stage we performed an optional GBS PCR to check the library quality, using the epiGBS PCR protocol (described below). Amplified GBS libraries were expected to be bigger on average than amplified epiGBS libraries. We assessed the quality of the PCR libraries by analyzing 1 µL of the PCR product on a high-sensitivity DNA chip on a 2100 Bioanalyzer system (Agilent).

Bisulfite treatment and purification. For bisulfite treatment, we used 20 µL of the nick-translated library. Bisulfite treatment was performed using the EZ DNA Methylation-Lightning kit (Zymo Research) with the following program: 8 min at 98 °C, then 1 h at 54 °C followed by up to 20 h at 4 °C, all according to the manufacturer's protocol.

EpiGBS PCR. Library amplification was performed per species in four individual 10-µL reactions containing 1 µL of ssDNA template, 5 µL of KAPA HiFi HotStart Uracil+ ReadyMix (Kapa Biosystems), and 3 pmol of each Illumina PE PCR primer (**Supplementary Fig. 1b**). Temperature cycling consisted of 95 °C for 3 min followed by 18 cycles of 98 °C for 10 s, 65 °C for 15 s and 72 °C for 15 s, with a final extension step at 72 °C for 5 min. Replicate PCR products were pooled and quantified using a Qubit dsDNA HS assay kit (Life Technologies). We assessed the quality of the libraries by analyzing 1 µL on a high-sensitivity DNA chip on a 2100 Bioanalyzer system (Agilent). Libraries were considered suitable for sequencing if the majority of DNA fragments were between 150 and 400 bp in size and no adaptor dimers were found. Typically, epiGBS PCR reactions of 18 cycles of a non-pooled plant sample yield 3–12 ng/µL of PCR product.

When the per-species pooled libraries passed quality control, they were pooled further according to the concentration and number of samples in the species pool such that individual samples were expected to yield an equal number of clusters on the Illumina flowcell. We carried out a 'nano run' on the Illumina MiSeq to quantify the per-sample read count. On the basis of the read counts obtained from that run, we pooled the individual nick-translated digestion-ligations in such a manner that an equal number of reads was expected per individual. Finally, we performed rapid run–mode paired-end sequencing on an Illumina HiSeq2500 sequencer using the HiSeq v4 reagents and the latest version of the HiSeq Control software (v2.2.38), which optimizes the sequencing of low-diversity libraries (https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote-hiseq-low-diversity.pdf). As the first five cycles of a sequencing run are used to calculate the color matrix, our barcode design achieves an almost perfect balance of the first 5 nt when equal numbers of sequences are obtained per forward read, or 'A', barcode. The reverse read, or 'B', barcodes do not have this requirement; hence only barcodes of 4 nt were used (**Supplementary Fig. 1a**).

Csp6I laboratory work. The Csp6I epiGBS libraries were constructed similarly to the PstI epiGBS libraries, with the following modifications: The restriction digestion reaction contained 1× FastDigest buffer and 40 units of Csp6I (Thermo Fisher Scientific; FD0214 in a 4-µl volume). The ligation reaction contained 2,400 pg of both A and B adaptors (both adjusted for the Csp6I sticky end). Whereas in the PstI protocol we used fully methylated adaptors (i.e., methylation of both strand I and strand II), for the Csp6I protocol we used hemimethylated adaptors. The adaptor strands that were resynthesized (incorporating 5_mC dNTPs) by nick translation were not methylated, as all cytosines are replaced by methylated 5_mC (**Supplementary Fig. 9**). Final amplification for the Csp6I library yielded 4–8 ng/µL product for an epiGBS PCR of 18 cycles of a library containing only *Arabidopsis* sample A29.

Bioinformatics analysis. All custom Python scripts mentioned are available on GitHub (https://github.com/thomasvangurp/epiGBS). Processed data are available at GenomeSpace (https://gsui.genomespace.org/jsui/gsui.html?pathOrUrl=/Home/thomasvangurp/epiGBS). The bioinformatics pipeline design is visualized in **Supplementary Figure 10**.

Demultiplexing. Paired-end reads for both the PstI-based titration MiSeq nano run and the HiSeq2500 run were demultiplexed using the custom Python script demultiplex.py with an appropriate barcode to sample file. Up to two mismatches were allowed in the barcode and enzyme-recognition site on both forward and reverse reads. During demultiplexing, read group attributes specifying the sample name were added to the read name, whereas the barcode was stripped from both forward and reverse reads. Additionally, a sample type tag (ST:Z:watson or ST:Z:crick) was added according

doi:10.1038/nmeth.3763

to the orientation of the enzyme-recognition sites in the paired reads (**Fig. 1b**).

De novo reference construction. The paired-end reads were split per species using Unix grep. *De novo* reference construction was carried out on a per-species level. Reads were merged using PEAR v0.9.5 (ref. 21) with the following settings: minimum *P* value for accepting an overlap, 0.001; minimum overlap, 10; no trimming; minimum assembly length, 0. This resulted in merged and nonmerged reads. Nonmerged reads were concatenated with 10 N-nucleotides between the forward and reverse read (**Supplementary Fig. 3**). Both merged and nonmerged reads were split in Watson and Crick reads using custom Unix grep queries. Methylation polymorphisms in Watson and Crick reads were removed using Unix sed, with all C's replaced with T's in Watson reads and all G's replaced with A's in Crick reads. Computationally derived 'demethylated' sequences were dereplicated using USEARCH (http://www.drive5.com/usearch/) with the dereplication command. Only clusters consisting of two or more reads were retained. To enable pairing of same-origin Watson and Crick reads, C's were converted to T's in Watson reads, and G's were converted to A's in Crick reads, yielding a binary AT-only sequence output that is identical for same-origin Watson and Crick reads (**Supplementary Fig. 2**). The reference sequence per pair is called for both Watson and Crick reads using combined BASH piped queries using SAMtools mpileup and BCFtools[22] for variant calling followed by vcfutils.pl vcf2fq (https://github.com/lh3/samtools/blob/master/bcftools/vcfutils.pl) for creating the Watson and Crick consensus sequence. Watson and Crick consensus sequences are used to recreate the original sequence from which the bisulfite-converted reads were derived. Bases from both Watson and Crick consensus sequences are processed simultaneously per position with create_consensus.py. In the case of Watson:T/Crick:C, a C is added to the reference sequence output, whereas in the case of Watson:G/Crick:A, a G is added. If paired nucleotides match, that nucleotide is added. In all other cases where paired nucleotides do not match, an ambiguous nucleotide (N) is added (**Supplementary Fig. 11**).

Trimming. Part of the read sequence in both Watson and Crick reads originates from fully methylated adaptors (**Supplementary Fig. 1c**). To exclude these adaptor nucleotides, the first 4 nt corresponding to the adaptor sequence ligated to the enzyme-recognition site are trimmed. For this, we trimmed Watson reads on forward (/1) reads and trimmed Crick reads on reverse (/2) reads. In merged Watson reads the first 4 nt were removed, whereas in merged Crick reads the last 4 nt were removed.

Mapping. Per species, we mapped merged and unmerged reads using a modified version of bwameth.py (https://github.com/thomasvangurp/epiGBS) with the default settings. The resulting BAM files for both merged and unmerged reads were sorted and merged. Read-group identifiers corresponding to the sample-name identifiers present in the mapped reads were added using the SAMtools command reheader. Both Watson and Crick BAM files were indexed using SAMtools.

Strand-specific variant calling. Variant calling was done with Freebayes[10] per species on both Watson and Crick BAM files separately using a custom Python script (map_reads.py module runFreebayes). This module runs Freebayes in parallel with settings that force variants to be called on all positions:

```
freebayes -f consensus_cluster.renamed.fa -F 0 -E 1 \
-C 0 -G 0 --haplotype-length 1 \
--report-all-haplotype-alleles --report-monomorphic\
--report-genotype-likelihood-max \
--haplotype-length 1 -KkXuiwaq 21
```

The resulting variant call files (VCFs) from the parallel runs done for both Watson and Crick BAM files were merged, compressed with bgzip, and indexed using the SAMtools module tabix (http://www.htslib.org/doc/tabix.html), resulting in a VCF for both the Watson and the Crick strand.

Methylation calling. The custom Python script methylation_calling.py uses the 'walktogether' method of the PyVCF package (https://github.com/jamescasbon/PyVCF) to simultaneously iterate over both Watson and Crick VCFs. In this way, SNPs and methylation polymorphisms were distinguished and split. C/T polymorphisms in the Watson strand combined with C on the Crick strand indicate a methylation polymorphism on the Watson strand, whereas a G/A polymorphism on the Crick strand combined with a G on the Watson strand indicates a methylation polymorphism on the Crick strand. Where combined SNP and methylation lead to a C/T or G/A polymorphism on both Watson and Crick strands, only the SNP is called, as the methylation ratio cannot be reliably determined. Per-site, per-individual methylation levels are available in various output formats, such as VCFs, IGV files and BED files, to provide for analysis and visualization of the methylation variation.

Methylation summary statistics calculation. Per-species similarity searches for reference sequences of the clusters with an appropriate proteome were done with USEARCH UBLAST[23]. A gene hit for a plant species was defined as a cluster having an e-value of <1e–5 after mapping to the reference proteome (RefSeq) of all eudicot plant species. For *Daphnia magna*, all *Daphnia*-related proteins (NCBI) were used as a reference. Per species, each cluster was labeled "gene" or "non-gene" on the basis of USEARCH UBLAST. Methylation is summarized for cytosines in symmetric CG and CHG contexts for both gene-related and non–gene-related clusters (**Supplementary Fig. 1e**).

The following criteria were used to determine whether positions should be considered:

(1) Per-position minimum read coverage for forward and reverse informative reads > 10.
(2) For cytosines on the top strand, only reads from the Watson pool are informative, whereas for cytosines on the bottom strand, only Crick reads are informative. Positions with SNPs or ambiguous contexts due to neighboring SNPs are excluded from this analysis.
(3) Minimum methylation ratio for both forward and reverse informative reads > 0.05; otherwise the position is considered nonmethylated.
(4) Only symmetric positions in CG and CHG contexts with sufficient coverage according to criterion 1 are taken into

account. For these symmetric positions, for both Watson and Crick strand nucleotides, criterion 3 needs to apply; otherwise the symmetric pair is considered nonmethylated.

(5) Methylation is summarized for symmetric pairs for all positions meeting the above criteria in up to four nontreated (control) samples having the highest coverage per species. (For *Mimulus guttatus*, *Daucus carota* and *Fallopia* × *bohemica*, only two individuals could be included.)

Methylation summary statistics calculation was performed for all species included with sufficient coverage. *Allium porrum* and *Homo sapiens* were excluded from the analysis because of insufficient sequence coverage. Statistics were computed in the submodule summary_methylation of the same Python script used for DMP detection in all species (all_DMP_detection.py).

Csp6I bioinformatics. For the Csp6I library, we mapped all merged paired-end reads to the *Arabidopsis* TAIR10 genome. As the origin (Watson or Crick) of the merged reads could not be established owing to the absence of a methylation-sensitive restriction enzyme–recognition site as is present for PstI, we adjusted the procedure for mapping the reads. All merged reads were computationally demethylated using Unix sed in two different ways, creating a C-to-T copy and a G-to-A copy for each read. These two different computationally derived copies were subsequently mapped with BWA-MEM to a C-to-T converted and G-to-A converted genome (also created using Unix sed). Only one of these four mapped copies had a meaningful and high-scoring match. To find that hit, we sorted the four resultant BAM files by name and retained the highest scoring match if it was matching over 80% of its length. This procedure mimics what BWA-METH normally does internally but could not do given our Csp6I-based epiGBS data.

We produced Watson and Crick BAM files by adding the original nonconverted sequence in its proper orientation to the appropriate BAM file. Methylation calling was done using SAMtools mpileup through a custom Python script (available at https://github.com/thomasvangurp/epiGBS under *Arabidopsis*_extra/*Arabidopsis*_Csp6i_analysis.py). Comparisons between Csp6I methylation estimates obtained for *Arabidopsis* generation 32 line 29 and corresponding generation 31 WGBS estimates from ref. 8 were obtained in the same way as the PstI-epiGBS data described above. BAM files with mapped reads for Csp6I are available on GenomeSpace (https://gsui.genomespace.org/jsui/gsui.html?pathOrUrl=/Home/thomasvangurp/epiGBS%20Nature%20Methods/Csp6i#).

DMP detection in all species. DMPs were detected in up to six pairwise comparisons per species for up to four individual specimens (**Supplementary Data 2**) using nontreated samples having the highest sequence coverage (Online Methods) with a procedure similar to that for *Arabidopsis* (**Supplementary Fig. 5**), except that in this analysis we did not require a minimum ratio difference of 70% or higher, as we also included the CHG context, in which differences are typically smaller. DMP frequency was established for cytosines in CG and CHG contexts in the gene-related and non–gene-related (other) clusters separately. We limited DMP detection to cytosines that were either methylated or nonmethylated symmetrically (on both Watson and Crick strands, with both below or both higher than 5%) in the CG or CHG context. To determine which positions to consider in the DMP analysis, we applied the same criteria as used for detection of methylation (see "Methylation Summary Statistics"). All calculations were done with the custom Python script all_DMP_detection.py, available in the GitHub repository (https://github.com/thomasvangurp/epiGBS).

Nonconversion and false methylation rate estimates. We computed nonconversion rates on the basis of nonmethylated phage λ DNA. A total of 8,475 cytosines of the phage λ genome were covered with 100 or more reads. On average, 0.39% of cytosines per position were not converted. We do not think this nonconversion rate had an impact on our methylation estimates, as in our analysis we required coverage to be higher than 10 and methylation rates to be higher than 5% in read pools of both forward and reverse informative reads (see "Methylation Summary Statistics"). With this criterion, not a single cytosine in phage λ was found to be methylated at a rate greater than 5%.

Code availability and source data. All custom Python scripts mentioned in the paper are available on GitHub (https://github.com/thomasvangurp/epiGBS). Source data for the Supplementary Figures can be found in **Supplementary Data 3**.

18. Vergeer, P., Wagemaker, N.C.A.M. & Ouborg, N.J. *Biol. Lett.* **8**, 798–801 (2012).
19. Swillen, I., Vanoverbeke, J. & De Meester, L. *Ecol. Evol.* **5**, 2712–2721 (2015).
20. Verhoeven, K.J.F., van Dijk, P.J. & Biere, A. *Mol. Ecol.* **19**, 315–324 (2010).
21. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. *Bioinformatics* **30**, 614–620 (2014).
22. Li, H. *et al. Bioinformatics* **25**, 2078–2079 (2009).
23. Edgar, R.C. *Bioinformatics* **26**, 2460–2461 (2010).

doi:10.1038/nmeth.3763