Title: The roles of Conserved Domains in DEMETER-Mediated Active DNA Demethylation *in planta*

Changqing Zhang[1,2,*], Yu-Hung Hung[1,2,*], Xiang-Qian Zhang[1,2,3,*], Dapeng Zhang[4,5], Wenyan Xiao[4], Lakshminarayan M. Iyer[6], L. Aravind[6], Jin Hoe Huh[7] and Tzung-Fu Hsieh[1,2,*]

Affiliations:

[1]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC 27695, USA

[2]Plants for Human Health Institute, North Carolina State University, North Carolina Research Campus, Kannapolis, NC 28081, USA

[3]Guangdong Engineering Research Center of Grassland Science, College of Forestry and Landscape Architecture, South China Agricultural University, Guangzhou 510642, China.

[4]Department of Biology, St. Louis University, St. Louis, MO 63103, USA

[5]Program of Bioinformatics and Computational Biology, St. Louis University, St. Louis, MO 63103, USA

[6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

[7]Department of Plant Science, Plant Genomics and Breeding Institute, and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul 08826, Republic of Korea

[*]These authors contribute equally to this study.

Correspondence and requests for materials should be addressed to T.-F.H. thsieh3@ncsu.edu

Abstract

DNA methylation plays critical roles in maintaining genome stability, genomic imprinting, transposon silencing, and development. In Arabidopsis genomic imprinting is established in the central cell by DEMETER (DME)-mediated active DNA demethylation, and is essential for seed viability. DME is a large polypeptide with multiple poorly characterized conserved domains. Here we show that the C-terminal enzymatic core of DME is sufficient to complement *dme* associated developmental defects. When targeted by a native DME promoter, nuclear-localized DME C-terminal region rescues *dme* seed abortion and pollen germination defects, and ameliorates CG hypermethylation phenotype in *dme-2* endosperm. Furthermore, targeted expression of the DME N-terminal region in wild-type central cell induces *dme*-like seed abortion phenotype. Our results support a bipartite organization for DME protein, and suggest that the N-terminal region might have regulatory function such as assisting in DNA binding and enhancing the processivity of active DNA demethylation in heterochromatin targets.

Double fertilization during sexual reproduction in flowering-plants is a unique process that underlies the distinctive epigenetic reprogramming of plant gene imprinting. In the ovule, a haploid megaspore undergoes three rounds of mitoses to produce a 7-celled, 8 nuclei embryo sac that consists of egg, central, and accessory cells [1]. During fertilization pollen grain elongates and delivers two sperm nuclei to the female gametophyte to fertilize the egg cell and the central cell, respectively. The fertilized egg cell forms the embryo that marks the beginning of the subsequent generation. Fertilization of the central cell initiates the development of endosperm that accumulates starch, lipids, and storage proteins and serves as a nutrient reservoir for the developing embryo [2, 3]. Endosperm is the major tissue where gene imprinting takes place in plant. Genomic imprinting is the differential expression of the two parental alleles of a gene depending on their parent-of-origin, and is an example of inheritance of differential epigenetic states. In Arabidopsis, MET1-mediated DNA methylation and DME demethylation are two modes of epigenetic regulation critical for imprinted expression of many genes [4, 5, 6, 7, 8]. For example, DEMETER (DME) is required for the expression of *MEA*, *FIS2*, and *FWA* in the central cell and in the endosperm while MET1 is responsible for the silencing of *FIS2* and *FWA* paternal alleles [4, 7]. Gene imprinting is essential for reproduction in Arabidopsis, and seeds that inherit a maternal *dme* allele abort due to failure to activate *MEA* and *FIS2*, essential components of the endosperm PRC2 complex required for seed viability, in the central cell [4, 9].

DME encodes a bifunctional 5mC DNA glycosylase/lyase required for active DNA demethylation in the central cell and the establishment of endosperm gene imprinting in Arabidopsis [5]. Additionally, paralogs of DME, REPRESSOR OF SILENCING 1 (ROS1), DML2, and DML3 are required to counteract the spread of DNA methylation mediated by the RNA-directed DNA methylation (RdDM) machinery into nearby coding genes [10, 11]. The three regions in the C-terminal half of DME protein (the A, Glycosylase, and the B regions, or as the AGB region hereafter) are conserved among the DME/ROS1 DNA glycosylase clade, and are required for DME 5mC excision activity *in vitro*. Thus, the AGB region comprise the minimal catalytic core for the enzymatic function, catalyzing direct excision of 5mC from DNA and initiating active DNA demethylation that influences transcription of nearby genes [5, 9, 12].

In Arabidopsis, DME-mediated DNA demethylation is preferentially targeted to small, AT-rich, and nucleosome-poor euchromatic transposons that flank coding genes [13]. Consequently, demethylation in the central cell influences expression of adjacent genes only in the maternal genome, and is a primary mechanism of gene imprinting in plant [5, 13, 14, 15]. In addition to small TEs near coding sequences, DME also targets gene-poor heterochromatin regions for demethylation [13]. The mechanism of DME recruitment to its target sites is not known. Studies in ROS1 have uncovered several players required in the ROS1 demethylation pathway [16, 17, 18]. Among them *IDM1* encodes a novel histone acetylase that preferentially acetylates H3K18 and H3K23 *in vitro*, and ROS1 target loci are enriched for H3K18 and K23 acetylation *in vivo* in an IDM1-dependent manner [19]. Thus, IDM1 marks ROS1 target sites by acetylating histone H3 to create a permissible chromatin environment for ROS1 function. The Arabidopsis SSRP1 (STRUCTURE SPECIFIC RECOGNITION PROTEIN1), a component of the FACT (facilitates chromatin transcription/transaction) histone chaperone complex, has been shown to regulate DNA demethylation and gene imprinting in Arabidopsis [20]. Linker histone H1 functions in chromatin folding and gene regulation [21, 22, 23, 24], and was shown to interact with DME in a yeast two-hybrid screen and in an *in vitro* pull-down assay [25]. Loss-of-function mutations in *H1* genes affect the imprinted expression of *MEA* and *FWA* in Arabidopsis endosperm, and impair demethylation of their maternal alleles, suggesting that H1 might participate in the DME demethylation process by interaction with DME [25].

Computational analysis showed that the DME/ROS1 like DNA glycosylases contain a core with multiple conserved globular domains, and except for the well-characterized glycosylase domain, very little is known about the function of the other domains. Here we show that the C-terminal region of DME necessary for 5-methylcytosine excision activity *in vitro* is sufficient to complement *dme* seed abortion and pollen germination defect, and partially rescue DNA hypermethylation phenotype in endosperm. We present evidence that the region N-terminal to the glycosylase domain can affect endogenous DME activity in a dominant negative manner when ectopically expressed in the nuclei of wild-type central cells. We propose a bipartite structural and functional organization model for the DME/ROS1 family of DNA glycosylases consisting the modular C-terminal AGB region that can substitute for DME's developmental function and the NTD region that might have regulatory functions such as assisting DNA binding and enhancing the processivity of demethylation in heavily methylated genomic regions.

Results

The DME catalytic core region is sufficient to complement *dme* associated developmental defects. Previous studies have revealed that the C-terminal half of DME comprising the three conserved A, Glycosylase, and B regions (the AGB region, as shown in Supplementary Fig. 1a) are required for *in vitro* 5mC excision activity [5], and deletion of the non-conserved linker between domain A and the glycosylase domain (interdomain 1; ID1) does not affect DME *in vitro* enzymatic activity [26, 27]. Thus, the AGB region is thought to be the core catalytic region for DME *in vitro* enzymatic activity. However, it is unknown whether the AGB region alone is sufficient for DME function *in vivo*. To determine if the AGB region is functional *in vivo*, we tested if expressing the AGB region in the central cell can complement *dme* seed abortion phenotype. A transgene carrying a 3.1-kb *DME* cDNA that encodes the C-terminal half of DME ($\mathrm{DME}^{\mathrm{CTD}}$, residue 936-1987) under the control of a native DME promoter was introduced into *DME/dme-2* heterozygous plants by using the floral dipping method [28]. Since $\mathrm{DME}^{\mathrm{CTD}}$ lacks a nuclear localization signal (data not shown), a classical SV40 nuclear localization signal (PKKKPKV) was introduced in front of the C-terminal fragment (designated as $nDME^{CTD}$, see Supplementary Fig. 1b) to ensure proper nuclear localization. We obtained multiple independent transgenic lines and assessed the transgene's ability to complement *dme-2* seed abortion phenotype.

The self-pollinated *DME/dme-2* plants produce 50% of normal seed that inherited wild type DME maternal allele, and the other 50% of aborted seed that inherited mutant *dme-2* maternal allele. In self-pollinated transgenic plants that carry a single locus of $nDME^{CTD}$ or DME^{FL} (full length DME.2 cDNA, major isoform of DME [29]) transgenes, we observed about 25% aborted seeds among independent transgenic lines, indicating that $nDME^{CTD}$ and DME^{FL} complement *dme* seed abortion phenotype (Fig. 1a, b, Supplementary Table 1). In addition, we also transformed $nDME^{CTD}$ and DME^{FL} into *dme-2/dme-2* homozygous plants (see Materials and Methods for isolation and characterization of *dme-2/dme-2* homozygous lines in Col-*gl*), both constructs produced T1 transgenic plants that displayed expected 50% seed abortion rate (Fig. 1b, Supplementary Table 1). Seed abortion caused by *dme* mutations is in part due to defects in activating imprinted PRC2 subunit genes required for endosperm development [5, 9, 30, 31, 32]. We use qRT-PCR to check if $\mathrm{nDME}^{\mathrm{CTD}}$ also restores DME target genes expression in the central cell. Indeed, *FIS2* and *FWA* expression is restored in the complemented lines (Fig. 1c). Thus nDME^{CTD} can substitute for the endogenous DME activity for seed viability, and active DME target genes expression.

In addition to maternal effects on seed viability [9], mutations in DME also affect pollen function in Col-0. When *DME/dme-2* heterozygous plants are self-pollinated, only about 20-30% of the viable F1 progeny are heterozygous (Supplementary Table 2), due to decreased *dme* pollen germination rate [33]. To test whether nDME^{CTD} can rescue *dme* pollen phenotype, we pollinated wild type Col-0 with pollen derived from transgenic lines that are homozygous for the *dme-2* allele and carry a single locus of the *nDME^{CTD}* transgene (*dme-2/dme-2; nDME^{CTD}*/~). If nDME^{CTD} does not complement *dme-2* pollen germination defects, we expect roughly half of the F1 progeny will carry the nDME^{CTD} transgene (hygromycin resistant) because mutant pollen with or without the transgene would germinate with equal frequency. Instead, we observed 65% - 90% of the F1 progeny are hygromycin resistant (Table 1), indicating that nDME^{CTD} complements *dme-2* pollen germination defect. These results show that expressing the C-terminal half of DME protein in the nucleus is sufficient to rescue *dme* visible phenotypes *in planta*.

nDME^{CTD} partially rescue *dme-2* CG hypermethylation phenotype in the endosperm. In Arabidopsis seed viability depends on the DME activity in the central cell to activate the MEDEA/FIS2/MSI1/FIE PRC2 complex required for endosperm development. In addition, DME is required to demethylate multiple maternally (*MEGs*) or paternally expressed imprinted genes (*PEGs*) to establish their parent-of-origin specific expression patterns in the endosperm [13, 15]. Thus, in *dme* mutant endosperm, discrete genomic loci targeted by DME for demethylation are hypermethylated [13]. Since nDME^{CTD} complements *dme* seed abortion, and activates DME target gene expression (Fig. 1), we assumed it does so by demethylating the central cell genome and activating PRC2 genes essential for seed development. To test this hypothesis, and to examine the extend of nDME^{CTD} demethylation activity *in vivo*, we manually isolated *nDME^{CTD}*-complemented endosperm (*dme-2/dme-2;nDME^{CTD}/nDME^{CTD}*), determined the DNA methylation profile by whole genome bisulfite sequencing, and compared the complemented methylomes to those of wild-type and *dme-2* endosperm. Methylomes from three independent lines were generated and compared with that of *dme-2* endosperm. We observed although the differentially methylated regions (DMRs) between each independent lines do not completely overlap, the DMRs unique to each line are also demethylated in other lines (Supplementary Fig. 2, 3), suggesting that the number of overlapped DMRs was underestimated due to the cutoff used in defining the DMRs, similar to what's observed in a recent study [34]. We therefore used the combined reads from three independent lines for the subsequent analyses so that all comparisons are confined to the same cutoff criteria (see Materials and Methods). As expected, several DME regulated *MEGs* and *PEGs* are demethylated compared to *dme-2* endosperm, indicating that nDME^{CTD} is correctly recruited to these loci for demethylation (Fig. 2a). We focused our analysis on previously determined differentially methylated sites between *dme-2* and wild-type endosperm (*dme* hyper-DMRs, the DME canonical targets) [13, 15]. Overall, the CG methylation levels in these canonical DME target sites are reduced in the complemented endosperm, indicating that nDME^{CTD} is directed to these endogenous DME target sites for demethylation. However, compared to wt endosperm, these *dme* hyper-DMRs are demethylated to a lesser degree by the nDME^{CTD} (Fig. 2b). Thus nDME^{CTD} only partially rescues the *dme* CG hypermethylation phenotype in the endosperm. The DMRs of *dme* relative to wild-type endosperm or to *nDME^{CTD}*-complemented endosperm partially overlap (Supplemental Fig. 4).

However, among the DMRs unique to nDME^{CTD}, we also observed decreased CG methylation in WT endosperm compared to *dme*, indicating that they are also demethylated by the endogenous DME. Similarly, among the DMRs unique to wt endosperm, these regions are also demethylated by the nDME^{CTD}. Thus nDME^{CTD} appears to partially demethylate the majority of the loci targeted by the endogenous DME. These observations also suggest that intact full-length DME protein is required for robust and complete demethylation *in vivo*.

We next examined the methylome of *dme-2* endosperm complemented by the full length *DME.2 cDNA* (designated as *DME^{FL}*). Unexpectedly, based on the number of DMRs between *dme* and *DME^{FL}*-complemented endosperm and the level of CG methylation within the DMRs (Fig. 2c), DME^{FL} appears to be less active compared to endogenous DME, or to nDME^{CTD}, albeit it being able to complement *dme* seed abortion (Fig. 1b) [9, 35]. Since the *DME^{FL}* transgene only differs from *nDME^{CTD}* by the N-terminal region, reduced activity of DME^{FL} compared to DME^{CTD} cannot be attributed to the lack of introns or 3’ flanking sequences that might be needed for robust DME protein production. Indeed, we found both transgenes are expressed at comparable levels in *DME^{FL}*- and *nDME^{CTD}*-complemented lines used in the methylome study (Supplemental Fig. 5), indicating lower activity of DME^{FL} compared to nDME^{CTD} is not due to their differential transcript abundance. Nevertheless, comparison of CG methylation levels in DMR regions unique to DME^{FL}, nDME^{CTD}, or endogenous DME also reveals that unique DMR regions are more or less hypomethylated in WT or in complemented endosperm relative to *dme* endosperm. Thus the methylome difference between wt, *DME^{FL}*-, and *nDME^{CTD}*-complemented endosperm appears to be more in the degree of demethylation, rather than in targeting specificity.

Function of the N-terminal region in DME-mediated active DNA demethylation. The *dme-2* allele is caused by an activation-tagging T-DNA insertion in the middle of the A region (Supplementary Fig. 1a)[9]. We found that in floral buds of *dme-2/dme-2* plants, the endogenous *DME* transcripts downstream of T-DNA insertion site is greatly reduced compared to wild-type Col-0 plants, but the level of DME transcripts upstream of the T-DNA insertion site is relatively high (Supplementary Fig. 6). We suspected these transcripts could produce truncated form of DME proteins that might interfere with the $\mathrm{DME}^{\mathrm{FL}}$ transgene activity. To test this hypothesis, we transformed wild-type Col-0 plants with an engineered GFP-tagged DME NTD (using the genomic DNA sequence upstream of T-DNA insertion site, encoding residues 1-1022, designated as DME^{NTD}*-GFP*) transgene mimicking the *dme-2* T-DNA insertion (Supplementary Fig. 1B). Clear GFP signals are observed in the central cell nuclei of transgenic lines (data not shown). We also observed about one third of transgenic lines showing apparent *dme-2* like seed abortion phenotype, with abortion rates ranging from 10% to ~ 40% (Supplementary Table 3, 4) in the T1 plants, suggesting that expression of $\mathrm{DME}^{\mathrm{NTD}}$ has a dominant negative effect on endogenous DME protein.

To minimize the possibility and the degree of transgene induced sense co-suppression, we reverse translated $\mathrm{DME}^{\mathrm{NTD}}$ protein sequence into cDNA sequence using the human codon usage table. As a result, the re-engineered “humanized” version of NTD ($\mathrm{mDME}^{\mathrm{NTD}}$) codes for the identical protein sequence but with no significant nucleotide sequence similarity to the original cDNA sequence to induce co-suppression (Supplementary Table 5). In addition, a GFP tag was added to the C-terminus ($\mathrm{mDME}^{\mathrm{NTD}}$-GFP) to monitor its expression (Fig. 3a). We generated 28 independent transgenic lines, and among them 16 lines showed seed abortion rate of 5% - 52% (Supplementary Table 3, 6). The aborted seeds resemble *dme* mutant seeds with abnormal endosperm, arrested embryo, and shriveled brown seeds (Fig. 3b, c). We selected four lines with high, medium, or no seed abortion rate (Fig. 3d), and assessed the endogenous DME transcript abundance. As shown in Fig. 3e, among lines with different seed abortion rate, the endogenous DME mRNA abundance is similar to that of the vector control line, indicating the severity of seed abortion phenotype is not due to interference of endogenous *DME* transcripts. Furthermore, the rate of seed abortion is positively correlated with the levels of *mDME*^{NTD}*-GFP* mRNA (Fig. 3f), suggesting the degree of seed abortion is likely due to the levels of transgene expression. We next tested whether expression of *nDME*^{CTD} or *DME*^{FL} in WT Col-0 can also induce seed abortion phenotype. For each construct, more than 25 independent transgenic lines were examined and none resulted in any seed abortion phenotype (Supplementary Table 3). Thus the dominant negative effect appears to be specific to the DME NTD region.

Evolutionary history and late acquisition of the N-terminal region of DME-like proteins.

We show the C-terminal half of DME is sufficient to complement *dme* mutant developmental phenotypes, and can be recruited to most of the DME target loci. Thus the DME$^{\mathrm{CTD}}$ most likely contains intrinsic targeting information. To gain insights from the evolution of the conserved domains in DME, we conducted sequence searches of the NR database with various homologs as query. The core of the DME-like proteins, as previously reported [36], comprises the catalytic glycosylase domain of the HhH (helix-hairpin-helix) modules followed by the FCL ([Fe4S4] cluster loop) motif and a divergent version of an RRM (RNA Recognition Motif) fold domain (Fig. 4). The DNA glycoslase and FCL domains span the A and G regions, whereas the RRM fold domain corresponds to the B region of angiosperm DME homologs. A diversity of domains associate with the basic DME core can be found across various clades. Land plants and charophytes (Streptophyta) possess a permuted divergent version of the umethylated CpG recognizing CXXC domain (containing only one of two structural repeats of the classical CXXC domain) between the FCL and RRM domains. By contrast, one or more copies of the CXXC domain can be found in chlorophyte and stramenopile algae at distinct positions. Some algal DME homologs (from Chlorophyte and stramenopile) also possess other chromatin-modification reader (Tudor and PHD domains), DNA binding (AT-hook motif), and the DnaJ domain which interacts with the chaperone Hsp70 [36, 37]. These accessory domains suggest a potential role for regulating the associated DNA glycosylase activity according to the DNA methylation (via CXXC) or chromatin status (via PHD, Tudor) of the cell in which they are expressed.

The N-terminal half of the DME consists of a large portion of unstructured, low complexity sequences (residues 346-947), a stretch of basic amino acid-rich direct repeats (residues 291-345), and a 120 amino-acid N-terminal domain (DemeN) of unknown function (residues 1-120)(see Supplementary Fig. 7 for sequence alignment). The DemeN domain and charged repeats are restricted to the angiosperm lineage and appears to be a late acquisition during land plant evolution.

In summary, the evolutionary history of the DME domains can be summarized as follows: bacterial versions of the HhH-FCL pair from a cyanobacterial source fused to an RRM-fold domain and further acquired an insert in the glycosylase domain to give the ancestral form in the plant lineage. This was likely then transferred to the stramenopiles from a secondary chlorophyte endosymbiont of this lineage. Finally, at the base of the streptophyte radiation, DME acquired a permuted CXXC, and later the DemeN domain and the associated charged repeats were acquired in the angiosperm lineage, possibly to facilitate and ensure a robust and thorough demethylation.

Discussion

We show for the first time that the core conserved region of the DME protein containing the DNA-glycosylase, FCL, divergent and permuted CXXC and divergent RRM domains is sufficient to rescue visible phenotypic defects caused by *dme* mutation. Although this truncated form of DME protein demethylates the majority of the canonical DME target sites, it does so in a less active and less efficient manner compared to the endogenous protein. We see two possibilities that might explain this lower activity and efficiency: 1) Critical cis-elements residing within introns or in 3'-end flanking sequences that are missing in the transgene might be required for robust transgene expression. 2) The N-terminal region might be required for full DME activity *in vivo*. Unfortunately, our attempt to assess the difference between DME^{FL} and $nDME^{CTD}$ was confounded by the possible interference from truncated NTD proteins due to T-DNA insertion in *dme-2* background. We suspect this might contribute to the reduced DMRs observed in $\mathit{DME}^{\mathit{FL}}$complemented endosperm. Therefore, we believe it is premature to draw any conclusion based on direct comparisons between $\mathit{DME}^{\mathit{FL}}$- and $\mathit{nDME}^{\mathit{CTD}}$-complemented endosperm methylomes (Fig. 2c).

Since the C-terminal AGB region is sufficient for DME's seed viability function in *planta*, and can be recruited to most of the canonical DME target sites, the CTD polypeptide most likely contains sufficient targeting information. *in vitro* studies of ROS1 suggest that the B region containing the CXXC and RRM domains is essential for the glycosylase and lyase activities, and might recognize modified DNA[38]. It is possible that the permuted CXXC domain is required to direct the protein to the target sites, or is involved in discriminating methylated vs un-methylated cytosines[39]. This is supported by mutation studies that implicate a potential role for this domain in DME *in vivo* function, but not in vitro enzymatic activity (Huh and Hsieh, unpublished results). Similarly, the role of the enigmatic divergent RNA-recognition motif (RRM) domain is also not fully understood. Mutagenesis screens for residues required for demethylation activity in bacteria identified multiple amino acid residues within the RRM domain[40]. Although the involvement of RNA species in the active DNA demethylation process has not been firmly established, an RRM protein ROS3 required for ROS1 demethylation suggests a potential role of non-coding RNAs in the active DNA demethylation pathway in Arabidopsis[41]. While it is tempting to speculate a role for RNA-binding, the DME RRM might also bind single-stranded DNA with methylated bases.

Based on the reduced demethylation activity of nDME^{CTD} on the canonical DME target sites, we suspect the NTD region might be required for full and robust demethylation activity probably to ensure that the imprinting network is properly activated and maintained (e.g., by subsequent PRC2 activity). To achieve this, the DME NTD might function to assist the glycosylase enzyme by tightly binding to DNA template for more complete and thorough demethylation. Supporting such model, *in vitro* study of ROS1 activity on 5mC excision revealed that the basic repeats (3DR, AT-hooks) region binds strongly to DNA template non-specifically, and removal of NTD region impairs the sliding capacity of the protein on DNA template[42], and significantly reduced ROS1 5mC excision activity[43]. We observed reduced degree of demethylation by nDME^{CTD} regardless of target length (Supplemental Fig. 8), suggesting that NTD is needed for complete demethylation in all the target sites.

Although DME preferentially targets smaller euchromatic transposons that flank coding genes, it also targets gene-poor heterochromatin regions for demethylation [13]. The biological significance of heterochromatin demethylation by DME is not known, but was speculated to involve reinforcing DNA methylation in egg cell and subsequently in the embryo [13]. These heterochromatin target sites are densely methylated, and demethylation by DME results in longer DMRs between *dme-2* and wt endosperm. Interestingly, the number of longer DMRs is significantly reduced between *dme-2* and *nDME*^{CTD}-complemented endosperm, suggesting that removal of NTD region also reduces the processivity of demethylation in long target sites (Supplemental Fig. 9a). Since heterochromatin regions are compacted, demethylation in such loci will require substantial chromatin remodeling such as eviction of nucleosomes for DME to gain access to the templates. It is tempting to speculate that the conserved motif in the DemeN domain might recruit other factor(s) via protein interaction to remodel local chromatins to permit DME demethylation. However, based on current data we cannot unequivocally ascribe NTD's function due to lack of proper full length DME transgenic comparison. Nevertheless, our results caution that peculiarity in certain genetic backgrounds (e.g., *dme-2*) might confound data interpretation. Future work on DME functional study could benefit from the generation of a clean loss-of-function background such as deleting the entire DME locus using CRISPR-assisted genome editing techniques.

We envision a possible model where the AGB region is sufficient for directing DME to target loci while NTD region is required for interacting with local chromatin environment, stabilizing binding to chromosomal templates, and assisting demethylating flanking sequences. In the absence of NTD, nDME^{CTD} can still demethylate majority of target sites, but in a less-efficient manner, likely due to the lack of non-specific DNA-binding by the basic AT-hook motifs. We surveyed wt DMRs that are longer than 1.5 kb, and found that these regions are also nDME^{CTD}'s DMRs, but are shorter in length (Supplemental Fig. 9b), possibly due to missing the DemeN domain. If NTD is needed for longer and more robust demethylation, why ectopic expression of NTD causes dominant negative (DN) effects on endogenous protein? Classical examples of dominant negative mutation often involve protein-protein interactions that are disrupted by mutated or truncated form of one particular partner or subunit. Although we do not have any evidence to suggest DME might homodimerize to become active, any weak physical interaction caused by ectopic NTD expression might induce conformational change that renders DME non-functional. Unfortunately our attempt to assess whether the NTD of DME can interact with each other was confounded by the self-activating activity of DME.2 NTD in yeast two-hybrid assay when fused to the GAL4 DNA binding domain (data not shown). Their possible interaction will need to be assessed by alternative strategies. Another possibility is that NTD binds and titrates out an important interacting partner required to activate DME through conformational change (allosteric interaction). By removing NTD, the AGB region is liberated from such conformational constrain and can demethylate its target sites. It is also possible that the non-specific DNA binding activity of NTD competes with DME for target sites, thereby reducing the overall efficiency of DME. The molecular underpinning of how NTD induces DN effect remains to be elucidated. From an evolutionary viewpoint, the use of an active DME-based demethylation appears to have been acquired early in the plant lineage. The presence of several accessory domains in addition to the conserved core suggests adjustments to the chromatin and methylation environment of the different species. The presence of additional domains such as the DemeN and basic repeats in angiosperms and the permuted CXXC domain in streptophyta lineage might reflect the adjustment to the unique methylation and chromatin environment of the larger Streptophyta and land plant genomes.

Materials and Methods

Molecular Cloning of Constructs Used in this Study.

All general molecular manipulations followed standard procedures (Sambrook et al. 1989). Q5 High Fidelity DNA polymerase (NEB, Ipswich MA, USA) was used for PCR amplifications. PCR products were purified using AMPure XP beads (Beckman Coulter, Indianapolis IN, USA). The sequences of all plasmid constructs were confirmed by sequencing (Eton, Research Triangle Park NC, USA). All PCR primers and double-stranded DNA fragments were synthesized by Integrated DNA Technologies (Coralville IA, USA), and sequences are listed in Supplementary File 1.

A binary plasmid vector, pFGAMh, was modified to facilitate the generation of plasmid constructs using the Gibson assembly method. In brief, the replication origins and T-DNA borders originated from pFGC5941 (GenBank Accession: AY310901). A hygromycin resistance gene (HPTII) under the control of the mannopine synthase promoter was installed for selection of transgenic seedlings. A Gateway attR cassette (rfa, Invitrogen, Carlsbad CA, USA), flanked with unique restriction sites XhoI and XbaI-SpeI was placed upstream octopine synthase polyadenylation signal (OCS3’). Plasmid pFGAMh, digested with restriction enzymes XhoI and XbaI, was used to generate plasmids pDME:DME^{CTD}, pDME:nDME^{CTD} and pDME:GFP::DME^{CTD} using the Gibson assembly method. The DME.2 upstream regulatory sequence (DMEpro; 2895 bp upstream of DME.2 translation start codon ATG) was PCR-amplified using primer pair VeDME/P3R and Col-0 gDNA as template. The coding sequence of linker-AGB (with a 6-Ala linker to its N-terminus; 3174 bp), was PCR-amplified using primer pair lnAGBF/CTDVeR and Col-0 cDNA as template. To bridge these two fragments (DMEpro and linker-AGB), one of the following three DNA fragments was used in the assembly reactions. For pDME:DME^{CTD}, a 50-bp fragment was generated by annealing DNA oligos ATGF and ATGR. For pDME:SV40NLS::AGB, a 71-bp fragment was generated by annealing DNA oligos

S40F and S40R followed by two rounds of PCR reactions. For pDME:GFP::DME^{CTD}, a 761-bp fragment was PCR-amplified using primer pair p3GFPF/dmGFPR and plasmid DNA pGFP-JS (Jen Sheen,Massachusetts GeneralHospital, Boston MA, USA) as template.

An intermediate plasmid vector, DME-P3-attR-AGB, was generated by digesting plasmid pDME:SV40NLS::AGB with restriction enzymes AflII and NcoI, and re-assembled with a 2800-bp fragment, which was produced through overlap PCR with 3 primer pairs, upAflII/P3attR, P3attF/attAGBR and attAGBF/dnNcoI, and Col-0 gDNA, attR cassette and Col-0 cDNA as templates. The resulting plasmid DME-P3-attR-AGB bears (1) the same 2895-bp regulatory sequence as the above constructs, (2) an attR cassette flanked by unique restrict sites XbaI and BglII, and (3) AGB coding sequence (3156 bp). To generate pDME:DME^{FL}, plasmid DME-P3-attR-AGB was digested with XbaI and BglII, and assembled with a 2985-bp sequence, which was generated through overlap PCR using primer pairs S1-5e/IN3R and IN3F/S1-5R, and Col-0 gDNA as template. The resulting plasmid pDME:DME^{FL} carries the complete DME.2 coding sequence and intron 2 sequence (6075 bp) immediately downstream of the 2895-bp regulatory sequence with no additional sequences.

The intermediate plasmid vector DME-P3-attR-AGB was digested with restriction enzymes BglII and SpeI (to completely remove the AGB coding sequence), and re-assembled with a 786-bp sequence, which included the coding sequence of GFP (with its start codon ATG changed to TTG) and was PCR-amplified using primers ttGFPF and SpeGFPR and plasmid DNA pGFP-JS as template. The resulting plasmid DME-P3-attR-GFP was used as an intermediate plasmid vector to generate constructs pDME:DME^{NTD}::GFP and pDME:mDME^{NTD}::GFP. Plasmid DME-P3-attR-GFP was digested with XbaI and BglII, and assembled with two DNA fragments: a 3289-bp sequence was PCR-amplified using primers S1-5F and dme2tR2 and Col-0 gDNA as template and a 158-bp synthetic DNA fragment (FragQ20) (Integrated DNA Technologies, Coralville IA, USA). The resulting construct pDME:DME$^{\mathrm{NTD}}$::GFP included the 2895-bp upstream regulatory sequence, the 3332-bp sequence downstream of translation start codon ATG, the coding sequence of 6-Ala linker, and the coding sequence of GFP. Note the NTD coding sequence included the first 86 bp of intron 4 of gene DME.2, and it was designed to mimic dme-2 T-DNA insertion. To generate pDME:mDME$^{\mathrm{NTD}}$::GFP, the sequence of the first 1012 amnio acid residues of DME.2 protein was converted to DNA sequence using program EMBOSS Backtranseq (http://www.ebi.ac.uk/Tools/st/emboss_backtranseq/) and the Homo sapiens codon usage table. The sequence was then analyzed using online programs SoftBerry FSPLICE (http://linux1.softberry.com/berry.phtml?topic=fsplice&group=programs&subgroup=gfind) and NetPlantGene2 (http://www.cbs.dtu.dk/services/NetPGene/), and manually edited to disrupt potential splicing donor sites or acceptor sites. The mDME$^{\mathrm{NTD}}$ sequence (3036 bp) and upstream- and downstream-overlapping sequence are broken into 4 fragments, and synthesized by Integrated DNA Technologies (Coralville IA, USA). The 4 DNA fragments were assembled with plasmid DME-P3-attR-GFP digested with XbaI and BglII, resulting construct pDME:mNTDh::GFP.

Whole-Genome Bisulfite Sequencing and DNA Methylome Analysis

Genomic DNA were isolated from hand dissected, 7-9 DAP *dme-2* endosperm that has been complemented by *DME*^{FL} or *nDME*^{CTD} (*dme-2*/*dme-2*;*DME*^{FL}/*DME*^{FL} or *dme-2/dme-2; nDME*^{CTD} *nDME*^{CTD}). Whole genome bisulfite sequencing library was constructed as described before [13, 44]. Approximately 20-50 ng of purified genomic DNA was spiked with 0.5ng of unmethylated cl857 *Sam7* Lambda DNA (Promega, Madison, WI) and sheared to about 300bp using Covaris M220 (Covaris Inc., Woburn, Massachusetts) under the following settings: target BP, 300; peak incident power, 75 W; duty factor, 10%; cycles per burst, 200; treatment time, 90 second; sample volume 50µl. The sheared DNA was cleaned up and recovered by 1.2x AMPure XP beads then followed by end repaired and A-tailing (NEBNext Ultra II DNA Library Prep Kit for Illumina, NEB) before ligation to NEBNext methylated multiplex adapters (NEBNext Multiplex Oligos for Illumina, NEB) according to the manufacturer's instructions. Adaptor-ligated DNA was cleaned up with 1x AMPure XP beads. The purified adaptor-ligated DNA was spiked with 50ng of unmethylated cl857 *Sam7* Lambda DNA and subjected to one round of sodium bisulfite conversion using the EZ DNA Methylation-Lightning Kit (Zymo Research Corporation, Irvine, CA) as outlined in the manufacturer's instructions with 80 min of incubation time. Half of the bisulfite-converted DNA molecules was PCR amplified with the following condition: 2.5 U of ExTaq DNA polymerase (Takara), 5 ul of 10 x Extaq reaction buffer, 25 µM dNTPs, 1 ul of index Primers (10 uM) in 50 uL reaction. The thermocyling condition was as follows: 95 °C for 2 min and then 10 cycles each of 95 °C for 30 s, 65 °C for 30 s, and 72 °C for 60 s. The enriched libraries were purified twice with 0.8x (v/v) AMPure XP beads to remove any adapter dimers. High throughput sequencing was performed by Novogene Corporation (USA). For each genotype, sequencing reads from three individual transgenic lines were combined. Sequenced reads were mapped to the TAIR10 reference genomes and DNA methylation analyses were performed as previously described (Supplementary Table 7) [13]. Fractional CG methylation in 50-bp windows across the genome was compared between *dme*, wild-type (GSE38935 [13]), *DME*^{FL}- *nDME*^{CTD}- complemented *dme-2* endosperm. Windows with a fractional CG methylation difference of at least 0.3 in the endosperm comparison (Fisher's exact test p-value < 0.001) were merged to generate larger differentially methylated regions (DMRs) if they occurred within 300 bp. DMRs were retained for further analysis if the fractional CG methylation across the whole DMR was 0.3 greater in *dme* endosperm than in wild-type endosperm (Fisher's exact test p-value < 10^{-10}), and if the DMR is at least 100-bp long. The merged DMR lists are in the Supplemental File 2. The *dme* and wild-type endosperm data used in this study were derived from crossed between *Col* (female parent) and *Ler* (male parent) (GSE38935, [13]). To avoid potential ecotype-specific methylation difference, *Ler* hyper-DMRs relative to Col-0 endosperms (GSE52814, [45]) were identified using the same criteria as described above and excluded from further analyses. For making the Venn diagram, merged DMR regions were converted into 50-bp windows. Only windows with methylation scores in all samples were retained for comparison in Venn diagram and boxplot analysis.

Plant Materials and Complementation Assays

We found we can easily obtained *dme-2/dme-2* Col-*gl* plants from *DME/dme-2* heterozygotes if we rescued seeds prior to desiccation on MS sucrose plates. This is consistent with the report that *fis* endosperm cellularization defect and embryo arrest can be rescued by culturing the developing seeds in sucrose media because *fis* seeds have reduced hexose level [46]. Using this method we generated multiple homozygous lines, and we did not detect any difference between individuals in terms of normal seed rate or visible phenotype. The adult *dme-2/dme-2* plants are morphologically indistinguishable from wild-type Col-*gl* plants but produce ~0.1% viable mature seeds. These *dme-2/dme-2* plants are not due to genetic mutation or heritable aberrant epigenetic effects that escape requirement of DME activity during gametogenesis because their subsequent progeny are phenotypically normal and produces same level (~0.1%) of normal seeds.

The *DME/dme-2* heterozygous or *dme-2/dme-2* homozygous lines in Col-*gl* background were subjected to Agrobacterium-mediated floral dipping transformation procedures [28]. Seeds were sterilized by 30% bleach solution and screened for T1 transgenic plants on a 0.5x MS nutrient medium with 1% sucrose, 0.8% agar and 40 μg/ml hygromycin. Germinated seedlings were transferred to soil and grown in the growth room under 16 hours of light and 8 hours of dark cycles at 23°C. Siliques from T1 transgenic plants were dissected 14-16 days after self-pollination using a stereoscopic microscope (SteREO Discovery.V12, Carl Zeiss, Wetzlar, Germany). The numbers of viable and aborted seeds in transgenic lines were statistically analyzed with the $\chi2$ test. The probability that deviates from a 1:1 or 3:1 segregation ratio for viable and aborted seeds was also calculated.

RNA extraction, cDNA synthesis and quantitative PCR analysis

Total RNA was extracted using TRIzol® Reagent (Invitrogen, Carlsbad, USA) and treated with TURBO DNase (Ambion, Austin TX, USA) according to the manufacturers' instructions. For cDNA synthesis, 5mg of total RNA was reverse-transcribed using Superscript III Reverse Transcriptase and oligo(dT) primer (Invitrogen). The cDNA was treated with RNase H (Invitrogen) at 37oC for 20min and diluted tenfold with H2O. For each 15-μl qPCR reaction, 1μl of diluted cDNA was used. The quantitative PCR was run on ABI 7500 Fast Real-Time PCR System (http://www.appliedbiosystems.com) using FastStart Universal SYBR Green Master Mix (Roche, http://www.roche.com). The quantitative PCR primers are listed in Supplementary File 1. The Ct values were normalized against *ACT2* (*At3g18780*) mRNA or *UBC* (*At5g25760*) mRNA. The abundance of mRNAs was expressed as relative to controls, with control values set to 1. The error bars represent the standard deviation of 4 biological replicates.

Protein domain analysis and phylogenetic inference

We utilized a domain-centric computational strategy to study DME and its related proteins. Specifically, we identify DME homologs by using the iterative profile searches with PSI-BLAST [47] from the protein non-redundant (NR) database at National Center for Biotechnology Information (NCBI). Multiple sequence alignments were built by the Promals [48] program, followed by careful manual adjustments. Consensus secondary structures were predicted using the PSIPRED [49] JPred program [50]. Conserved domains were further characterized based on the comparison to available domain models from pfam [51] and sequence/structural features. The PhyML program [52] was used to determine the maximum-likelihood tree using the Jones–Taylor–Thornton (JTT) model for amino acids substitution with a discrete gamma model (four categories with gamma shape parameter: 1.096). The tree was rendered using MEGA Tree Explorer [53].

Acknowledgments

We thank Drs. Ping-Hung Hsieh and Jer-Young Lin for assistance in methylome analysis, Robert Goldberg, Robert Fischer and Matthew Bauer for discussions during the course of this study. This work is partially supported by the Hatch Project 02413 (to T.-F.H.), National Institute of Food and Agriculture, U.S. Department of Agriculture, U.S.A., by the National Science Foundation Grant MCB-1715115 to T.-F.H. and W.Y.X., and by the State of NC appropriations as distributed by the University of North Carolina General Administration and the NC Agricultural Research Service Office at NC State University. LMI and LA are supported by the Intramural Research Program of the National Library of Medicine, NIH, USA.

Author contribution.

C.Z., Y.-H.H., X.-Q.Z., J.H.H. and T.-F.H. designed the research. C.Z., Y.-H.H., X.-Q.Z. performed the experiments. D.Z., L.M.I, and L.A. performed the evolutionary analysis. C.Z., Y.-H.H., and T.-F.H. wrote the article. T.-F.H., C.Z., Y.-H.H., W.X., J.H.H. interpreted and commented the article.

References

1. Yang WC, Shi DQ, Chen YH. Female gametophyte development in flowering plants. *Annu Rev Plant Biol* **61**, 89-108 (2010).

2. Gehring M. Genomic imprinting: insights from plants. *Annu Rev Genet* **47**, 187-208 (2013).

3. Kohler C, Wolff P, Spillane C. Epigenetic mechanisms underlying genomic imprinting in plants. *Annu Rev Plant Biol* **63**, 331-352 (2012).

4. Jullien PE, Kinoshita T, Ohad N, Berger F. Maintenance of DNA Methylation during the Arabidopsis Life Cycle Is Essential for Parental Imprinting. *Plant Cell* **18**, 1360-1372 (2006).

5. Gehring M*, et al.* DEMETER DNA glycosylase establishes MEDEA polycomb gene self-imprinting by allele-specific demethylation. *Cell* **124**, 495-506 (2006).

6. Xiao W*, et al.* Imprinting of the MEA Polycomb gene is controlled by antagonism between MET1 methyltransferase and DME glycosylase. *Dev Cell* **5**, 891-901 (2003).

7. Kinoshita T*, et al.* One-way control of FWA imprinting in Arabidopsis endosperm by DNA methylation. *Science* **303**, 521-523 (2004).

8. Tiwari S*, et al.* MATERNALLY EXPRESSED PAB C-TERMINAL, a novel imprinted gene in Arabidopsis, encodes the conserved C-terminal domain of polyadenylate binding proteins. *Plant Cell* **20**, 2387-2398 (2008).

9. Choi Y*, et al.* DEMETER, a DNA Glycosylase Domain Protein, Is Required for Endosperm Gene Imprinting and Seed Viability in *Arabidopsis*. *Cell* **110**, 33-42 (2002).

10. Penterman J, Zilberman D, Huh JH, Ballinger T, Henikoff S, Fischer RL. DNA demethylation in the Arabidopsis genome. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 6752-6757 (2007).

11. Lister R*, et al.* Highly integrated single-base resolution maps of the Arabidopsis genome. *Cell* **133**, 395-397 (2008).

12. Gong Z, Morales-Ruiz T, Ariza RR, Roldan-Arjona T, David L, Zhu J-K. ROS1, a Repressor of Transcriptional Gene Silencing in Arabidopsis, Encodes a DNA Glycosylase/Lyase. *Cell* **111**, 803-814 (2002).

13. Ibarra CA*, et al.* Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science* **337**, 1360-1364 (2012).

14. Hsieh TF*, et al.* Regulation of imprinted gene expression in Arabidopsis endosperm. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 1755-1762 (2011).

15. Hsieh TF*, et al.* Genome-wide demethylation of Arabidopsis endosperm. *Science* **324**, 1451-1454 (2009).

16. Wang X*, et al.* RNA-binding protein regulates plant DNA methylation by

controlling mRNA processing at the intronic heterochromatin-containing gene IBM1. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15467-15472 (2013).

17. Lei M, Zhang H, Julian R, Tang K, Xie S, Zhu JK. Regulatory link between DNA methylation and active demethylation in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 3553-3557 (2015).

18. Lang Z*, et al.* The methyl-CpG-binding protein MBD7 facilitates active DNA demethylation to limit DNA hyper-methylation and transcriptional gene silencing. *Mol Cell* **57**, 971-983 (2015).

19. Qian W*, et al.* A histone acetyltransferase regulates active DNA demethylation in Arabidopsis. *Science* **336**, 1445-1448 (2012).

20. Ikeda Y*, et al.* HMG domain containing SSRP1 is required for DNA demethylation and genomic imprinting in Arabidopsis. *Dev Cell* **21**, 589-596 (2011).

21. Bustin M, Catez F, Lim JH. The dynamics of histone H1 function in chromatin. *Mol Cell* **17**, 617-620 (2005).

22. Fan Y*, et al.* Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell* **123**, 1199-1212 (2005).

23. Graziano V, Gerchman SE, Schneider DK, Ramakrishnan V. Histone H1 is located in the interior of the chromatin 30-nm filament. *Nature* **368**, 351-354 (1994).

24. Hashimoto H*, et al.* Histone H1 null vertebrate cells exhibit altered nucleosome architecture. *Nucleic Acids Res* **38**, 3533-3545 (2010).

25. Rea M*, et al.* Histone H1 affects gene imprinting and DNA methylation in Arabidopsis. *Plant J* **71**, 776-786 (2012).

26. Brooks SC, Fischer RL, Huh JH, Eichman BF. 5-methylcytosine recognition by Arabidopsis thaliana DNA glycosylases DEMETER and DML3. *Biochemistry* **53**, 2525-2532 (2014).

27. Jang H, Shin H, Eichman BF, Huh JH. Excision of 5-hydroxymethylcytosine by DEMETER family DNA glycosylases. *Biochem Biophys Res Commun* **446**, 1067-1072 (2014).

28. Clough SJ, Bent AF. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J* **16**, 735-743 (1998).

29. Park JS*, et al.* Control of DEMETER DNA demethylase gene transcription in male and female gamete companion cells in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 2078-2083 (2017).

30. Kohler C, Hennig L, Bouveret R, Gheyselinck J, Grossniklaus U, Gruissem W. Arabidopsis MSI1 is a component of the MEA/FIE Polycomb group complex and required for seed development. *The EMBO journal* **22**, 4804-4814 (2003).

31. Grossniklaus U, Vielle-Calzada J-P, Hoeppner MA, Gagliano WB. Maternal control of embryogenesis by *MEDEA*, a polycomb-group gene in *Arabidopsis*. *Science* **280**, 446-450 (1998).

32. Luo M, Bilodeau P, Dennis ES, Peacock WJ, Chaudhury A. Expression and parent-of-origin effects for FIS2, MEA, and FIE in the endosperm and embryo of

developing Arabidopsis seeds. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 10637-10642 (2000).

33. Schoft VK*, et al.* Function of the DEMETER DNA glycosylase in the Arabidopsis thaliana male gametophyte. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 8042-8047 (2011).

34. Lang Z*, et al.* Critical roles of DNA demethylation in the activation of ripening-induced genes and inhibition of ripening-repressed genes in tomato fruit. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E4511-E4519 (2017).

35. Choi Y, Harada JJ, Goldberg RB, Fischer RL. An invariant aspartic acid in the DNA glycosylase domain of DEMETER is necessary for transcriptional activation of the imprinted MEDEA gene. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 7481-7486 (2004).

36. Iyer LM, Abhiman S, Aravind L. Natural history of eukaryotic DNA methylation systems. *Progress in molecular biology and translational science* **101**, 25-104 (2011).

37. Walsh P, Bursac D, Law YC, Cyr D, Lithgow T. The J-protein family: modulating protein assembly, disassembly and translocation. *EMBO reports* **5**, 567-571 (2004).

38. Hong S, Hashimoto H, Kow YW, Zhang X, Cheng X. The Carboxy-Terminal Domain of ROS1 Is Essential for 5-Methylcytosine DNA Glycosylase Activity. *J Mol Biol*, (2014).

39. Long HK, Blackledge NP, Klose RJ. ZF-CxxC domain-containing proteins, CpG

islands and the chromatin connection. *Biochem Soc Trans* **41**, 727-740 (2013).

40. Mok YG*, et al.* Domain structure of the DEMETER 5-methylcytosine DNA glycosylase. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 19225-19230 (2010).

41. Zheng X*, et al.* ROS3 is an RNA-binding protein required for DNA demethylation in Arabidopsis. *Nature* **455**, 1259-1262 (2008).

42. Ponferrada-Marin MI, Roldan-Arjona T, Ariza RR. Demethylation initiated by ROS1 glycosylase involves random sliding along DNA. *Nucleic Acids Res* **40**, 11554-11562 (2012).

43. Ponferrada-Marin MI, Martinez-Macias MI, Morales-Ruiz T, Roldan-Arjona T, Ariza RR. Methylation-independent DNA binding modulates specificity of Repressor of Silencing 1 (ROS1) and facilitates demethylation in long substrates. *The Journal of biological chemistry* **285**, 23032-23039 (2010).

44. Hsieh TF. Whole-genome DNA methylation profiling with nucleotide resolution. *Methods in molecular biology* **1284**, 27-40 (2015).

45. Pignatta D, Erdmann RM, Scheer E, Picard CL, Bell GW, Gehring M. Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *Elife* **3**, e03198 (2014).

46. Hehenberger E, Kradolfer D, Kohler C. Endosperm cellularization defines an important developmental transition for embryo development. *Development* **139**, 2031-2039 (2012).

47. Altschul SF*, et al.* Gapped BLAST and PSI-BLAST: a new generation of protein

database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

48. Pei J, Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics* **23**, 802-808 (2007).

49. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res* **41**, W349-357 (2013).

50. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. *Bioinformatics* **14**, 892-893 (1998).

51. Finn RD*, et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-285 (2016).

52. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321 (2010).

53. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596-1599 (2007).

Figure Legends

Figure 1 Complementation of *dme* seed abortion phenotype by the truncated DME nAGB.

(a) Siliques were dissected and photographed 14 days after self-pollination. In *dme-2/dme-2* silique greater than 99% of seeds are aborted. A single copy of *nDME*^{CTD} transgene reduces seed abortion rate to 50%; and in the *dme-2/dme-2; nDME*^{CTD}*/nDME*^{CTD} silique, all the *dme-2* seeds are rescued and developed normally. Scale bar = 0.5 mm. (b) Complementation of *dme-2* seed abortion phenotype by *nDME*^{CTD} *and DME*^{FL}. (c) The *nDME*^{CTD} transgene restores DME target genes *FWA* and *FIS2* expression. WT: Col-0; *nDME*^{CTD}: *dme-2/dme-2; nDME*^{CTD}*/ nDME*^{CTD}; *dme-2*: *dme-2/dme-2*. Total RNA was isolated from stage F1 to F12 floral buds.

Figure 2 Endosperm methylome analysis. (a) Genome browser snapshots of CG DNA methylation at selected imprinted gene loci. Top two tracks are coding genes (magenta) and TEs (orange) with Tair10 chromosome coordinates. For the bottom seven tracks, each track represents fractional CG methylation levels for different genotype: black trace, *dme-2* endosperm; dark green trace, WT endosperm; dark blue trace, *DME*^{FL}-complemented endosperm; dark purple trace, *nDME*^{CTD}-complemented endosperm; light green trace, WT endosperm subtracted from *dme-2* mutant endosperm; light blue trace, *DME*^{FL}-complemented endosperm subtracted from *dme-2* endosperm; light purple trace, *nDME*^{CTD}-complemented endosperm subtracted form *dme-2* endosperm. DNA CG hypomethylation at selected maternally expressed (*FIS2* and *SDC*) and paternally expressed (*SUV7*, *YUC10*, *and PHE1*) imprinted genes is restored in *DME*^{FL}- and *nDME*^{CTD}-complemented endosperm. (b) Boxplot of CG methylation levels among canonical DME target sites in *dme-2* mutant (grey), WT (white), *DME*^{FL}- (blue), or *nDME*^{CTD} - (red) complemented endosperm. (c) Venn Diagram (top panel) of CG hyper-DMRs in 50-bp windows between *dme-2* endosperm relative to WT, *DME*^{FL}-complemented or *nDME*^{CTD}-complemented endosperm. Boxplot (bottom panel) of CG methylation levels in *dme-2* mutant (grey), WT (white), *DME*^{FL}- (blue) or *nDME*^{CTD}- (red) complemented endosperm in WT only (left panel), *DME*^{FL} only, or *nDME*^{CTD} only (right panel) DMRs.

Figure 3 Expression of DME NTD region in wild-type central cell induces *dme*-like seed abortion phenotype. (a) Confocal microscopy image of ovule in F12 floral bud shows the expression of mDME^{NTD}-GFP in the central cell. Scale bar, 20 µm. (**b-c**) Ectopic expression of *DME^{NTD}* in WT central cell induces *dme-2* like seed abortion phenotype in silique (**b**) and in developing seeds (**c**). Total RNA was isolated from stage F1 to F12 floral buds from independent lines with different seed abortion ratios (**d**) to assess transgene and endogenous DME expression. (**e**) Endogenous DME transcript levels in independent transgenic lines are comparable to the control line, but the transgene expression level varies among these independent lines with different seed abortion rates. Error bars indicate SD. NS, $p > 0.2$ (Ctrl vs 23), $p > 0.5$ (Ctrl vs 15), $p > 0.3$ (Ctrl vs 25), $p > 0.4$ (Ctrl vs 8), not significant (two-tailed t test). (**f**) Correlation analysis shows that the transcript abundance of the transgene, but not that of the endogenous DME transcripts, correlates with seed abortion rates (by linear regression).

Figure 4. Evolution of plant DME-like proteins. A phylogenetic tree was reconstructed using the PhyML program. Only node supporting values >0.80 from ML bootstrap analyses are shown. The representative domain architectures of DME homologs in major plant clades are shown along the tree, demonstrating domain fusions during evolution. Domain abbreviations: DemeN, N-terminal domain of DEME-like proteins in angiosperms; DnaJ, DnaJ molecular chaperone homology domain (Pfam: PF00226); FCL, [Fe4S4] cluster loop motif (also called Iron-sulfur binding domain of endonuclease III; Pfam: PF10576); HhH-GL, HhH-GPD superfamily base excision DNA repair protein (Pfam: PF00730); PHD, PHD finger (Pfam: PF00628); RRM, RNA recognition motif (Pfam: PF00076); Tudor, Tudor domain (Pfam: PF00567).

Supplemental Information

Figure Legends

Fig. S1. Diagrams of DME protein structure and transgene constructs.

(a) DME protein domain architecture. The positions of conserved domains along DME protein. Numbers represent amino acid position relative to the translation start sites. DME.1 is shorter than DME.2 by 258 amino acids due to alternative splicing, missing the very N-terminal DemeN domain. DemeN is a domain of unknown function conserved among angiosperm DME-like protein. 3DR is the stretch of basic rich amino acid direct repeats, resembling AT-hook motifs, and serves as a nuclear localization signal; per-CXXC is the permuted CXXC zinc finger motif; RRM is the RNA recognition motif; FCL is a [Fe4S4] cluster loop following the HhH module. The *dme-2* allele harbors a T-DNA insertion in region A at amino acid position 1012. ID1 and ID2 are variable, low complexity sequences between the glycosylase domain and the conserved B region. (b) Transgene constructs used in this study. DMEpro refers to the upstream regulatory sequence (2895 bp upstream of the translation start codon ATG) of DME.2. SV40NLS: PKKKRKV. A polypeptide linker comprising 6 alanine residues is placed between any protein fragment fusions.

Fig. S2. DNA methylomes of three independent *nDME*^{CTD}-complemented *dme-2* endosperm.

(a) Venn diagram showing partial overlap of *dme* CG hyper-DMRs relatives to each nAGB-complemented endosperm (*nDME*^{CTD}*-1* to *nDME*^{CTD}-3). (b) Boxplot of CG methylation levels among canonical DME target sites in *dme-2* mutant (black), *nDME*^{CTD}-1 (pink), *nDME*^{CTD}*-2* (magenta), or *nDME*^{CTD}*-3* (red) complemented endosperm, in *nDME*^{CTD}*-1* specific (left panel), *nDME*^{CTD}*-2* specific (middle panel), and *nDME*^{CTD}*-3* specific DMRs. These results show that the combined DMRs are more or less hypomethylated in each independent line compared to *dme-2* endosperm.

Fig. S3. DNA methylomes of three independent DME^{FL-}complemented *dme-2* endosperm. (a) Venn diagram showing partial-overlap of *dme* CG hyper-DMRs relatives to each *DME^{FL}*-complemented endosperm (*DME^{FL}-1* to *DME^{FL}-3*). (b) Boxplot of CG methylation levels among canonical DME target sites in *dme-2* mutant (black), *DME^{FL}*-1 (light blue), *DME^{FL}-2* (medium blue), or *DME^{FL}-3* (dark blue) complemented endosperm, in *DME^{FL}-1* specific (left panel), *DME^{FL}-2* specific (middle panel), and *DME^{FL}-3* specific DMRs. These results show that the combined DMRs are more or less hypomethylated in each independent line compared to *dme-2* endosperm.

Fig. S4. The DMRs of *dme* relative to WT endosperm or nDME^{CTD}- complemented endosperm. Venn Diagram (top) and Boxplot analysis (bottom) of CG hyper-DMRs in 50-bp windows between *dme-2* endosperm relative to *nDME^{CTD}*-complemented or WT endosperm. CG methylation levels of DMRs unique to *nDME^{CTD}*-complemented endosperm are also demethylated in the WT endosperm (left panel). Similarly, DMRs unique to WT endosperm are demethylated in *nDME^{CTD}*-complemented endosperm (right).

Fig. S5. DME^{FL} and nDME^{CTD} transgenes are expressed at comparable levels among independent complementation lines. *DME^{FL}* and *nDME^{CTD}* expression levels are comparable between the four of the six complementation lines used in the methylome study. Total RNA was isolated from stage F1 to F12 floral buds. The results show that there is no significant difference

in expression level between these two transgenes(t-test, *p*>0.4).

Fig. S6. The effects of T-DNA insertion on endogenous DME transcript abundance in *dme-2/dme-2* plants. Total RNA was isolated from stage F1 to F12 floral buds. Equal amount of total RNA from WT and *dme-2/dme-2* were used for reverse transcription and quantitative PCR. Six paired of primers (PN1-PN6) correspond to the N-terminal region before the T-DNA insertion site, and three pairs of C-terminal region primers (PC1-PC3) were used to assess endogenous DME transcript level in *dme-2/dme-2* mutant plants. The position of each primer pair is indicated in the DME diagram where T-DNA insertion site is shown.

Fig. S7. Alignment of angiosperm DME-like proteins showing the conserved DemeN domain and the basic rich 3DR repeats. Bioinformatics analysis using available DME-like sequences identified a ~ 120-amino-acid-long conserved region at the very N-termini among DME-like proteins in angiosperms. This sequence is characterized by a highly conserved WxPxTPxK motif that might function in protein-protein interactions. Further toward the C-terminus is a stretch of basic amino acids rich region that serves as a nuclear localization signal. This sequence consists of three direct repeats (3DR) reminiscent of the AT-hook motifs that may bind DNA.

Fig. S8. Boxplot of CG methylation levels among canonical DME target sites in different DMR length category, in *dme-2* mutant (black), wild-type (white), or *nDME*^{CTD}-complemented endosperm

Fig. S9. (a) Merged DMR length distribution in WT and *nDME*^{CTD}-complemented endosperm. (b) Genome Browser examples of long WT DMRs. Tracks are as labeled. The DMR regions are indicated as horizontal bars according to their length in each sample (bottom two tracks). Even though *nDME*^{CTD} complemented endosperm lack longer DMRs, these regions are also shorter DMRs in *nDME*^{CTD}-complemented endosperm.

Table 1. Rescue of the reduced paternal *dme-2* allele transmission by the *nDME*^{CTD} transgene.

Female	Male parent	F1, DME/*dme-2*	F1, *DME/dme-2*; *nDME*^{CTD}	*nDME*^{CTD} transmission rate (%)	*p* for 1:1†
Col-0	*dme-2/dme-2; nDME*^{CTD} /~ Line 1	32	62	66	2.0E-3
Col-0	*dme-2/dme-2; nDME*^{CTD} /~ Line 2	3	50	94.3	1.1E-10
Col-0	*dme-2/dme-2; nDME*^{CTD} /~ Line 3	8	34	81	6.0E-5
Col-0	*dme-2/dme-2; nDME*^{CTD} /~ Line 4	9	44	83	1.5E-6

† Probability that that the deviation from the indicated segregation ration (1:1 inheritance of paternal genome with or without *nDME*^{CTD} transgene in the F1 generation) is due to chance.

Figure 1

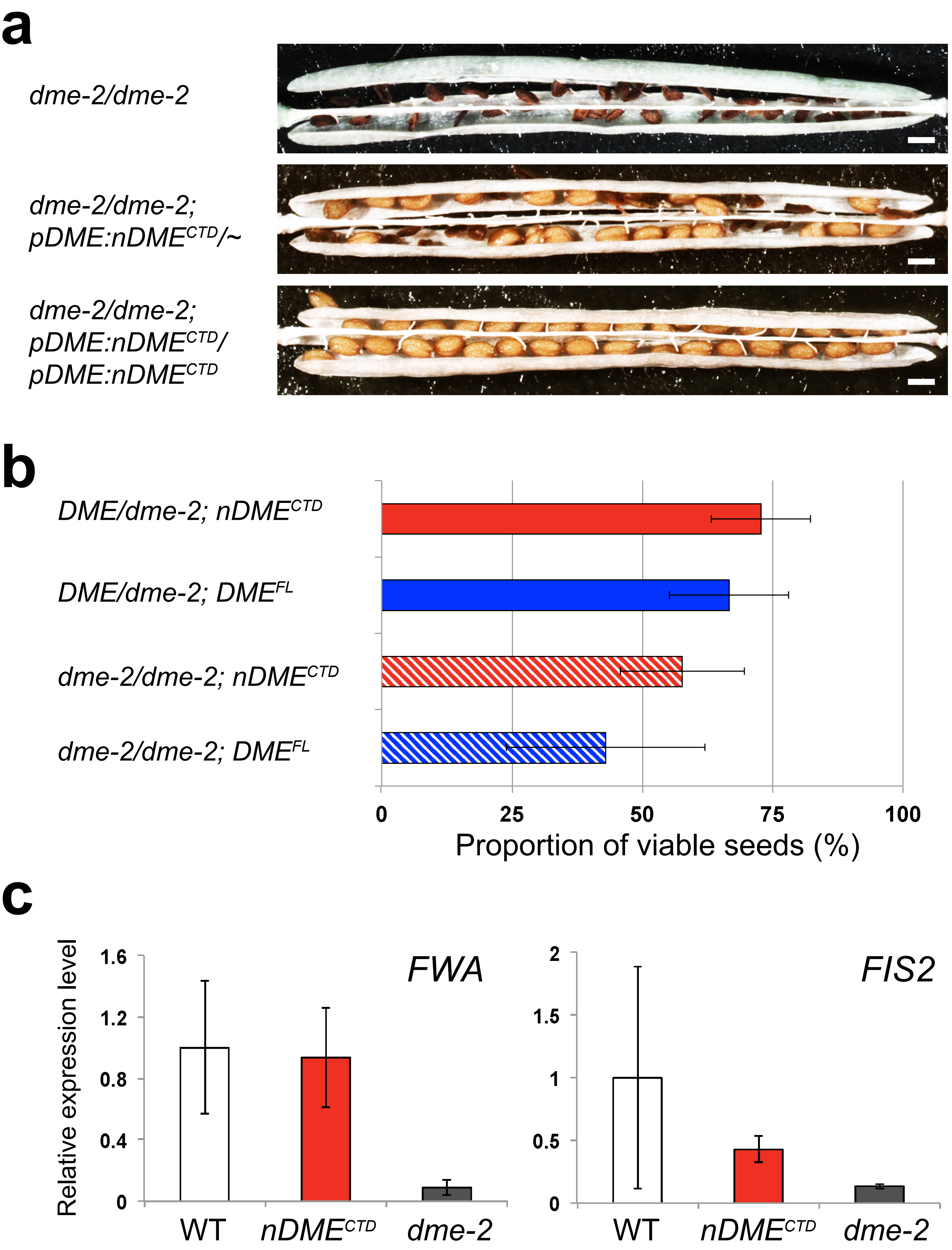

a
dme-2/dme-2
dme-2/dme-2; pDME:nDME^{CTD}/~
dme-2/dme-2; pDME:nDME^{CTD}/ pDME:nDME^{CTD}
b
DME/dme-2; nDME^{CTD}
DME/dme-2; DME^{FL}
dme-2/dme-2; nDME^{CTD}
dme-2/dme-2; DME^{FL}
0
25
50
75
100
Proportion of viable seeds (%)
c
Relative expression level
1.6
1.2
0.8
0.4
0
FWA
WT
nDME^{CTD}
dme-2
2
1.5
1
0.5
0
FIS2
WT
nDME^{CTD}
dme-2

Figure 2

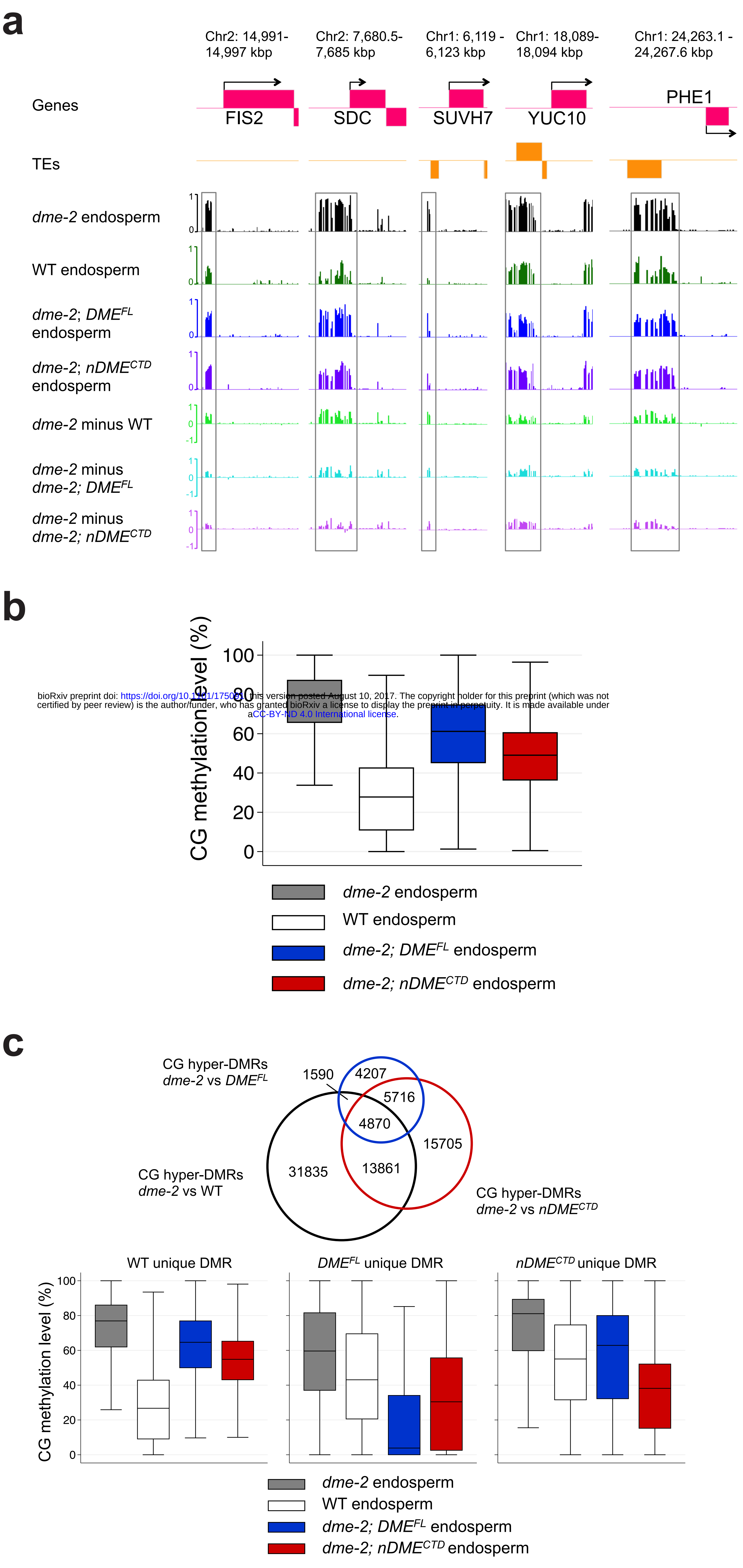
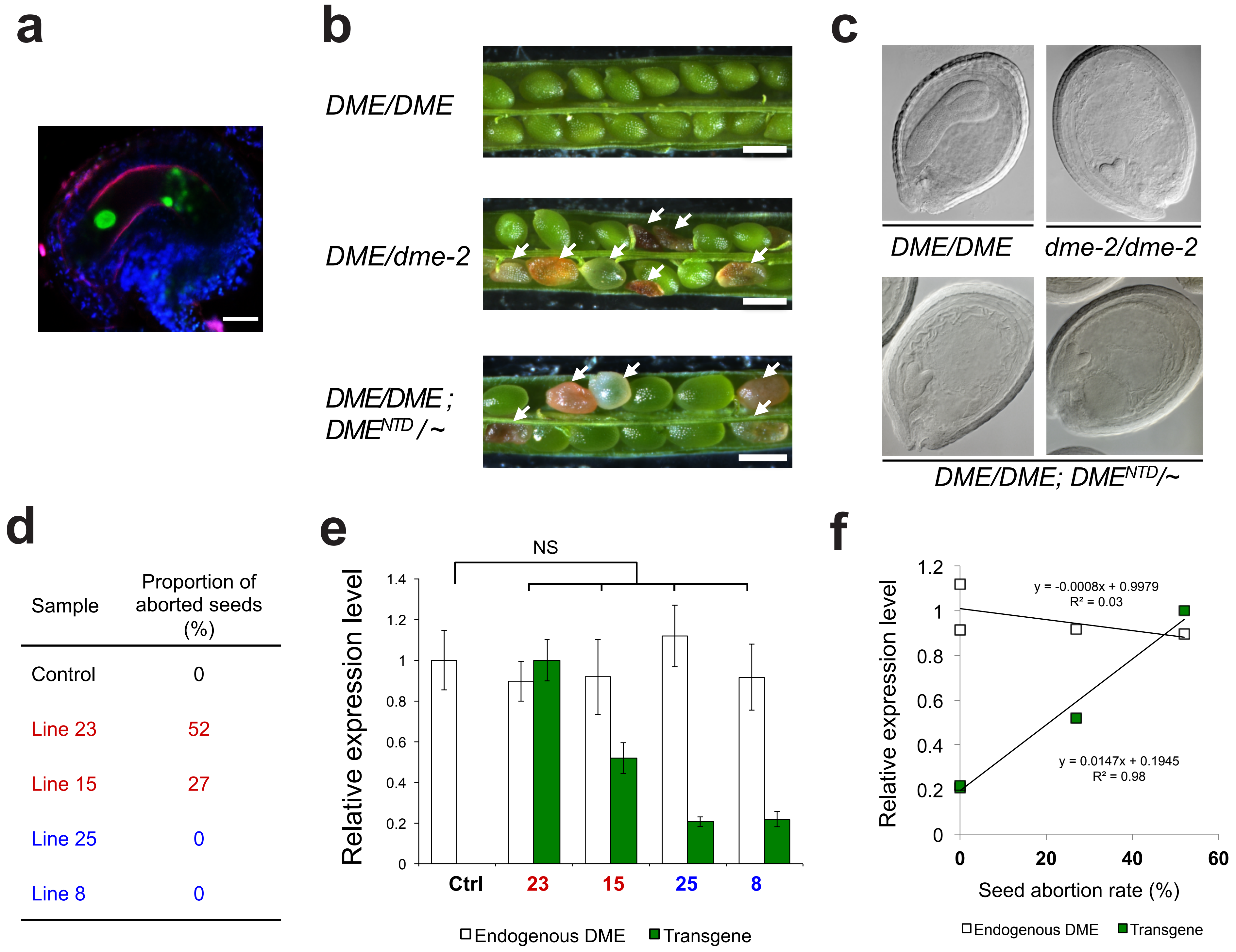
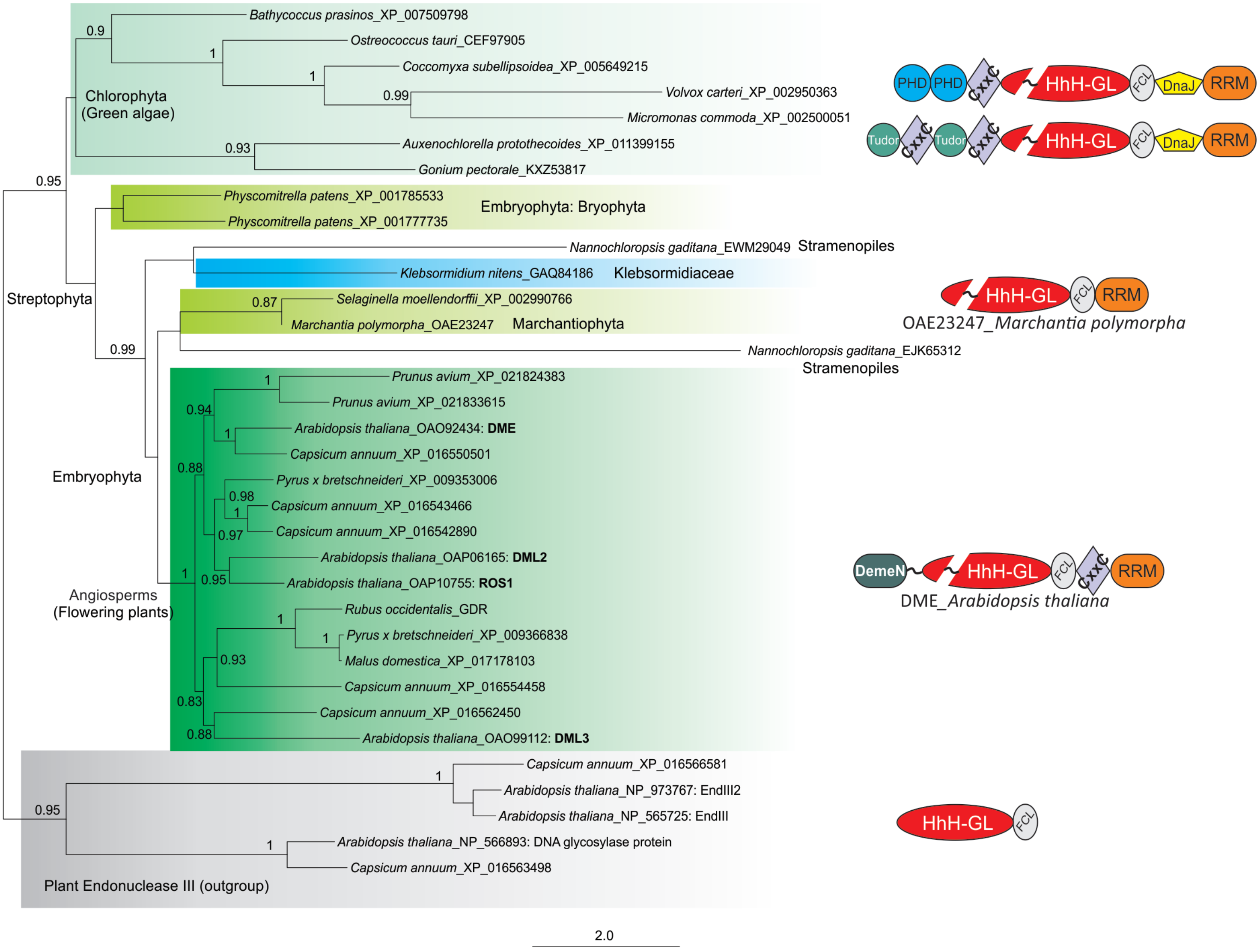

a
Chr2: 14,991-14,997 kbp
Chr2: 7,680.5-7,685 kbp
Chr1: 6,119 - 6,123 kbp
Chr1: 18,089-18,094 kbp
Chr1: 24,263.1 - 24,267.6 kbp
Genes
FIS2
SDC
SUVH7
YUC10
PHE1
TEs
dme-2 endosperm
WT endosperm
dme-2; DME^FL endosperm
dme-2; nDME^CTD endosperm
dme-2 minus WT
dme-2 minus dme-2; DME^FL
dme-2 minus dme-2; nDME^CTD
1
0
-1
b
CG methylation level (%)
100
80
60
40
20
0
▨ dme-2 endosperm
□ WT endosperm
🟦 dme-2; DME^FL endosperm
🟥 dme-2; nDME^CTD endosperm
c
CG hyper-DMRs dme-2 vs DME^FL
CG hyper-DMRs dme-2 vs WT
CG hyper-DMRs dme-2 vs nDME^CTD
1590
4207
5716
4870
15705
31835
13861
WT unique DMR
DME^FL unique DMR
nDME^CTD unique DMR
CG methylation level (%)
100
80
60
40
20
0
▨ dme-2 endosperm
□ WT endosperm
🟦 dme-2; DME^FL endosperm
🟥 dme-2; nDME^CTD endosperm

Figure 3

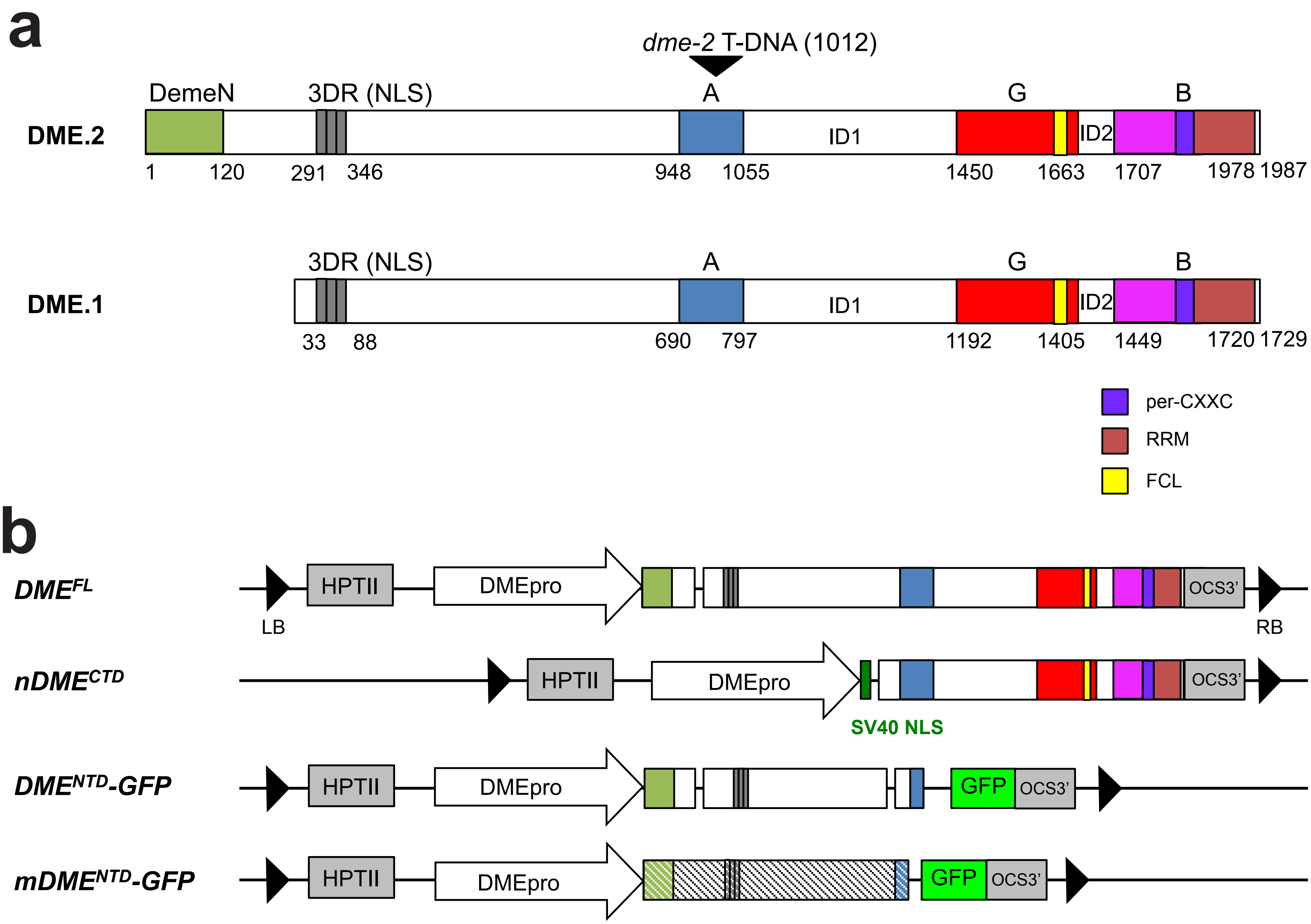


Sample	Proportion of aborted seeds (%)
Control	0
Line 23	52
Line 15	27
Line 25	0
Line 8	0

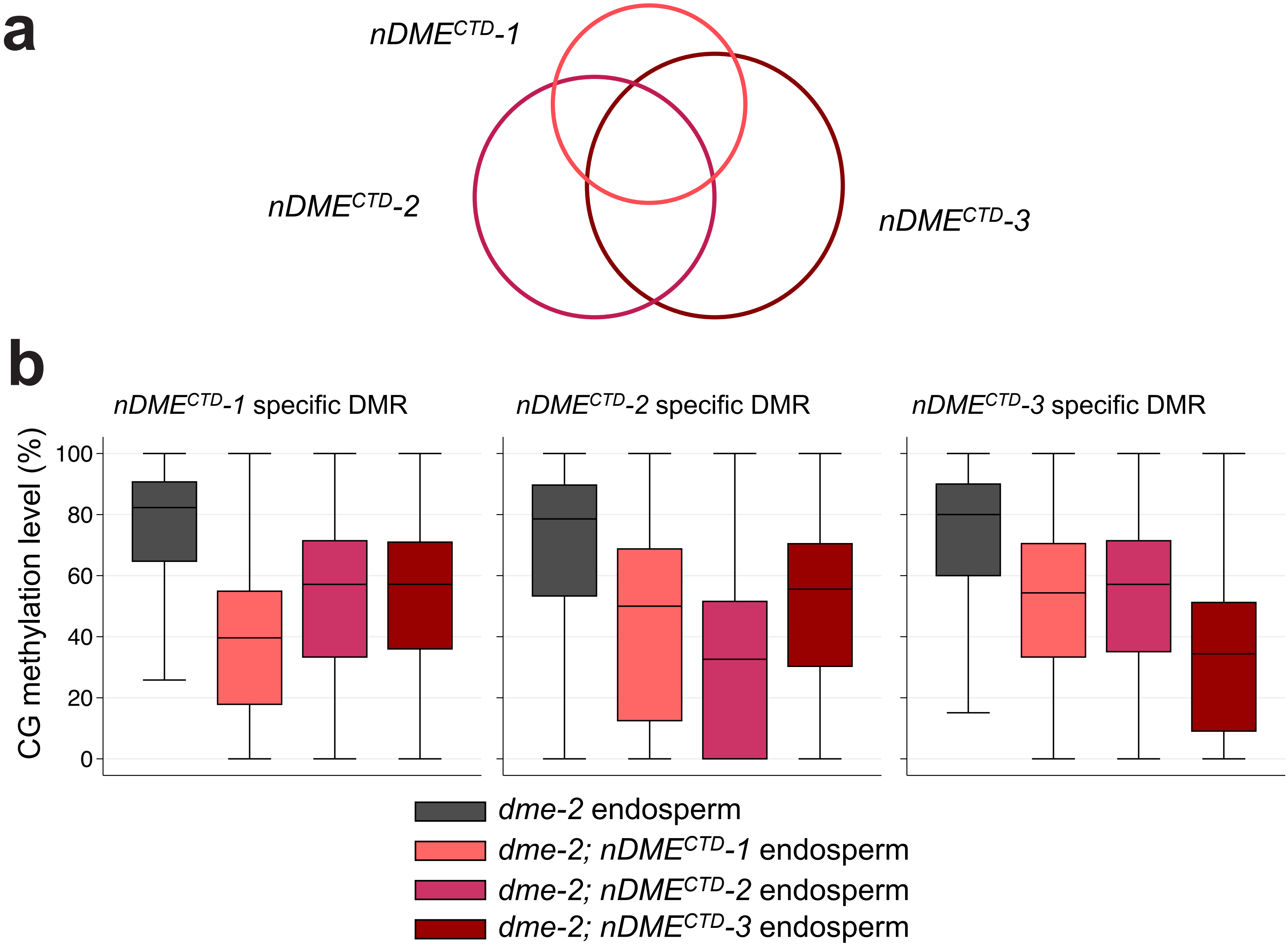
Figure 4


Bathycoccus prasinos_XP_007509798
0.9
Ostreococcus tauri_CEF97905
1
Coccomyxa subellipsoidea_XP_005649215
1
Chlorophyta
(Green algae)
0.99
Volvox carteri_XP_002950363
Micromonas commoda_XP_002500051
0.93
Auxenochlorella protothecoides_XP_011399155
Gonium pectorale_KXZ53817
PHD PHD CxxC HhH-GL FCL DnaJ RRM
Tudor CxxC Tudor CxxC HhH-GL FCL DnaJ RRM
0.95
Physcomitrella patens_XP_001785533
Physcomitrella patens_XP_001777735
Embryophyta: Bryophyta
Nannochloropsis gaditana_EWM29049
Stramenopiles
Klebsormidium nitens_GAQ84186
Klebsormidiaceae
Streptophyta
0.87
Selaginella moellendorffii_XP_002990766
Marchantia polymorpha_OAE23247
Marchantiophyta
HhH-GL FCL RRM
OAE23247_Marchantia polymorpha
0.99
Nannochloropsis gaditana_EJK65312
Stramenopiles
1
Prunus avium_XP_021824383
Prunus avium_XP_021833615
0.94
1
Arabidopsis thaliana_OAO92434: DME
Capsicum annuum_XP_016550501
0.88
Embryophyta
Pyrus x bretschneideri_XP_009353006
0.98
1
Capsicum annuum_XP_016543466
Capsicum annuum_XP_016542890
0.97
Arabidopsis thaliana_OAP06165: DML2
DemeN HhH-GL FCL CxxC RRM
1
0.95
Arabidopsis thaliana_OAP10755: ROS1
DME_Arabidopsis thaliana
Angiosperms
(Flowering plants)
Rubus occidentalis_GDR
1
1
Pyrus x bretschneideri_XP_009366838
0.93
Malus domestica_XP_017178103
Capsicum annuum_XP_016554458
0.83
Capsicum annuum_XP_016562450
0.88
Arabidopsis thaliana_OAO99112: DML3
Capsicum annuum_XP_016566581
1
Arabidopsis thaliana_NP_973767: EndIII2
0.95
Arabidopsis thaliana_NP_565725: EndIII
HhH-GL FCL
1
Arabidopsis thaliana_NP_566893: DNA glycosylase protein
Capsicum annuum_XP_016563498
Plant Endonuclease III (outgroup)
2.0

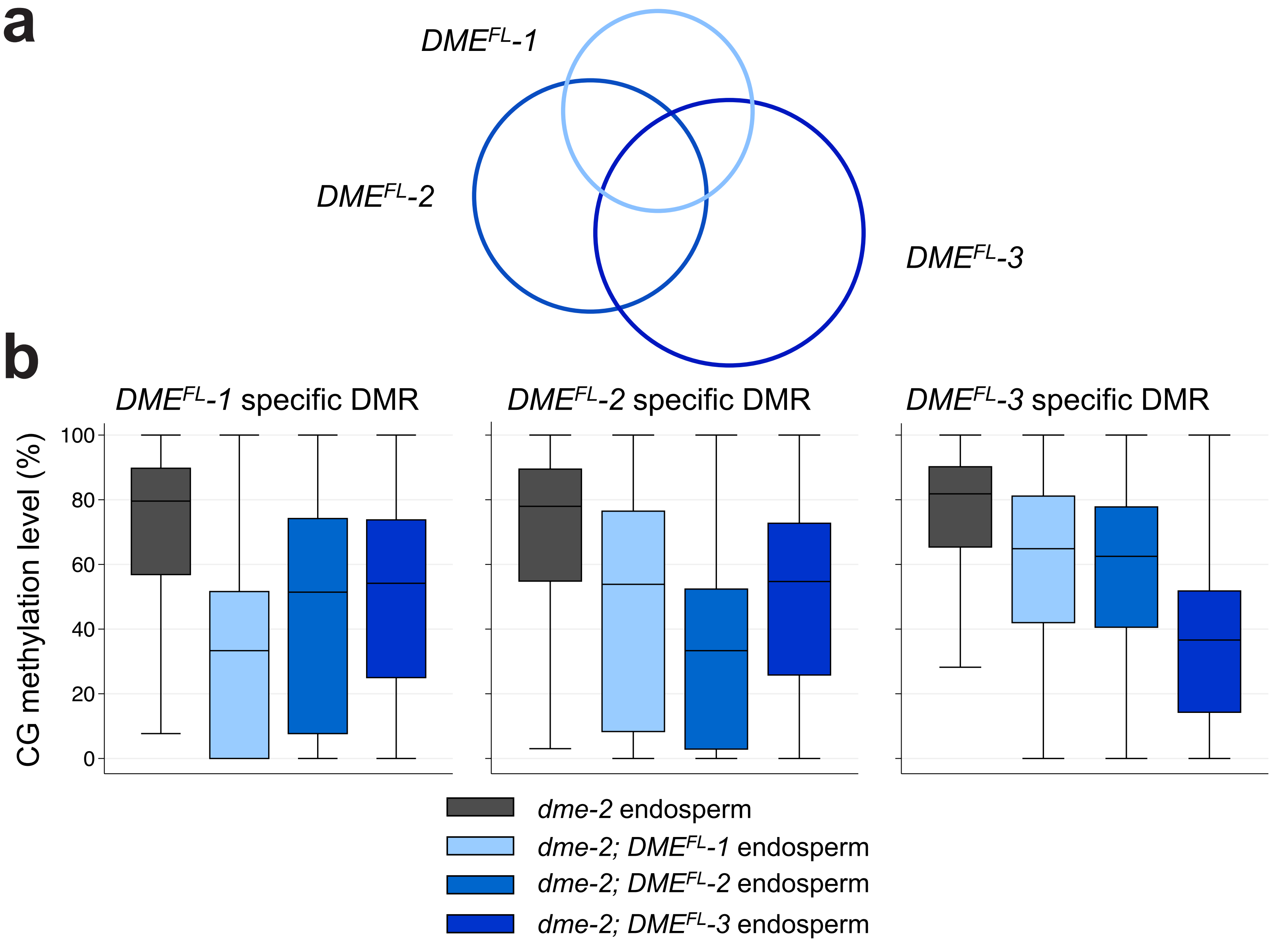
Supplementary Figure S1



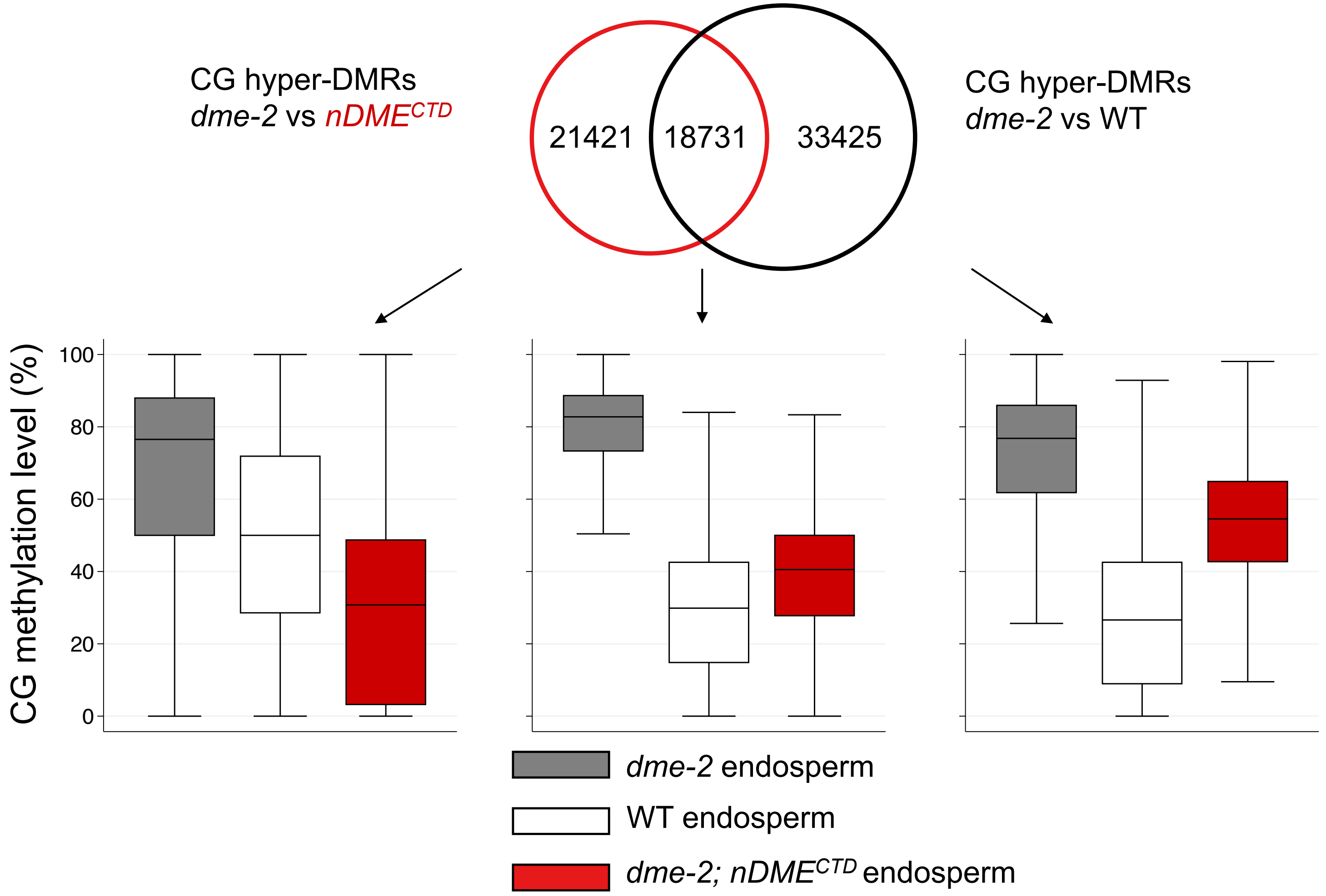
Supplementary Figure S2



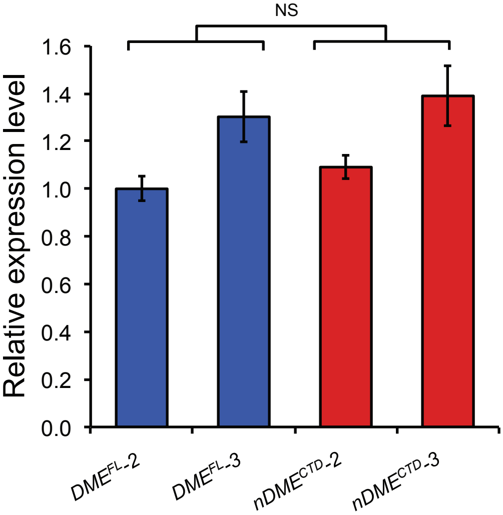
Supplementary Figure S3



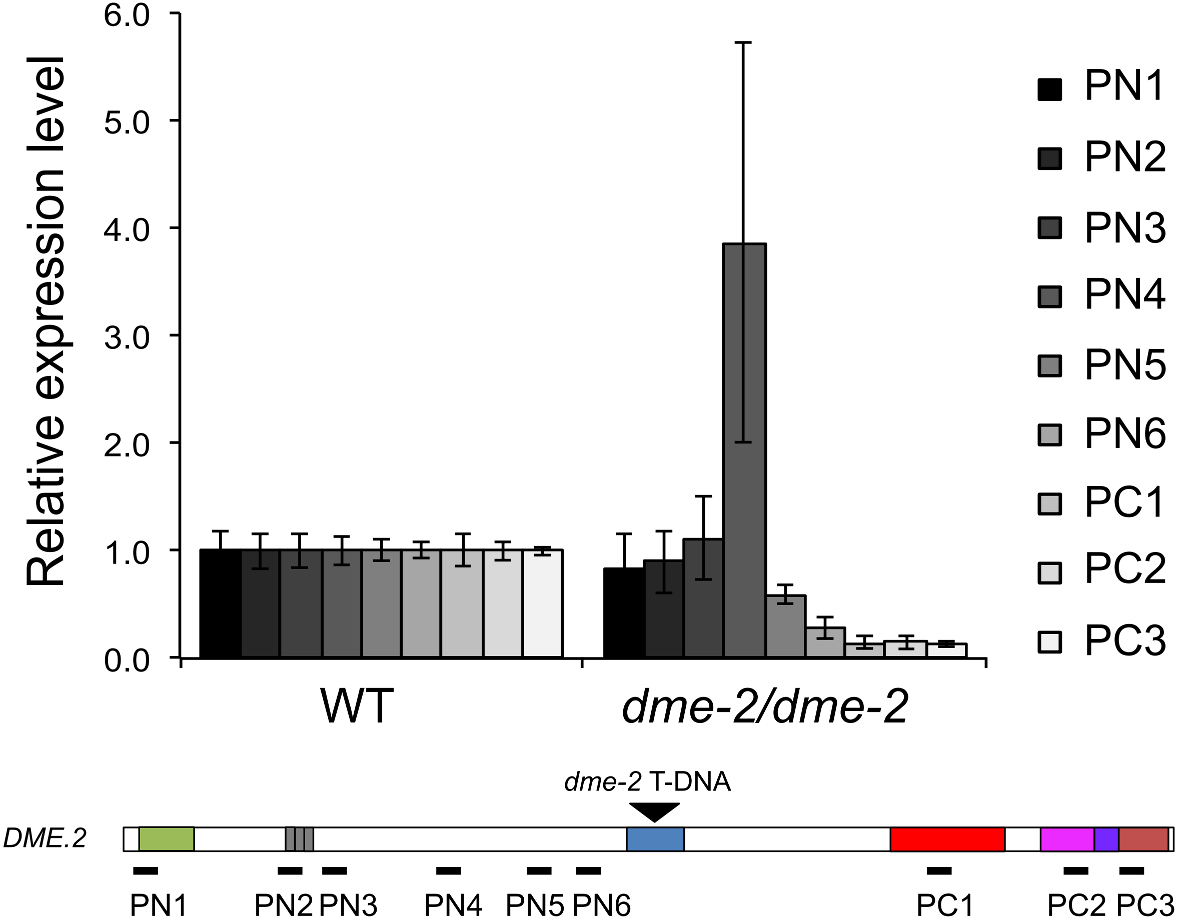
Supplementary Figure S4



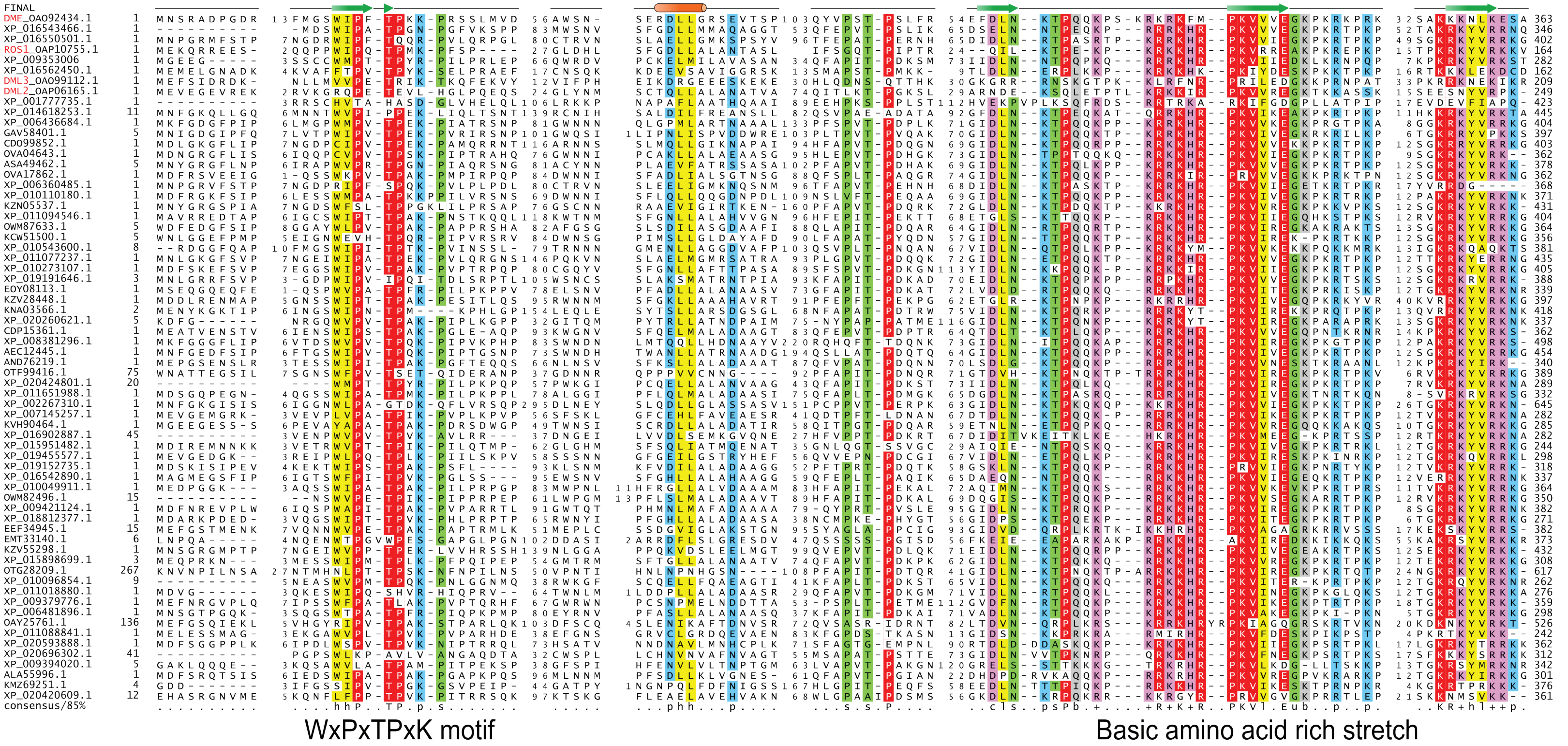
Supplementary Figure S5



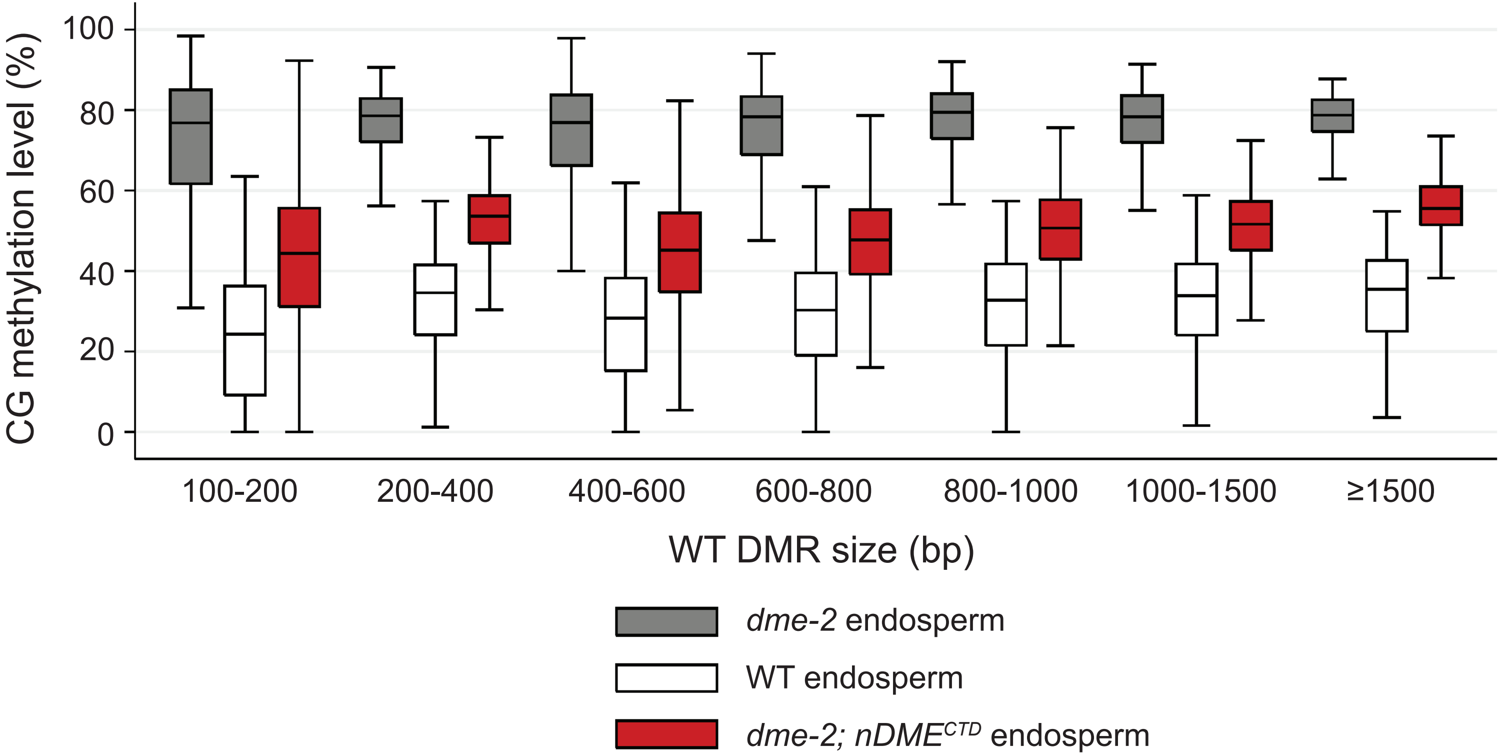
Supplementary Figure S6



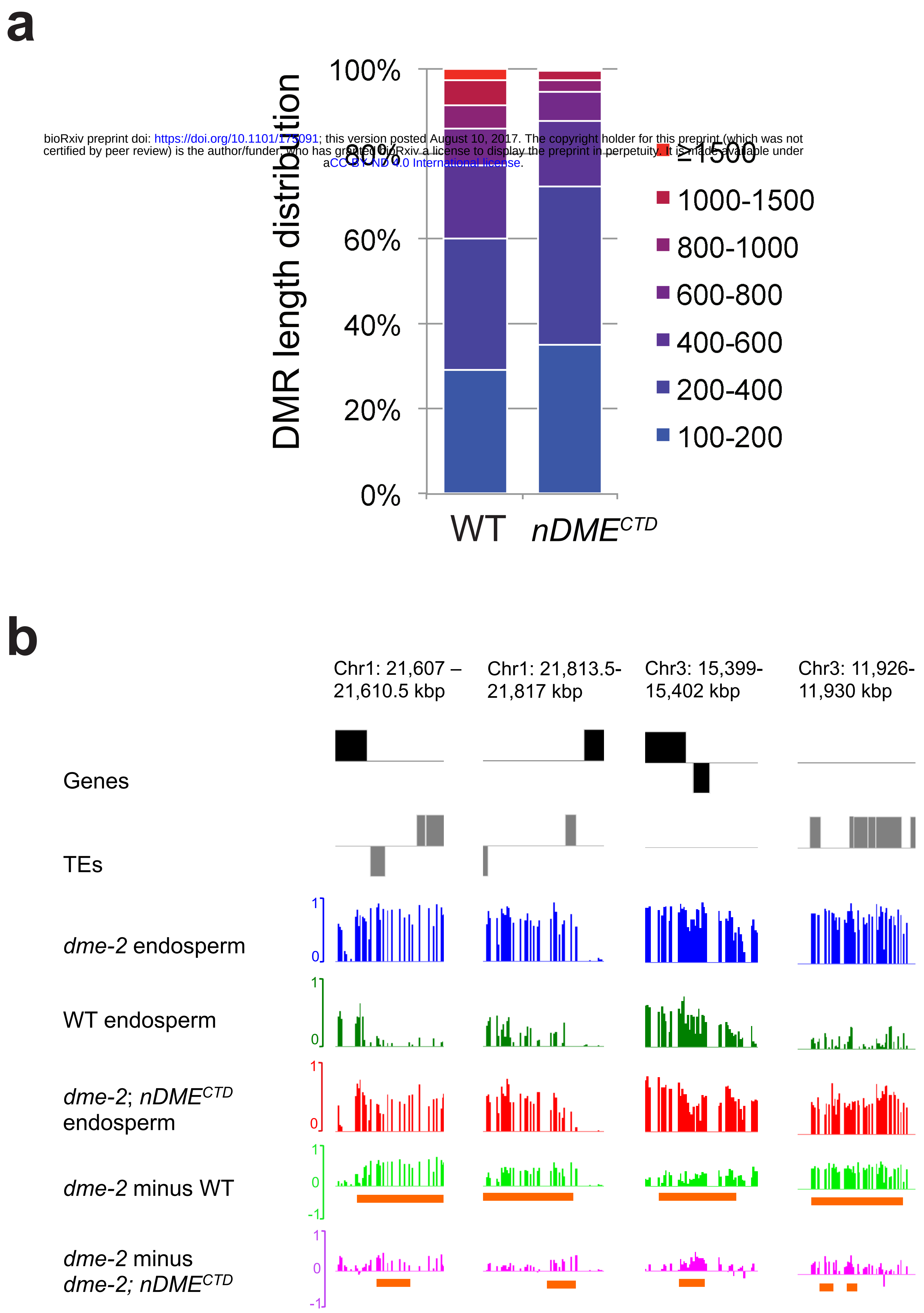
Supplementary Figure S7



Supplementary Figure S8



Supplementary Figure S9


a
DMR length distribution
WT
nDME^CTD
≥1500
b
Chr1: 21,607 – 21,610.5 kbp
Chr1: 21,813.5-21,817 kbp
Chr3: 15,399-15,402 kbp
Chr3: 11,926-11,930 kbp
Genes
TEs
dme-2 endosperm
WT endosperm
dme-2; nDME^CTD endosperm
dme-2 minus WT
dme-2 minus dme-2; nDME^CTD