Large-scale comparative epigenomics reveals hierarchical regulation of non-CG methylation in *Arabidopsis*

Yu Zhang, C. Jake Harris, Qikun Liu, Wanlu Liu, Israel Ausin, Yanping Long, Lidan Xiao, Li Feng, Xu Chen, Yubin Xie, Xinyuan Chen, Lingyu Zhan, Suhua Feng, Jingyi Jessica Li, Haifeng Wang, Jixian Zhai, and Steven E. Jacobsen

*Institute of Plant and Food Science, Department of Biology, Southern University of Science and Technology, 518055 Shenzhen, China; *Institute for Advanced Studies, Wuhan University, 430072 Wuhan, China; *College of Life Science, Wuhan University, 430072 Wuhan, China; *Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA 90095; *Basic Forestry and Proteomics Research Center, Fujian Agriculture and Forestry University, 350002 Fuzhou, China; *UCLA-FAFU Joint Research Center on Plant Proteomics, University of California, Los Angeles, CA 90095; *Department of Statistics, University of California, Los Angeles, CA 90095; *State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Fujian Agriculture and Forestry University, 350002 Fuzhou, China; and *Howard Hughes Medical Institute, University of California, Los Angeles, CA 90095

Contributed by Steven E. Jacobsen, December 16, 2017 (sent for review September 15, 2017; reviewed by Paoyang Chen and Daniel Schubert)

Whole-genome bisulfite sequencing (WGBS) has become the gold standard for DNA methylation analysis, and a tremendous amount of WGBS data has been generated by the research community. However, the systematic comparison of DNA methylation profiles to identify regulatory mechanisms has yet to be fully explored. Here we reprocessed the raw data of over 500 publicly available *Arabidopsis* WGBS libraries to reveal unanticipated connections between well-known genes. For instance, MET1 and CMT3 were found to be required for the maintenance of asymmetric CHH methylation at nonoverlapping regions of CMT2 targeted heterochromatin. Our comparative methylation approach has established a framework for extracting biological insights via large-scale comparison of methylation sites and can also be adopted for other genomics datasets.

**Significance**

In plants, DNA cytosine methylation plays a central role in diverse cellular functions, from transcriptional regulation to maintenance of genome integrity. Vast numbers of whole-genome bisulfite sequencing (WGBS) datasets have been generated to profile DNA methylation at single-nucleotide resolution, yet computational analyses vary widely among research groups, making it difficult to cross-compare findings. Here we reprocessed hundreds of publicly available *Arabidopsis* WGBS libraries using a uniform pipeline. We identified high-confidence differentially methylated regions and compared libraries using a hierarchical framework, allowing us to identify relationships between methylation pathways. Furthermore, by using a large number of independent wild-type controls, we effectively filtered out spontaneous methylation changes from those that are biologically meaningful.


Reviewers: P.C., Academia Sinica; and D.S., Freie Universität Berlin.

The authors declare no conflict of interest.

This open access article is distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The sequences reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo (accession no. GSE98872).

Y.Z., C.J.H., and Q.L. contributed equally to this work.

To whom correspondence may be addressed. Email: hafengwang@fasu.edu.cn, zhajix@usc.edu.cn, or jacobson@ucla.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1716300115/-/DCSupplemental.
Our analysis in different mutants revealed a previously overlooked hierarchical framework regulating non-CG methylation and established connections between different epigenetic regulators. For example, MOM1 and MORC family proteins coordinately target a small but specific subset of RNA-directed DNA methylation (RdDM) regions, whereas MET1 and CMT3 are each required for CHH methylation at unique subsets of CMT2 targeted regions. This framework could be adopted to perform large-scale methylome comparisons in other model organisms or for other types of NGS data.

Results and Discussion

Uniform Processing and Quality Check of Arabidopsis WGBS Libraries in GEO. We retrieved 503 Arabidopsis whole-genome bisulfite sequencing libraries from the National Center for Biotechnology Information GEO database (17), which included a wide spectrum of genotypes, tissues, and treatments (Dataset S1). Considering the variation in sequencing depth and quality among these libraries, we developed a uniform procedure to process all libraries and assess their quality (Fig. 1; see Materials and Methods for more details). We excluded libraries with low efficiency of bisulfite conversion, libraries with low coverage, libraries that were not in the reference Col-0 background, and libraries that represent duplicated GEO entries (Fig. 1 B and C and Dataset S2). In total, these quality control steps filtered out 189 libraries. The remaining 314 high-quality WGBS libraries, including 54 designated as “control” libraries (this set includes all libraries of wild-type leaf or seedling tissue—these are the most common tissue types submitted for WGBS analysis) and 260 “test” libraries (this set includes all non-WT genotypes, treatments, or nonleaf/seedling tissue types), were selected for further analysis.

Identification of High-Confidence Differentially Methylated Regions.

To establish connections among the different WGBS libraries, we first evaluated the changes in methylation (differentially methylated regions (DMRs)) for each genotype/tissue/treatment by comparing each of the 260 test libraries to each of the 54 control libraries (see Materials and Methods for more details and Datasets S3 and S4). In brief, we defined six types of DMR (hyper- or hypo-CG/CHG/CHH) and performed DMR calling with 100-bp bin resolution for each test library. A DMR is only valid when the test library differs from at least 33 of the 54 control libraries (see Materials and Methods for more details; SI Appendix, Figs. S1 and S2; and Dataset S3). These DMRs were designated as “high-confidence” DMRs (hcDMRs) and are listed in Datasets S5–S10, and the hcDMR calling pipeline is available for download at https://github.com/yu-z/hcDMR_caller. By design, hcDMRs should filter out spontaneous DMRs that occur in wild-type plants (18, 19) through comparison with a large number of control libraries. To validate the hcDMRs, we sought to complement a mutant because true DMRs arising as a consequence of the genetic knockout should be restored after the reintroduction of a functional copy of the gene, whereas DMRs that arise spontaneously should not. We chose to complement morc6 because this mutant is known to cause very modest changes in methylation (20, 21), allowing us to assess both the accuracy and sensitivity of the hcDMR calling method. We reintroduced a FLAG tagged version of MORC6 into the morc6 mutant background and performed RNAseq in the first (T1) generation alongside wild-type and morc6 mutant controls. This confirmed the functional activity of the MORC6 transgene because nearly all genes derepressed in the morc6 mutant were restored to wild-type levels (SI Appendix, Fig. S3). Next, we performed WGBS on the complementation line and found that whereas only 3% of hcDMRs were not complemented (9311), ∼26% of DMRs derived by comparing morc6 to a matched wild-type control (22) failed to regain methylation by 10% or more (3551,391) (Fig. 2 A and B). The increased rate of false positives using the matched control method could be the result of the different life histories of the laboratory strain wild-type and morc6 T-DNA plant. Each acquires independent spontaneous methylation changes through generations, which are filtered out using the hcDMR method, but would be identified using a matched control method because the methylation differences are real, although unconnected to the underlying genetic mutation. These results indicate that our pipeline is both sensitive and robust, identifying true DMRs even for a mutant with relatively weak methylation defects.

Validation and Comparison of hcDMRs to Existing DMR Calling Methods.

To extend our comparison of hcDMRs to the standard method for DMR detection of comparing matched wild-type controls, we utilized three nrpe1 mutant libraries which have been sequenced by independent laboratories alongside wild-type controls from the same studies (23–25). Although the total number of DMRs identified using the laboratory matched control method was higher compared with hcDMRs, the intersect of DMRs among these three nrpe1 libraries was much lower than that of the hcDMR method (∼30% in laboratory matched compared with ∼70% in hcDMRs) (Fig. 2C). This indicates that a high proportion of laboratory matched control identified DMRs may represent false positives. Additionally, the hcDMR method identified 5,296 DMRs that are commonly shared among all three nrpe1 samples, whereas the matched control method identified only 2,901 common DMRs (Fig. 2C), suggesting that hcDMRs more accurately reflect DMRs that result from the mutation. In conclusion, the hcDMR method allows us to extract and identify a robust set of DMRs from a mutant genotype that is broadly independent of the laboratory of origin.

Statistical and Quantitative Measurement of Overlapping of DMRs and Library Clustering. Having established hcDMRs for each library, we sought to use this information to gain insight into the relatedness and relationships between the libraries. To evaluate the degree of overlap between libraries, we first calculated the probability of obtaining the observed number of shared DMRs under the assumption of obtaining the observed number of shared DMRs under the same studies (23–25). Thus, this probability was calculated in a pairwise manner for each test library against every other test library, and we refer to this method as S-MOD (Materials and Methods and Fig. 3A). Thus, for
The Arabidopsis H3K9 demethylase, IBM1, removes H3K9me2 in gene bodies to prevent the establishment of CHG methylation (26–29), and our analysis showed that loss-of-function *ibm1* mutants contain over 109,000 hyper-CHG hcDMRs (*SI Appendix*, Fig. S16B). In contrast to gene bodies, heterochromatic regions in Arabidopsis are marked by the H2A variant H2A.W, which is encoded by the gene HTA6. *hta6* loss-of-function mutants result in TE derepression and elevated levels of CHG methylation (30). Consistent with the clear separation between these two distinct pathways, Q-MOD analysis showed that the overlap of hyper-CHG DMRs between *ibm1* and *hta6* was minimal (*SI Appendix*, Fig. S16B). Pollen samples, which include microspore, sperm cell, and vegetative nucleus, ranked very highly based on the total number of hyper-CHG DMRs (*SI Appendix*, Fig. S16B). These pollen hyper-CHG DMRs almost exclusively overlapped with the *hta6* hyper-CHG DMRs and not with the *ibm1*. In line with our observation, the expression level of *HTA6* in pollen is among the lowest compared with that from other types of Arabidopsis tissue (31) (*SI Appendix*, Fig. S17). This suggests that during pollen development, TEs rather than protein-coding genes tend to be hypermethylated in the CHG context, which is partially due to the down-regulation of *HTA6*. This is also consistent with the observation that pollen cells undergo epigenetic reprogramming to reinforce TE silencing via small RNA reactivation (32).

In addition, we observed extensive clustering of root tissue samples, indicating a distinct methylation profile for this tissue type, and confirmed extensive hypermethylation in the CHG and CHH contexts in columella cells (Dataset S12) as described (33). On the other hand, we did not detect any distinct clustering patterns for stress treated samples, perhaps suggesting that any methylation changes observed are not consistent between treatments.

We also observed a connection between the methylation profile of *mom1* and *morc* mutants. MOM1 is a plant-specific transcriptional silencer (34, 35). It acts synergistically with MORC6 to silence transcription at a subset of TEs by recruiting the RNA-binding protein HTA6, a RING finger/bilisulfite sequencing suggested that MOM1 has minimum effects on DNA methylation: TEs that are suppressed by MOM1 display no changes in DNA methylation in *mom1* mutants (15, 34, 36, 37). However, our large-scale BS-Seq analysis revealed that *mom1* indeed affects DNA methylation at a small number of loci (53 hypo-CHH, 271 hypo-CHG, and 721 hypo-CHH). These loci show significant overlap with those of *morc1* and *morc* mutants (*SI Appendix*, Figs. S10–S15; and Dataset S12). We refer to this method as Q-MOD. Using Q-MOD, we were able to detect some fine differences in the methylation profiles between the *morc* mutants and the controls (*SI Appendix*, Figs. S4–S9).

**Nonredundancy Within CHH Methylation Pathways and Within CHG Demethylation Pathways and a Connection Between MOM1 and MORC.** Previous studies have shown that the CMT2 and RdDM pathways act nonredundantly to control genome-wide CHH methylation in Arabidopsis (11, 12). This is consistent with the results from both our S-MOD and Q-MOD analyses, which revealed that the overlap between hyper-CHH DMRs observed in *cmt2* and RdDM mutants was minimal (*SI Appendix*, Fig. S16A). We also noticed interesting patterns of nonoverlapping at hyper-CHG DMR clustering. Ectopic gains of CHG methylation over gene bodies versus transposable elements (TEs) are prevented by distinct pathways, and our analysis confirmed this nonredundancy (*SI Appendix*, Fig. S16B).
levels of mCHG (Fig. 4C). Furthermore, the CG density at met1 sites is more than twice of that at cmt3 sites (6.5% vs. 2.7%), whereas conversely, the CHG density is greater at cmt3 sites than met1 sites (6.9% vs. 5.5% at met1 sites) (Fig. 4D). These higher densities are consistent with the well-established in vivo function of these enzymes because MET1 and CMT3 are primarily associated with maintenance of CG and CHG levels, respectively, throughout the genome.

However, these features do not explain why CHH methylation levels are reduced in met1 or cmt3 because CMT2 itself remains the most likely candidate for deposition of CHH at these loci (11, 12). We hypothesized that the loss of CHH methylation may be occurring indirectly, through loss of H3K9me2, which is needed for CMT2 function (11). Using a met1 H3K9me2 Chip–Chip data set (38), we found that TEs with met1-dependent CHH methylation indeed experienced a striking coincident reduction in H3K9me2 (Fig. 4E and F). At cmt3 sites, H3K9me2 levels are somewhat increased in the met1 background. This is consistent with the ectopic gains of H3K9me2 previously observed in the met1 background (38) and perhaps suggests that although global levels of H3K9me2 are maintained, there may be an antagonistic relationship between H3K9me2 levels at met3 vs. met1 subset sites. The coincident loss of CHH and H3K9me2 is consistent with a model whereby symmetric cytosine methylation at a subset of CMT2 target sites is required to maintain sufficient levels of H3K9me2 for CMT2 function. In support of this hypothesis, the met1/cmt3 double mutant shows a near-complete loss of CHH methylation (see ref. 15 and Fig. 4A). Although the interdependence of CHG methylation and H3K9me2 can be explained through the well-established feedback loop between CMT3 and KRYPTONITE/SUHV4 (39, 40), the connection between CG methylation and H3K9me2 at these sites is less obvious. Although the hypo-CHH DMRs in the suvh4 mutant reside clearly within the cmt3 subset block, the suvh4/cmt3 mutant resides in the cmt2 block along with the met1/cmt3 double mutant (Fig. 4A). This perhaps suggests that SUVH5 and/or SUVH6 may be responsible for linking CG methylation to H3K9me2 at these sites (11, 12). Using a met1 H3K9me2 Chip–Chip data set (38), we found that TEs with met1-dependent CHH methylation indeed experienced a striking coincident reduction in H3K9me2 (Fig. 4E and F). At cmt3 sites, H3K9me2 levels are somewhat increased in the met1 background. This is consistent with the ectopic gains of H3K9me2 previously observed in the met1 background (38) and perhaps suggests that although global levels of H3K9me2 are maintained, there may be an antagonistic relationship between H3K9me2 levels at met3 vs. met1 subset sites. The coincident loss of CHH and H3K9me2 is consistent with a model whereby symmetric cytosine methylation at a subset of CMT2 target sites is required to maintain sufficient levels of H3K9me2 for CMT2 function. In support of this hypothesis, the met1/cmt3 double mutant shows a near-complete loss of CHH methylation (see ref. 15 and Fig. 4A). Although the interdependence of CHG methylation and H3K9me2 can be explained through the well-established feedback loop between CMT3 and KRYPTONITE/SUHV4 (39, 40), the connection between CG methylation and H3K9me2 at these sites is less obvious. Although the hypo-CHH DMRs in the suvh4 mutant reside clearly within the cmt3 subset block, the suvh4/cmt3 mutant resides in the cmt2 block along with the met1/cmt3 double mutant (Fig. 4A). This perhaps suggests that SUVH5 and/or SUVH6 may be responsible for linking CG methylation to H3K9me2 at these sites (11, 12). Using a met1 H3K9me2 Chip–Chip data set (38), we found that TEs with met1-dependent CHH methylation indeed experienced a striking coincident reduction in H3K9me2 (Fig. 4E and F). At cmt3 sites, H3K9me2 levels are somewhat increased in the met1 background. This is consistent with the ectopic gains of H3K9me2 previously observed in the met1 background (38) and perhaps suggests that although global levels of H3K9me2 are maintained, there may be an antagonistic relationship between H3K9me2 levels at met3 vs. met1 subset sites. The coincident loss of CHH and H3K9me2 is consistent with a model whereby symmetric cytosine methylation at a subset of CMT2 target sites is required to maintain sufficient levels of H3K9me2 for CMT2 function. In support of this hypothesis, the met1/cmt3 double mutant shows a near-complete loss of CHH methylation (see ref. 15 and Fig. 4A). Although the interdependence of CHG methylation and H3K9me2 can be explained through the well-established feedback loop between CMT3 and KRYPTONITE/SUHV4 (39, 40), the connection between CG methylation and H3K9me2 at these sites is less obvious. Although the hypo-CHH DMRs in the suvh4 mutant reside clearly within the cmt3 subset block, the suvh4/cmt3 mutant resides in the cmt2 block along with the met1/cmt3 double mutant (Fig. 4A). This perhaps suggests that SUVH5 and/or SUVH6 may be responsible for linking CG methylation to H3K9me2 at these sites (11, 12). Using a met1 H3K9me2 Chip–Chip data set (38), we found that TEs with met1-dependent CHH methylation indeed experienced a striking coincident reduction in H3K9me2 (Fig. 4E and F). At cmt3 sites, H3K9me2 levels are somewhat increased in the met1 background. This is consistent with the ectopic gains of H3K9me2 previously observed in the met1 background (38) and perhaps suggests that although global levels of H3K9me2 are maintained, there may be an antagonistic relationship between H3K9me2 levels at met3 vs. met1 subset sites. The coincident loss of CHH and H3K9me2 is consistent with a model whereby symmetric cytosine methylation at a subset of CMT2 target sites is required to maintain sufficient levels of H3K9me2 for CMT2 function. In support of this hypothesis, the met1/cmt3 double mutant shows a near-complete loss of CHH methylation (see ref. 15 and Fig. 4A). Although the interdependence of CHG methylation and H3K9me2 can be explained through the well-established feedback loop between CMT3 and KRYPTONITE/SUHV4 (39, 40), the connection between CG methylation and H3K9me2 at these sites is less obvious. Although the hypo-CHH DMRs in the suvh4 mutant reside clearly within the cmt3 subset block, the suvh4/cmt3 mutant resides in the cmt2 block along with the met1/cmt3 double mutant (Fig. 4A). This perhaps suggests that SUVH5 and/or SUVH6 may be responsible for linking CG methylation to H3K9me2 at these sites (11, 12). Using a met1 H3K9me2 Chip–Chip data set (38), we found that TEs with met1-dependent CHH methylation indeed experienced a striking coincident reduction in H3K9me2 (Fig. 4E and F). At cmt3 sites, H3K9me2 levels are somewhat increased in the met1 background. This is consistent with the ectopic gains of H3K9me2 previously observed in the met1 background (38) and perhaps suggests that although global levels of H3K9me2 are maintained, there may be an antagonistic relationship between H3K9me2 levels at met3 vs. met1 subset sites. The coincident loss of CHH and H3K9me2 is consistent with a model whereby symmetric cytosine methylation at a subset of CMT2 target sites is required to maintain sufficient levels of H3K9me2 for CMT2 function. In support of this hypothesis, the met1/cmt3 double mutant shows a near-complete loss of CHH methylation (see ref. 15 and Fig. 4A). Although the interdependence of CHG methylation and H3K9me2 can be explained through the well-established feedback loop between CMT3 and KRYPTONITE/SUHV4 (39, 40), the connection between CG methylation and H3K9me2 at these sites is less obvious. Although the hypo-CHH DMRs in the suvh4 mutant reside clearly within the cmt3 subset block, the suvh4/cmt3 mutant resides in the cmt2 block along with the met1/cmt3 double mutant (Fig. 4A). This perhaps suggests that SUVH5 and/or SUVH6 may be responsible for linking CG methylation to H3K9me2 at these sites (11, 12). Using a met1 H3K9me2 Chip–Chip data set (38), we found that TEs with met1-dependent CHH methylation indeed experienced a striking coincident reduction in H3K9me2 (Fig. 4E and F). At cmt3 sites, H3K9me2 levels are somewhat increased in the met1 background. This is consistent with the ectopic gains of H3K9me2 previously observed in the met1 background (38) and perhaps suggests that although global levels of H3K9me2 are maintained, there may be an antagonistic relationship between H3K9me2 levels at met3 vs. met1 subset sites. The coincident loss of CHH and H3K9me2 is consistent with a model whereby symmetric cytosine methylation at a subset of CMT2 target sites is required to maintain sufficient levels of H3K9me2 for CMT2 function. In support of this hypothesis, the met1/cmt3 double mutant shows a near-complete loss of CHH methylation (see ref. 15 and Fig. 4A). Although the interdependence of CHG methylation and H3K9me2 can be explained through the well-established feedback loop between CMT3 and KRYPTONITE/SUHV4 (39, 40), the connection between CG methylation and H3K9me2 at these sites is less obvious.
Materials and Methods

Bioinformatic Analysis.

BS-seq analysis and DMR calling. Raw sequencing data (SRA files) of 503 Arabidopsis WGBS libraries were downloaded from GEO (Dataset S1). RNA-seq and BS-seq data from MORC6 transgenic materials are deposited to GEO under accession GSE98872 (reviewer access token: odwdowmwtvitpwv).

BS-seq reads from each library were mapped to Arabidopsis TAIR10 genome using BSMAP (42), allowing only uniquely mapped reads, with up to 4% mismatches. We also applied a CHH filter, which discarded reads with three or more consecutive CHH sites to remove reads with low conversion efficiency (13, 15). Fractional DNA methylation levels were computed by CG (\((# + \#T)\) using 100-bp bin. Libraries with a high proportion of unconverted-C in chloroplast (>1%) and low genome coverage (<5%) were discarded. Libraries that are not in Col-0 background or have duplicated GEO entries were also filtered out from further analysis.

The remaining 314 high-quality WGBS libraries, including 54 designated wild-type control (Col-0 leaf or seedling) and 260 test libraries, were selected for DMR analysis. Standard DMR calling between test and control libraries was performed as previously described (15), where the difference in CG, CHG, and CHH methylation in each bin needed to be at least 0.4, 0.2, and 0.1, respectively (15). DMRs were defined as 100-bp different methylated genomic fragments.

To identify the number of shared control sets required for robust DMR designation, we assessed DMR calling in each control library (in addition to test type libraries) against the other 53 control libraries (Dataset S3). For each 100-bp bin, setting the cutoff for DMR designation to one control library results in a large number of DMRs being called from both the test and the control libraries (SI Appendix, Fig. S1), many of which are spontaneous or false positive DMRs. As the cutoff increases, the number of DMRs in each of the control libraries drops significantly, whereas the number of DMRs from the test libraries remains substantial. However, requiring a cutoff of 54—meaning that the locus in the test library must be different from every other control library—appeared too stringent because most libraries only retained a small number of DMRs using this criterion (Dataset S3). To find the balance between stringency and false positive DMR designation, we calculated the deceleration of number of DMRs identified in control libraries as the increase of the number of supporting control libraries using the equation $a = \left(\frac{n_1 - n_{1-1}}{n_1} - \frac{n_{1-2} - n_{1-3}}{n_1}\right)$, where $n_i$ is the average of number of DMRs identified from 54 control libraries, which are supported by at least $c$ control libraries ($c \geq 3$). In all six contexts (hypo-/hyper-CG/CHG/CHH), the deceleration rate of the amount of DMRs becoming called becomes steady at around 33 control sets (SI Appendix, Fig. S2). Thus, a DMR is only defined when the test library bin differs from at least 33 out of the 54 control libraries. Note that it is not necessarily the same 33 libraries of 54 that are taken as supported controls for each bin. Although these stringent criteria afford a reduced the rate of false positives, this comes at a trade-off with false negatives (Fig. 2 A and B). We recorded both the number of “Goodbin” (i.e., 100-bp bin with sufficient coverage, $\geq 5$) and the number of DMRs for all libraries in CG, CHG, and CHH contexts. For the matched WT hypo-CHH DMRs (Fig. 2), we used the same 100-bp bins and 0.1 methyl ratio change cutoff, then applied Fisher’s exact test requiring mutant backgrounds. The TE requiring met1 for CHH methylation maintenance also shows loss of H3K9me2 in the met1 background. (B) H3K9me2 levels in WT vs. met1 over met1 subset hypo-CHH hCDMRs (Left) and over cmt3 subset hypo-CHH hCDMRs (Right), ***p < 2.2e-16 (Wilcoxon rank sum test).

Comparative analysis of DMRs across libraries. Given two test libraries, we compare them by testing the dependence of DMR sets $A$ and $B$ to determine if they are independent. An observed number of overlapping DMRs $n$ between $A$ and $B$ is independent. We use the number of overlapping DMRs $n$ between $A_1$ and $A_2$ as test statistic with performance of hypergeometric test. The null hypothesis will be rejected when the test statistics are high. The statistical significance for the overlap of DMRs between two test libraries was calculated using the equation $a = \left(\frac{n_1 - n_{1-1}}{n_1} - \frac{n_{1-2} - n_{1-3}}{n_1}\right)$, where $n_i$ is the average of number of DMRs identified from 54 control libraries, which are supported by at least $c$ control libraries ($c \geq 3$). In all six contexts (hypo-/hyper-CG/CHG/CHH), the deceleration rate of the amount of DMRs becoming called becomes steady at around 33 control sets (SI Appendix, Fig. S2). Thus, a DMR is only defined when the test library bin differs from at least 33 out of the 54 control libraries. Note that it is not necessarily the same 33 libraries of 54 that are taken as supported controls for each bin. Although these stringent criteria afford a reduced the rate of false positives, this comes at a trade-off with false negatives (Fig. 2 A and B). We recorded both the number of “Goodbin” (i.e., 100-bp bin with sufficient coverage, $\geq 5$) and the number of DMRs for all libraries in CG, CHG, and CHH contexts. For the matched WT hypo-CHH DMRs (Fig. 2), we used the same 100-bp bins and 0.1 methyl ratio change cutoff, then applied Fisher’s exact test requiring mutant backgrounds. The TE requiring met1 for CHH methylation maintenance also shows loss of H3K9me2 in the met1 background. (B) H3K9me2 levels in WT vs. met1 over met1 subset hypo-CHH hCDMRs (Left) and over cmt3 subset hypo-CHH hCDMRs (Right), ***p < 2.2e-16 (Wilcoxon rank sum test).

**PLANT BIOLOGY**

**PLOS PLUS**

**PNAS PLUS**

**PNAS EARLY EDITION**

Zhang et al.

**Zhang et al.**
\( p_{\text{value}} = \frac{\sum_{i=1}^{n_1} \binom{n_1}{i} \binom{n_2}{N-i}}{\binom{N}{N_1}} \), in which \( n_1 \) and \( n_2 \) are the number of DMRs in the two libraries being compared, \( N \) is the number of common DMRs shared by the two libraries, and \( N \) is defined as the size of the union of all possible DMRs identified in the whole dataset of test libraries. Hence, the \( p_{\text{value}} \) indicates the level of their dependence between the two libraries.

The S-MOD score was then transformed into an S-MOD score using the equation S-MOD score = −\( l\log(p_{\text{value}}) \).

Pairwise S-MOD scores between all test libraries were stored in a symmetric matrix, of which the horizontal and vertical coordinate are library IDs. Before Q-MOD clustering, we filtered out libraries in which the S-MOD score was below 100 in all pairwise comparisons, indicating that the library has low relatedness to all other libraries (see Supplementary Table S1). We performed separately for hyper-pyr-GpCG/ChG/ChH.

We chose 100 as the S-MOD cutoff score because it can filter out most libraries with no/weak relatedness and libraries with low number (\( <40 \)) of DMRs (SI Appendix, Fig. S18).

Also, to offset the effect of library quality on number of hcDMRs (libraries with lower sequencing depth typically yield fewer hcDMRs), the total number of hcDMRs in the denominator was adjusted to remove hcDMRs that do not have sufficient coverage in both mutant libraries (SI Appendix, Fig. S19). Q-MOD overlap percentages were calculated using the equation overlap\%(\( i \)) = \( \frac{\text{card}(A \cap n1) / \text{card}(A) \text{, where A and } A_i \text{ represent the set of DMRs of the two test libraries being compared. The total number of hcDMRs in the denominator was adjusted to remove hcDMRs that do not have sufficient coverage in both test libraries (Goodbin filter). The matrix of overlap percentage was clustered using the Ward method in R.

Experimental Procedures.

**Transgenic plant material and constructs.** Wild-type and morc6-3 mutant lines (gene AT1G19100, GABI_599B06, aka crh6-5) are from the ecotype Columbia (Col-0) and were grown under 16-h light, 8-h dark cycles. The pATMORC6::ATMORC6-FLAG construct was previously described (43). Agrobacterium tumefaciens AGLO strain carrying this construct was used to transform morc6-3 using the floral dip method. For Western blot, HRP-coupled FLAG-secondary antibody (A8592; Sigma) was used.

**RNA-seq.** RNA was extracted from unopened floral bud tissue using TRIzol (Thermo Fisher) and the Direct-Zol RNA MiniPrep kit (R2050; Zymo Research) including in-column DNase treatment. Seventy-five nanograms total RNA was used as input for the TrueSeq Stranded mRNA Library Prep Kit for Neoprep (NP-202-1001; Illumina). Libraries were sequenced on a HiSeq 2000 (Illumina). Reads were aligned with TopHat, including the fr-firststrand parameter. Cufflinks was used to generate count data using annotation from TAIR10 that was fetched to the DESeq2 package in R for differential expression analysis.

**BS-seq.** Genomic DNA from leaf tissue from a T1 (transgenic generation one) pATMORC6::ATMORC6-FLAG in morc6-3 plant was isolated using DNeasy Plant Mini kit (69106; Qiagen). Five hundred nanograms genomic DNA starting material was sheared using the Covaris instrument. Libraries were generated using the KAPA hyper prep kit (KK802) with EZ DNA methylation lightening kit for bisulfite conversion (D5030; Zymo Research) and MyTaQ HS mix (BioLine BIO-25045) for amplification. Libraries were sequenced on a HiSeq 2000 (Illumina), and reads were aligned to the TAIR10 genome using BSMAP.

**Acknowledgments.** The authors thank the UCLA-FAFU Joint Research Center on Plant Proteomics for institutional support and Life Sciences Editors for assistance with manuscript editing. Y.Z. was supported by National Natural Science Foundation of China Grant 31701092 and Southern University of Science and Technology Start-up Postdoctoral Fellowship. C.J.H. was supported by an European Molecular Biology Organization Long-Term Fellowship (ALTF 1138-2014). H.W. was supported by National Natural Science Foundation of China Grant 31570102. S.E.J. is an investigator of the Howard Hughes Medical Institute. The group of J.Z. is supported by the Thousand Talents Program for Young Scholars and by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT065172). Work in the S.E.J. laboratory was supported by NIH Grant GM0398.


Zhang et al.