A Thesis for the Degree of Master of Science

Comparative phylogenetic and biochemical analysis on plant DNA demethylases

계통분석 및 생화학 분석을 통한 식물 특이적 DNA 탈 메틸화 효소간 비교연구

FEBRUARY, 2015

HOSUB SHIN

MAJOR IN HORTICULTURAL BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF

SEOUL NATIONAL UNIVERSITY

Comparative phylogenetic and biochemical analysis on plant DNA demethylases

UNDER THE DIRECTION OF DR. JIN HOE HUH
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
SEOUL NATIONAL UNIVERSITY

BY
HOSUB SHIN

MAJOR IN HORTICULTURAL BIOTECHNOLOGY
DEPARTMENT OF PLANT SCIENCE
GRADUATE SCHOOL SEOUL NATIONAL UNIVERSITY

FEBRUARY, 2015

APPROVED AS A QUALIFIED THESIS OF HO SUB SHIN
FOR THE DEGREE OF MASTER OF SCIENCE
BY THE COMMITTEE MEMBERS

CHAIRMAN _____

Doil Choi, Ph.D.

VICE-CHAIRMAN _____

Jin Hoe Huh, Ph.D.

MEMBER _____

Byoung-Cheorl Kang, Ph.D.

Comparative phylogenetic and biochemical analysis on plant DNA demethylases

HOSUB SHIN

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

ABSTRACT

DNA methylation is one of the major epigenetic marks essential for gene regulation in many eukaryotes. Arabidopsis DEMETER (DME) and its homologs play important roles in active DNA demethylation by directly removing 5-methylcytosine from DNA. DME family genes are exclusively found in the plant lineage but their counterparts have not been reported in animals. Therefore, it is of great interest not only to infer the origin and the function of an ancestral form of DME but also to understand how it has been changed during plant evolution. I performed comparative phylogenetic analysis on 135 DME homologs identified from 42 different species including green algae, mosses and seed plants. Three domains essential for catalytic function are highly conserved among distantly related nonvascular and vascular plants, but algal homologs display limited similarity. Biochemical analysis using recombinant proteins demonstrates that DME homologs derived from mosses have canonical DNA demethylase activity but algal DME lacks such a function. These findings suggest that the DNA

demethylase function is essential even in lower plant species but algal DME may have distinct functions. This study provides evolutionary insight into the establishment of DNA demethylation system and its biological significance unique to the plant lineage.

Key Words**:** DNA demethylation, DNA demethylase, DEMETER.

Student number: 2013-21107

CONTENTS

ABSTRACT ····· i

CONTENTS ····· iii

LIST OF TABLES ····· vi

LIST OF FIGURES ····· vii

LIST OF ABBREVIATIONS ····· viii

INTRODUCTION ····· 1

LITERATURE REVIEWS ····· 5

1. DNA methylation

2. Active DNA demethylation pathway in plants and mammals

3. Plant specific DNA demethylase

MATERIAL AND METHODS ····· 9

Assessment of DME homologs from public databases

Sequence analysis and phylogenetic analysis

Plant and algae materials

Plasmid construction

Protein expression and purification

DNA substrates for labeling

5mC excision activity test

5mC excision activity test by cell extract

RESULTS ····· 19

DME family has highly conserved domains from mosses to angiosperms

Differences of DME homologs in ancient plant lineages compared to land plants

Four conserved cysteine residues in domain B are not present in basal land plants

Classification of DME homologs by the phylogenetic analysis

Sequence diversity observed in DML3 clade

Estimation of differentiation of DME, ROS1 and DML3 homologous genes

Catalytic activity of removing 5mC is conserved in all subclasses of land plants

The 5mC excision activity is not detected in DME homolog in algae

Glycosylase domain of algal DME homologs cannot excise

5mC with domain A and B of AtDME

DISCUSSION ····· 42

REFERENCES ····· 44

ABSTRACT IN KOREAN ····· 52

LIST OF TABLES

Table 1. List of DME homologs used in amino acid alignments

Table 2. List of oligos for PCR amplification and *In vitro* 5mC excision assay

Table 3. Classification of DME homologs from 33 sequenced angiosperms

LIST OF FIGURES

Figure 1. Amino acid alignments of conserved domains of DME homologs

Figure 2. Graphical representation of the sequence conseravation of DME homologs in plants

Figure 3. Amino acid alignment in domain B algae to angiosperm

Figure 4. The phylogenetic trees showing the evolutionary relationship among DNA demethylases in plant lineage

Figure 5. The 5mC excision activity of DME homologs isolated from diverse plant species

Figure 6. The 5mC excision activity of algal DME homologs

Figure 7. Base excision of activity test of OspDME

Figure 8. Base excision activity of chimeric enzyme

LIST OF ABBREVIATIONS

5caC	5-carboxylcytosine
5fC	5-formylcytosin
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
AID	Activation-induced cytidine deaminase
AP	Apurinic/apyrimidinic
APOBEC	Apolipoprotein B mRNA-editing catalytic polypeptides
BER	Base excision repair
CMT3	CHROMOMETHYLASE 3
DME	DEMETER
DML2	DEMETER-LIKE 2
DML3	DEMETER-LIKE 3
DNMT3	DNA METHYTRANSFERASE 3
DRM2	DOMAINS REARRANGED METHYLTRASNSFERASE 2
EndoIII	Endonuclease III
GPD	Glycine/proline-rich loop with a conserved aspartate
HhH	Helix-hairpin-helix
hOGG1	Human 8-oxoguanine DNA glycosylase
MEA	MEDEA
MET1	METHYLTRANSFERASE1

RdDM	RNA-directed DNA methylation
ROS1	REPRESSOR OF SILENCING 1
SMUG1	Single-strand-selective monofunctional uracil-DNA glycosylase 1
TDG	Thymine-DNA glycosylase
TET	Ten-eleven translocation methylcytosine dioxygenase 1

INTRODUCTION

Methylcytosine (5mC) was identified before recognition of DNA as genetic materials (Johnson et al., 1925). 5mC occupies approximately 2~8% in mammalian genome and much higher in plant genome, but rarely detected in budding yeasts (*Saccharomyces cerevisiae*), nematodes (*Caenorhabditis elegans*) and fruit fly (*Drosophila melanogaster*) (Doerfler, 1983). In most bacterial species, DNA methylation serves as a defense function, protecting bacteria from bacteriophage infection by removing unmethylated foreign DNA (Wilson, 1988).

One of the important mechanisms of gene regulation in eukaryotes is an epigenetic regulation by DNA methylation. DNA methylation plays roles as an epigenetic mark for gene silencing, X-chromosome inactivation and gene imprinting in higher eukaryotes (Huh et al., 2008; Kim and Zilberman, 2014). 5mC is produced by DNA methyltransferases, which transfer a methyl group from S-adenosyl methionine to C5 of cytosine. DNA methyltransferases from mammals and plants contain a conserved methyltrasnferase catalytic domain in the C-terminal region (Bestor, 2000; Chan et al., 2005; Law and Jacobsen. 2010). It is speculated that the DNA methylation system has evolved from prokaryotes (Zemach et al., 2010).

The removal of DNA methylation, called DNA demethylation, can be divided into passive and active pathways. Passive DNA demethylation occurs when DNA methyltransferases are inactive during DNA replication, whereas active DNA demethylation is induced by DNA demethylases (Gehring et al., 2009; Wu and Zhang, 2010). Active DNA demethylation is important to regulate DNA

methylation level in several cellular processes during development. In plants and mammals, active DNA demethylation pathways are distinct from each other (Zhang and Zhu, 2012).

In mammals, indirect DNA demethylation pathways are initiated with conversions of 5mC into 5-hydroxymethylcytosine (5hmC) by Ten eleven translocate (TET) family proteins or into thymine by Activation-induced cytidine deaminase/apolipoprotein B mRNA-editing catalytic polypeptides (AID/APOBEC), respectively. TET proteins convert 5hmC to 5-formylcytosine (5fC) and 5fC to 5-carboxylcytosine (5caC), and AID/APOBEC proteins convert 5hmC to 5hmU. These modified bases are excised by thymine-DNA glycosylase (TDG) or single-strand-selective monofunctional uracil-DNA glycosylase 1 (SMUG1) via the BER pathway to complete DNA demethylation (Ito et al., 2011; He et al., 2011; Williams et al., 2011, Zhang and Zhu, 2012).

On the other hand, the DEMETER (DME) family proteins that consist of DME, REPRESSOR OF SILENCING 1 (ROS1), DEMETER-LIKE 2 (DML2) and DML3 were found as 5mC-specific DNA demethylases in *Arabidopsis thilana* (Gehring el al., 2006; Morales-Ruiz et al., 2006; Ortega-Galisteo et al., 2008). These enzymes are bifunctional DNA glycosylases that exhibit not only the glycosylase activity by recognizing and removing 5mC, but also the lyase activity to cleave a phosphodiester backbone (Huh et al., 2008; Law and Jacobsen, 2010; Zhu, 2009). DME family proteins directly excise 5mC to form the abasic (apurinic/apyrimidinic, AP) site and the following base excision repair (BER) enzymes finish DNA demethylation (Gehring et al., 2009; Jang et al., 2014)

These DME family proteins have three conserved domains, including a DNA glycosylase domain and two unknown domains A and B (Mok et al., 2010; Morales-Ruiz et al., 2006; Penterman et al., 2007). The glycosylase domain of DME contains several motifs such as a helix-hairpin-helix (HhH) motif, a glycine/proline-rich loop with a conserved aspartic acid (GPD) and four cysteine residues comprising the 4F-4S cluster, all of which are essential for the glycosylase activity (Mok et al., 2010). In contrast to the glycosylase domain, the characteristics of two additional conserved domains A and B have not been revealed.

DME is dominantly expressed in the central cell before fertilization and is indispensable for endosperm development and seed viability (Choi et al., 2002; Gehring et al., 2006; Huh et al., 2008). In contrast, *ROS1*, *DML2* and *DML3* are expressed in nearly all vegetative tissues and presumably induce locus-specific DNA demethylation (Penterman et al., 2007; Qian et al., 2012; Stroud et al., 2013). However, the functional deficiency of these genes did not display any morphological defects (Penterman et al., 2007).

The establishment of plant-specific DNA demethylation system may have dramatic effects on the plant evolution. Therefore, it is of great interest not only to infer the origin and function of an ancestral form of the DME family, but also to understand how it has been changed during the plant evolution. In this study, I performed phylogenetic and biochemical analyses to study the *DME* gene family in an evolutionary context. The phylogenetic analysis was performed by comparing the DME family proteins in angiosperms, gymnosperm, lycophyte, bryophyte and green algae. And the biochemical analysis was performed to discern the

conservation of the 5mC excision activity in these family proteins. These analyses revealed that the three essential domains were derived from algal DME homologs. And the DME homolog from moss has canonical DNA demethylase activity but algal DME lacks such a function. These results provide evolutionary insight on the establishment of plants specific DNA demethylation system and its biological meaning to the plant lineage.

LITERATURE REVIEWS

1. DNA methylation

DNA methylation is the conversion of cytosine to 5-methylcytosine (5mC). DNA methylation is an abundant epigenetic modification in both plants and mammals (Law and Jacobsen, 2010) and has essential roles for genome stability, imprinting, X chromosome inactivation and tissue specific gene regulation (Bender, 2004, Bird, 2002).

To generate DNA methylation, both *de novo* and maintenance DNA methyltransferases transfer the methyl group from S-adenosyl-methionine to cytosine. These two classes of methyltransferases are found in plants and mammals and contain a conserved catalytic domain in C-terminal region (Bestor, 2000; Chan et al., 2005). DNA METHYTRANSFERASE 3 (DNMT3) and DOMAINS REARRANGED METHYLTRASNSFERASE 2 (DRM2), known as *de novo* DNA methyltransferases in plants and mammals, generate a new methylation patterns and DNMT1 and its ortholog, METHYLTRANSFERASE1 (MET1), known as maintenance DNA methyltransferases, maintain CpG methylation patterns (Chan et al., 2005; Bestor, 2000; Law and Jacobsen, 2010). In plants, there are also non-CpG methylations. CpHpG (H is A or T) methylation is maintained by CHROMOMETHYLASE 3 (CMT3) and CpHpH methylation is carried out by DRM2 mediated by small RNAs via RNA-directed DNA methylation (RdDM) (Chan et al., 2005; Law and Jacobsen, 2010; Matzke et al., 2005).

2. Active DNA demethylation pathway in plants and mammals

Passive DNA demethylation occurs when maintenance methyltransferases are inactivated during DNA replication. However active DNA demethylation is carried out independent of DNA replication. This active DNA demethylation is mediated by several enzymes, and then 5mC is excised from DNA. In this step, mammals and plants have developed different DNA demethylation mechanism (Zhang and Zhu, 2012).

In plants, the DEMETER (DME) family, which consists of DME, REPRESSOR OF SILENCING 1 (ROS1), DEMETER-LIKE 2 (DML2) and DML3, directly excises 5mC to form an abasic (apurinic/apyrimidinic, AP) site (Choi et al., 2002; Gong et al., 2002; Ortega-Galisteo et al., 2008). These enzymes are bifunctional DNA glycosylases that not only recognize and remove 5mC in double strand DNA, but also have lyase activity as cleave a phosphordiester bond (Gehring et al., 2006; Law and Jacobsen, 2010; Zhu, 2009). After the AP site is created, the base excision repair (BER) pathway enzymes complete DNA demethylation (Gehring et al., 2009). The phosphodiester bond at 3' end of the AP site is excised by AP endonuclease and single strand break with a 3'-hydroxyl group is generated. After that, the gap is filled and sealed by DNA polymerase and DNA ligase (David et al., 2007; Gehring et al., 2009).

In mammals, equivalent 5mC-specific glycosylases have not been identified. By contrast, there appear to be indirect demethylation pathways (Bhutani et al., 2011). First, 5mC could be hydroxylated into 5-hydroxymethylcytosine (5hmC) by the Ten-eleven translocation methylcytosine

dioxygenase 1 (TET) family proteins and also these proteins oxidize 5hmC to 5-formylcytosin (5fC) or 5-carboxylcytosine (5caC) (Ito et al., 2010; Ito et al., 2011). 5hmC or 5mC can be deaminated into 5-methyluracil (5mU) or 5-hydroxymethylurail (5hmU) by the Activation-induced cytidine deaminase/apolipoprotein B mRNA-editing catalytic polypeptides (AID/APOBEC) family (Bhutani et al., 2009). After these successive modifications on 5mC, the intermediates of 5mC were converted into cytosine by the Uracil-DNA glycosylase (UDG) family of the BER glycosylase like Thymine-DNA glycosylase (TDG) or Single-strand-selective monofunctional uracil-DNA glycosylase 1 (SMUG1) (Cortellino et al., 2011; Guo et al., 2011; He et al, 2011; Zhang and Zhu, 2012).

3. Plant specific DNA demethylase

Plant specific DNA demethylases in *Arabidopsis thaliana* consist of DME, ROS1, DML2 and DML3. DME possesses three conserved domains including the glycosylase domain and two unknown domains A and B (Mok et al., 2010). The glycosylase domain includes four cysteine residues forming the 4F-4S cluster, a helix-hairpin-helix (HhH) motif and a glycine/proline-rich loop with a conserved aspartate (GPD), which are present in other glycosylases such as MutY, Endonuclease III (EndoIII) in *E. coil* and human 8-oxoguanine DNA glycosylase (hOGG1). These motifs are reported to be crucial for glycosylase activity (Bruner et al., 2000; Guan et al., 1998; Kuo et al., 1992; Mok et al., 2010).

DME is essential for endosperm development and gene imprinting, which is primarily expressed in the central cell and synergids of the female gametophyte

(Choi et al., 2002; Huh et al. 2008). The biological function of DME is to induce the maternal allele-specific hypomethylation of *MEDEA* (*MEA*), which leads to its transcriptional activation and endosperm development (Choi et al., 2004; Gehring et al., 2006).

ROS1 is another member of the DME family. ROS1 was identified in screening for mutants with misregulated expression of the repetitive *RD29A-LUC* transgene. ROS1 is required for the release of transcriptional silencing of a hypermethylated transgene (Gong et al., 2002; Zhu, J.K. 2009). Also, ROS1 regulates DNA methylation level in genome. Compared to temporally and spatially static expression of DME in central cell, ROS1 is expressed in nearly plant tissues (Gong et al., 2002; Penterman et al., 2007). In *ros1-3* and *ros1-4*, locus-specific increase of DNA methylation level is detected (Penterman et al., 2007; Qian et al. 2012). DML2 and DML3 are also expressed in all organs of plants (Ortega-Galisteo et al., 2007).

MATEIRIALS AND METHODS

Assessment of DME homologs from public databases

Protein sequences of all DME homologs identified from a total of 42 plant species ranging from unicellular green algae to angiosperms were obtained from public databases such as Phytozome v9.0 (http://www.phytozome.net/), DOE Joint-Genome institute (JGI; http://genome.jgi-psf.org/) and ConGenIE (http://congenie.org/start/) (Table 1). Proteins displaying significant levels of conservation of the glycosylase domain and domains A and B were chosen for further analysis.

Sequence analysis and phylogenetic analysis

Amino acid alignment of DME homologs was performed using ClustalX v2.1 (Larkin et al., 2007) and GeneDox v2.7.00 (Nicholas and Nicholas, 1997). Sequences were derived from DME domain A, the glycosylase domain and domain B, respectively (Mok et al., 2010). The conserved protein motif search was performed using weblogo3 (Crooks et al., 2004). The phylogenetic tree was constructed by the neighbor-joining method using MEGA v6 (Tamura et al., 2013) with 2,000 bootstrap replications.

Plant and algae materials

Solanum lycopersicum var. MicroTom was used to prepare CDS of DME homologs.

The plants were grown with a 16 hr of light and 8 hr of dark cycle at 26 ^{o}C. *Ostreococcus sp* and *Ostrecoccus tauri* were received form Roscoff Culture Collection (http://roscoff-culture-collection.org/). These algae were used to prepare CDS of DME homologs. Algae were cultured with 16 hr of light and 8 hr of dark at 26 ^{o}C for two weeks in f/2 medium (Guillard et al., 1962).

Plasmid construction

Total RNA of floral buds and leaves of MicroTom was extracted with TriZol reagent (Life Technologies). RNA was treated with DNase I (Takara) and 5 μg of total RNA was subjected to first-strand cDNA synthesis using SuperScript III RTase (Life Technologies) with oligo-dT (Life Technologies) according to the manufacturer's protocol. PCR amplification was performed to obtain DME homologs (Solyc09g009080.1.1, Solyc10g083630.1.1 and Solyc11g007580.1.1) of MicroTom using designated primers (Table 2). In a similar way, DME homologs of *Ostreococcus sp* (OspDME; jgi|OstRCC809| 36503|fgenesh1_pg.C_scaffold_5000084), *Ostrecoccus tauri* (OtDME; gi|Ostta4|32774|0500010139), and OspDMEΔ were isolated (Table 2). The Solyc09g009080.1.1Δ was cloned into the *Bam* HI and *Eag* I sites of the pLM302 vector (provided by Prof. Brandt, Vanderbilt University). The Solyc10g083630.1.1Δ was cloned into the *Bam* HI and *Xho* I sites of the pLM302 vector. The Solyc11g007580.1.1Δ was cloned into the *Bam* HI and *Sal* I sites of the pLM302 vector, The OspDME was cloned into the *Bam* HI and *Eco* RI sites of the pLM302 vector. The OtDME was cloned into the *Bam* HI and *Hind* III sites of the

pLM302 vector. The OspDMEΔ was cloned into the *Bam* HI and *Hind* III sites of the pLM302 vector. Chimeric proteins between AtDME and algal DME homologs were produced by physical combination of the domain A and domain B from modified AtDME (Jang et al., unpublished) and glycosylase domain of OspDME using the *Hind* III and *Not* I sites (Table 2). The construct of PpDMEΔ (Pp1s14_101v6.1) was cloned into the *Eco* RI and *Sal* I sites of the pLM302 vector (provided by prof. Ohad, Tel Aviv University). And also, the sequence of DME homolog in *Oryza sativa* (OsDME; Loc_os02g29238) was cloned into the *Bam* HI and *Sal* I sites of the pLM302 vector.

Protein expression and purification

The plasmids for *in vitro* 5mC excision activity test were transformed into *E. coli* Rosetta2 (DE3) strain (Merck Millipore) and cells were grown at 28 °C in 2 liters of LB media containing 50 μg/ml of kanamycin and 50 μg/ml of chloramphenicol until the OD_{600} value reached at 0.4. Expression of protein was induced by adding 100 μM of IPTG and incubated at 28 °C for 3 hr. Cells were harvested by centrifugation at 6,000 rpm for 15 min at 4 °C and resuspended into 30 ml of lysis buffer (50 mM Tris-HCl, pH 7.4, 100 mM NaCl, 10% glycerol, 0.1 mM DTT). The 0.1 mM PMSF was added to prevent protease activity. Cells were sonicated for 5 min on ice. Cell extracts were centrifuged at 12,000 rpm for 25 min at 4 °C. The supernatant was filtered through nylon membrane with 0.45 μm pore (Advantec) and loaded onto Histrap FF column (GE Healthcare). Eluted fractions were

collected and loaded onto Heparin HP column (GE Healthcare). Finally gel filtration was performed by Superdex 200-pg column (GE Healthcare). Eluted fractions were concentrated and stored in a storage buffer (20 mM Tris-HCl, pH 7.4, 40 mM NaCl, 4% glycerol, 0.1 mM DTT) at -80 ^{o}C.

DNA substrates for labeling

The 35-mer oligonucleotides containing cytosine and 5-methylcytosine were purchased by Midland Certified (TX) and thymine was purchased by Integrated DNA technologies (IDT) (Table 2). Forty pmol of oligonucleotides were labeled by T4 polynucleotide kinase (Takara) with 30 μCi of [γ-^{32}P] ATP. Labeled oligonucleotides were purified by Qiaquick Nucleotide Removal Kit (Qiagen) and annealed with complementary oligonucleotides. The mixture was boiled at 95 ^{o}C for 10 min and slowly cooled down to room temperature overnight for annealing.

5mC excision activity test

Twenty five nM oligonucleotides were incubated with 200 ng of purified DME homolog proteins in the glycosylase reaction buffer (10 mM Tris-HCl, pH 7.4, 50 mM NaCl, 0.5 mM DTT) at 37 ^{o}C for 1 hr. Reaction was stopped by adding an equal volume of stop solution (98% formamide, 10 mM EDTA, 0.2% Xylen cyanol FF, 0.2% of bromophenol blue) and reaction mixture was boiled at 95 ^{o}C for 10 min. Reaction products were separated on a 15% denaturing polyacrylamide gel containing 7.5 M urea and 1× TBE (Tris/ Boric acid/ EDTA) for 3 hr in 1200 V.

5mC excision activity test by cell extract

The plasmids for *in vitro* 5mC excision activity test were transformed into *E. coli* Rosetta2 (DE3) strain (Merck Millipore) and cells were grown at 28 °C in 100 ml of LB containing 50 μg/ml of kanamycin and chloramphenicol until the OD_{600} value reached at 0.4. Expression of protein was induced with 100 μM of IPTG at 28 °C for 3 hr. Cells were harvested by centrifugation at 6,000 rpm for 15 min at 4 °C and resuspended in 10 ml of lysis buffer (50 mM Tris-HCl, pH 7.4, 100 mM NaCl, 10% glycerol, 0.1 mM DTT). 0.1 mM PMSF was added to prevent protease activity and cells were sonicated for 1 min on ice. After that 25 nM oligonucleotides were incubated with 2 μl of cell extract in the glycosylase reaction buffer (10 mM Tris-HCl, pH 7.4, 50 mM NaCl, 0.5 mM DTT) and 2 mM EDTA at 37 °C for 1 hr. Terminating reaction and separation of samples were performed as described above.

Table 1. List of DME homologs used in amino acid alignments.

Protein ID*	Length (aa)	Organism Name
Aquca_009_00981.1	2169	*Aquilegia coerulea*
341139	1079	*Arabidopsis lyrata*
353424	1073	*Arabidopsis lyrata*
482629	1433	*Arabidopsis lyrata*
487239	1998	*Arabidopsis lyrata*
902767	1620	*Arabidopsis lyrata*
AT2G36490.1	1394	*Arabidopsis thaliana*
AT3G10010.1	1333	*Arabidopsis thaliana*
AT4G34060.1	1045	*Arabidopsis thaliana*
AT5G04560.2	1988	*Arabidopsis thaliana*
Bradi2g23797.1	1768	*Brachypodium distachyon*
Bradi3g43692.1	1916	*Brachypodium distachyon*
Bradi3g43720.1	993	*Brachypodium distachyon*
Bradi4g08870.1	1975	*Brachypodium distachyon*
Bradi4g16620.1	1681	*Brachypodium distachyon*
Brara.A00413.1.p	1151	*Brassica rapa*
Brara.E00835.1.p	1401	*Brassica rapa*
Brara.J02727.1.p	1927	*Brassica rapa*
Bostr.13129s0501.1.p	2025	*Boechera stricta*
Bostr.22252s0064.1.p	1379	*Boechera stricta*
Bostr.23794s0441.1.p	1221	*Boechera stricta*
Bostr.23794s0442.1.p	1344	*Boechera stricta*
Bostr.7867s1258.1.p	952	*Boechera stricta*
Ciclev10010892m	1965	*Citrus clementina*
Ciclev10027676m	1763	*Citrus clementina*
Ciclev10030474m	2030	*Citrus clementina*
Cagra.0228s0039.1.p	1580	*Capsella grandiflora*
Cagra.0228s0040.1.p	1484	*Capsella grandiflora*
Cagra.1305s0031.1.p	926	*Capsella grandiflora*
Cagra.4686s0012.1.p	1291	*Capsella grandiflora*
Cagra.9217s0036.1.p	2001	*Capsella grandiflora*
evm.model.supercontig_128.52	1890	*Carica papaya*
Carubv10000024m	2001	*Capsella rubella*
Carubv10007707m	1063	*Capsella rubella*
Carubv10012825m	1294	*Capsella rubella*
Carubv10024677m	1484	*Capsella rubella*

(Countinued)

Table 1. (Continued)

Protein ID*	Length (aa)	Organism Name
Carubv10025379m	1561	*Capsella rubella*
Cucsa.083110.1	1353	*Cucumis sativus*
Cucsa.308950.1	1845	*Cucumis sativus*
Cucsa.378580.1	1777	*Cucumis sativus*
Cucsa.385370.1	1098	*Cucumis sativus*
orange1.1g000414m	1542	*Citrus sinensis*
orange1.1g002141m	961	*Citrus sinensis*
orange1.1g002245m	948	*Citrus sinensis*
62092	1569	*Coccomyxa subellipsoidea*
Eucgr.A01547.1	1852	*Eucalyptus grandis*
Eucgr.J01016.1	1885	*Eucalyptus grandis*
Thhalv10012472m	1229	*Eutrema salsugineum*
Thhalv10016139m	1469	*Eutrema salsugineum*
Thhalv10019961m	1058	*Eutrema salsugineum*
Thhalv10026936m	1022	*Eutrema salsugineum*
mrna01635.1-v1.0-hybrid	1937	*Fragaria vesca*
mrna30143.1-v1.0-hybrid	1766	*Fragaria vesca*
mrna30462.1-v1.0-hybrid	1647	*Fragaria vesca*
Glyma.03G190800.1.p	1742	*Glycine max*
Glyma.10G065900.1.p	2015	*Glycine max*
Glyma.10G202200.1.p	1430	*Glycine max*
Glyma.13G151000.1.p	1994	*Glycine max*
Glyma.20G188300.1.p	1851	*Glycine max*
Gorai.002G145100.1	1936	*Gossypium raimondii*
Gorai.006G158300.1	1959	*Gossypium raimondii*
Gorai.009G088400.1	1686	*Gossypium raimondii*
Gorai.009G408300.1	1894	*Gossypium raimondii*
Lus10014367	1813	*Linum usitatissimum*
Lus10023864	1755	*Linum usitatissimum*
Lus10027713	1598	*Linum usitatissimum*
Lus10035576	2280	*Linum usitatissimum*
MDP0000207280	1778	*Malus domestica*
MDP0000233406	1960	*Malus domestica*
MDP0000319132	1779	*Malus domestica*
cassava4.1_000084m	1872	*Manihot esculenta*
cassava4.1_000129m	1721	*Manihot esculenta*

(Countinued)

Table 1. (Continued)

Protein ID*	Length (aa)	Organism Name
cassava4.1_000140m	1688	*Manihot esculenta*
cassava4.1_024541m	1471	*Manihot esculenta*
cassava4.1_024834m	1245	*Manihot esculenta*
cassava4.1_031075m	1900	*Manihot esculenta*
Migut.D02337.1.p	1348	*Mimulus guttatus*
Migut.F00888.1.p	1884	*Mimulus guttatus*
Migut.J00987.1.p	1887	*Mimulus guttatus*
Migut.L00609.1.p	896	*Mimulus guttatus*
56174	2193	*Micromonas sp. RCC299*
Medtr1g061220.1	2842	*Medicago truncatula*
Medtr7g103680.1	1779	*Medicago truncatula*
LOC_Os01g11900.1	1953	*Oryza sativa*
LOC_Os02g29230.1	1637	*Oryza sativa*
LOC_Os02g29380.1	1208	*Oryza sativa*
LOC_Os04g28860.1	961	*Oryza sativa*
LOC_Os05g37350.1	1848	*Oryza sativa*
LOC_Os05g37410.1	1830	*Oryza sativa*
36503	1070	*Ostreococcus sp. RCC809*
32774	1063	*Ostreococcus tauri*
Pp1s14_101V6.1	1933	*Physcomitrella patens*
Pp1s480_4V6.1	1849	*Physcomitrella patens*
Pp1s226_20V6.2	1814	*Physcomitrella patens*
ppa000163m	1557	*Prunus persica*
ppa000207m	1470	*Prunus persica*
ppa020575m	1747	*Prunus persica*
Potri.006G116000.1	1664	*Populus trichocarpa*
Potri.008G025900.1	1790	*Populus trichocarpa*
Potri.010G234400.1	1868	*Populus trichocarpa*
Pavir.Da00998.1.p	1963	*Panicum virgatum*
Pavir.Ea00774.1.p	1914	*Panicum virgatum*
Pavir.Eb00820.1.p	1963	*Panicum virgatum*
Pavir.Ib03800.1.p	634	*Panicum virgatum*
Pavir.J00215.1.p	1797	*Panicum virgatum*
Pavir.J00370.1.p	1741	*Panicum virgatum*
Pavir.J05638.1.p	958	*Panicum virgatum*
Pavir.J10064.1.p	749	*Panicum virgatum*

(Countinued)

Table 1. (Continued)

Protein ID*	Length (aa)	Organism Name
Phvul.001G186500.1	1690	*Phaseolus vulgaris*
Phvul.007G223600.1	2210	*Phaseolus vulgaris*
29092.m000452	1634	*Ricinus communis*
29428.m000327	1876	*Ricinus communis*
29991.m000647	1712	*Ricinus communis*
Sobic.001G188500.1.p	844	*Sorghum bicolor*
Sobic.004G149800.1.p	1915	*Sorghum bicolor*
Sobic.006G224100.1.p	1041	*Sorghum bicolor*
Sobic.008G085300.1.p	1874	*Sorghum bicolor*
Sobic.009G155900.1.p	1790	*Sorghum bicolor*
Si016069m	1955	*Setaria italica*
Si016216m	987	*Setaria italica*
Si020961m	1605	*Setaria italica*
Si028642m	1930	*Setaria italica*
Si038691m	602	*Setaria italica*
Solyc03g123440.2.1	1539	*Solanum lycopersicum*
Solyc09g009080.2.1	1703	*Solanum lycopersicum*
Solyc10g083630.1.1	1825	*Solanum lycopersicum*
Solyc11g007580.1.1	1870	*Solanum lycopersicum*
1110044	1717	*Selaginella moellendorffii*
750020	1063	*Selaginella moellendorffii*
PGSC0003DMP400004883	1835	*Solanum tuberosum*
PGSC0003DMP400016173	1361	*Solanum tuberosum*
PGSC0003DMP400060928	880	*Solanum tuberosum*
Thecc1EG022433t1	1923	*Theobroma cacao*
Thecc1EG043681t2	2000	*Theobroma cacao*
GSVIVT01031400001	907	*Vitis vinifera*
GSVIVT01033777001	1471	*Vitis vinifera*
GSVIVT01034713001	1622	*Vitis vinifera*
GRMZM2G123587_P01	1907	*Zea mays*
GRMZM2G131756_P01	910	*Zea mays*
GRMZM2G422464_P02	1905	*Zea mays*
GRMZM5G828460_P01	651	*Zea mays*

* Protein IDs are used from pytozome DB identifier

Table 2. List of oligos for PCR amplification and *in vitro* 5mC excision assay

Name	Sequence (5' → 3')	Usage
solyc09g009080.1.1_F	AATTGGATCCTATCAAAGAGATGGAAGTATTGTCCCT	RT-PCR for cloning
solyc09g009080.1.1_R	AATTCGGCCGCTAGTTTTCATCTGGCTTTCCTTTT	RT-PCR for cloning
solyc10g083630.1.1_F	AATTGGATCCTATCAAAGAGATGGAAGTATTGTCCC	RT-PCR for cloning
solyc10g083630.1.1_R	AATTGTCGACTCAGGAGGCTACTCCTTTGTCTTC	RT-PCR for cloning
solyc11g007580.1.1_F	AATTGGATCCTACAAGGGCAGTGGCACAAT	RT-PCR for cloning
solyc11g007580.1.1_R	AATTCTCGAGCTACTTTGTATGGACAGATATTGGTGA	RT-PCR for cloning
Ost_sp_F	AATTGGATCCATGACGACGGAGGTGCGAGAG	RT-PCR for cloning
Ost_sp_F	AATTGAATTCTCACGCCACAGACGTCATTGAG	RT-PCR for cloning
Ost_tauri_F	AATTGGATCCCCAATTCGCGCGCTTTGG	RT-PCR for cloning
Ost_tauri_R	AATTAAGCTTTCATGCAATCATTGGTGCCGT	RT-PCR for cloning
G_domain_Ost_sp_F	AATTAAGCTTCCGTTCGACACACCCGTTCGGC	PCR for cloning
G_domain_Ost_sp_R	AATTGCGGCCGCATGGTTGTAAAGAAGCCAGCCCCGAGTG	PCR for cloning
MEDEA_pro_35mer_C_F	CTATACCTCCTCAACTCCGGTCACCGTCTCCGGCG	5mC excision activity test
MEDEA_pro_35mer_5mC_F	CTATACCTCCTCAACTCCGGTCACCGTCTCCGGCG (C = 5mC)	5mC excision activity test
MEDEA_pro_35mer_T_F	CTATACCTCCTCAACTCTGGTCACCGTCTCCGGCG	5mC excision activity test
MEDEA_pro_35mer_G_R	CGCCGGAGACGGTGACCGGAGTTGAGGAGGTATAG	5mC excision activity test

RESULTS

DME family has highly conserved domains from mosses to angiosperms.

To understand how plant-specific DNA demethylases have been changed and established during the plant evolution, the alignment analysis was performed with amino acid sequences of DME homologs (Figure 1). Sequences of DME homologs collected from public databases were mispredicted fairly. Thus, sequences with relatively a small coverage to full-length DME are excluded. Finally, a total of 141 DME homologs were obtained from 42 plant species including angiosperms, moss and green algae (Table 1). The amino acid alignment was performed with clustalX2 with parameters of BLOSUM 30 matrix, gap opening penalty of 5, and gap extension penalty of 0.05. From the previous study, DME homologs in *Arabidopsis thaliana* have three conserved domains essential for the 5mC excision activity (Mok et al., 2010). Especially, the HhH motif, the GPD and four cysteine residues for 4F-4S cluster in the glycosylase domain were reported as essential (Guan et al., 1998; Kuo et al., 1992). The HhH glycosylase domain is the conserved element of the HhH superfamily (Scharer et al., 2010) and it was reported that two catalytically important residues (K1286 and D1304) in the HhH motif are necessary for the bifunctional glycosylase/lyase activity of DME (Gehring et al., 2006). Also, the 4F-4S cluster and the GPD found in human 8-oxoguanine DNA glycosylase and *E. coli* adenine DNA glycosylase are known to have critical roles for the 5mC excision (Guan et al., 1998; Kuo et al., 1992). The alignment of 141 DME homologs shows that these three conserved domains are universal in various plant species and also, the HhH motif, the GPD and four cysteine residues for 4F-4S cluster are highly conserved. A number of amino acid

sequences of domain A and domain B are also highly conserved (Figures 1, 2), suggesting the importance of these regions for the 5mC excision activity. In the sum of alignment results, all DME homologs in land plants have approximately 80.8% of overall sequence similarity. The sequence similarity of the glycosylase domain (~85.1%) is higher than those of domain A (~77.8%) and domain B (~79.5%) (Figure 2). These data present that the three domains of DME are highly conserved throughout land plants, suggesting the necessity of DME family genes over plant lineage evolution**.**

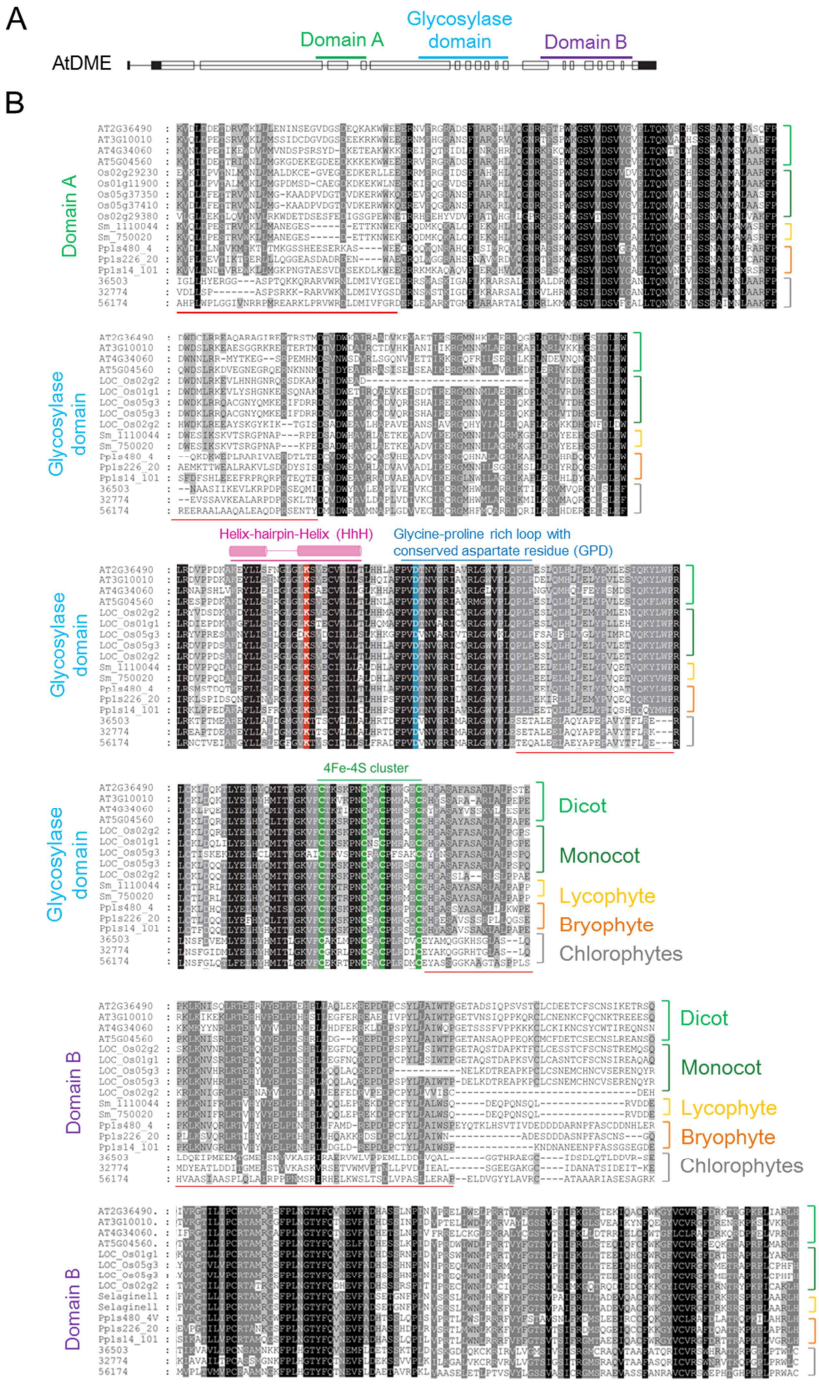

A
Domain A
Glycosylase domain
Domain B
AtDME
B
Domain A
Glycosylase domain
Helix-hairpin-Helix (HhH)
Glycine-proline rich loop with conserved aspartate residue (GPD)
4Fe-4S cluster
Dicot
Monocot
Lycophyte
Bryophyte
Chlorophytes
Domain B

Figure 1. Amino acid alignments of three conserved domains of DME homologs. (A) Schematic view of *Arabidopsis thaliana DME* with three essential domains for catalytic function (domain A, Glycosylase domain and domain B; Mok et al, 2010) (B) Amino acid alignments of 17 DME homologs in *A. thaliana*, *Oryza sativa*, *Selaginella moellendorffii*, *Physcomitrella patens*, *Ostreococcus sp.* and *Micromonas sp.*. The HhH, GPD, and catalytic (K in red and D in blue) and cysteine residues (green) for Fe-S cluster assembly are denoted in boxes (Mok et al. 2010). Dissimilar regions between algae and land plants are denoted by red bars.

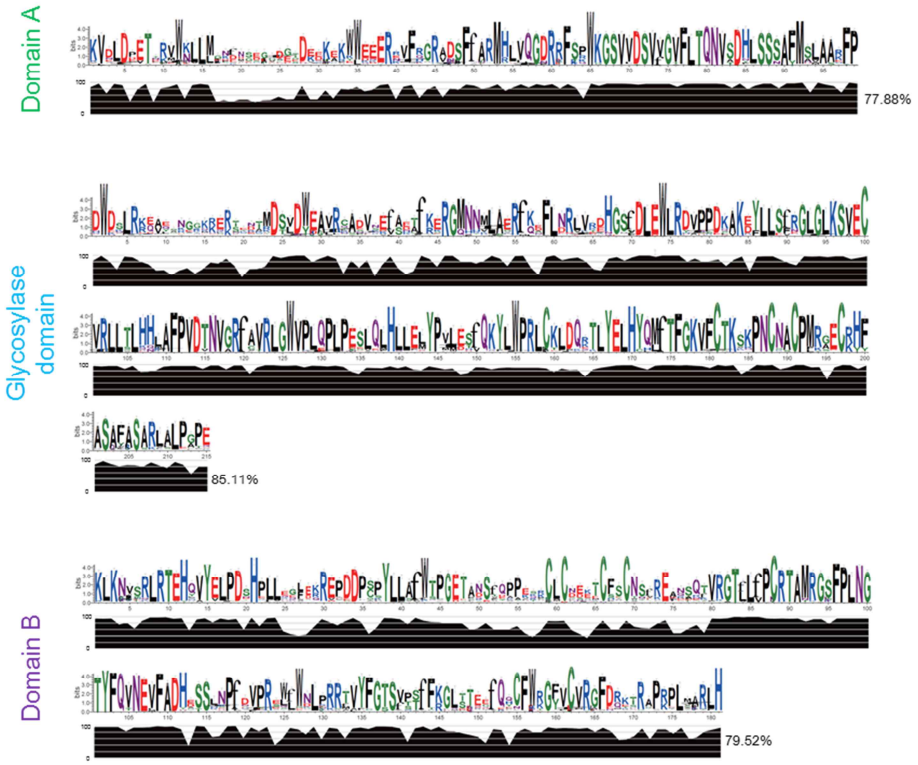


Figure 2. Graphical representation of the sequence conservation of DME homologs in plants. The sequence logos were generated by amino acid alignments of individual domains from 141 DME homologs using WebLogo3 (Crooks et al., 2004). The height of symbols within the stack represents the relative frequency of each amino acid at the position. The overall height of the stack under the logo shows the degree of conservation. The average degree of conservation of each domain is represented to the right. Color scheme of amino acid is divided according to its chemical properties (black, hydrophobic; green, polar; purple, neutral; blue, basic; red, acidic).

Differences of DME homologs in ancient plant lineages compared to land plants

Viridiplantae (green plants) including land plants as well as green algae have descended from a single eukaryotic ancestor that had acquired photosynthetic cyanobacterium. In green algae such as *Ostreococcus sp.*, *Micromonas sp.* and *Coccomyxa sp.,* only a single copy of DME homolog was found. These algal DME homologs have also three conserved domains like other DME family proteins, but less conserved compared to other DME homologs from land plants including *Physcometrella patens* and *Selaginella moellendorffii* (Figure 1). However, aforementioned HhH, GPD and 4Fe-4S cluster in the glycosylase domain could be detected in all algal DME homologs (Figure 1). Taken together, the alignment data shows that three functional domains of DME family proteins would be derived from algae.

Four conserved cysteine residues in domain B are not present in basal land plants.

In DME family proteins in *A. thaliana*, the conservation of four cysteine residues in domain B called the perm-CXXC motif (Cys-X_8-Cys-X_4-Cys-X_2-Cys) was identified (Figure 3). The perm-CXXC motif is a permutated version of a single unit of the zinc finger-CXXC motif and is homologous to the structural-zinc binding motif (Iyer et al., 2011). However, the function of this motif is largely unknown. Interestingly, in DME homologs in *P. patens*, *S. moellendorffii* and algae,

these four cysteine residues in domain B are not conserved, despite the existence of the perm-CXXC motif in the other DME homologs (Figure 3). In order to estimate when the perm-CXXC motif has emerged, the DME homologous sequences were obtained from the spruce genome database (Nystedt et al., 2013). The spruce genome has DME homologs with two, three or four conserved cysteine residues (Figure 3). This implies that the perm-CXXC motif has been acquired during evolution of spermatophytes from bryophytes. However, the face that monocots and asterid dicots have both DME homologs with or without the conserved perm-CXXC motif suggests that the motif might have gene through spontaneous mutations and degenerated probably because it is dispensable for biological functions.

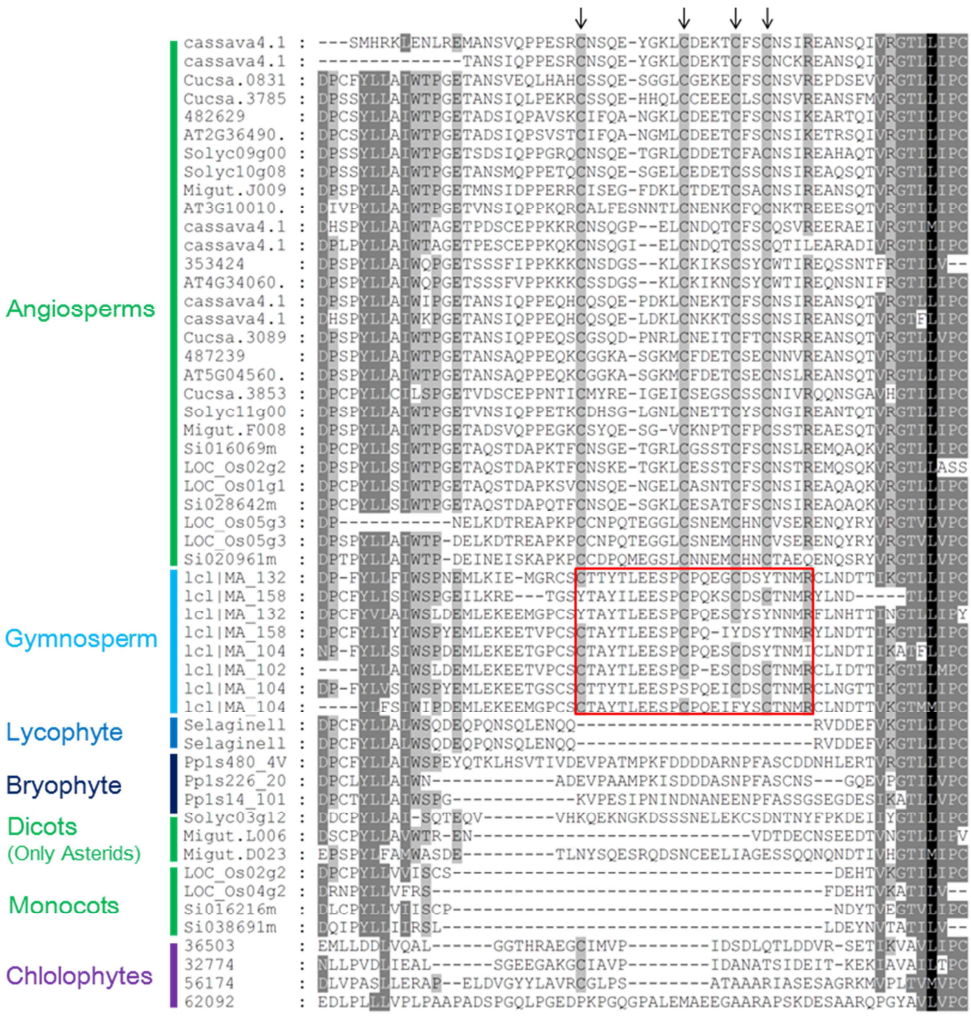


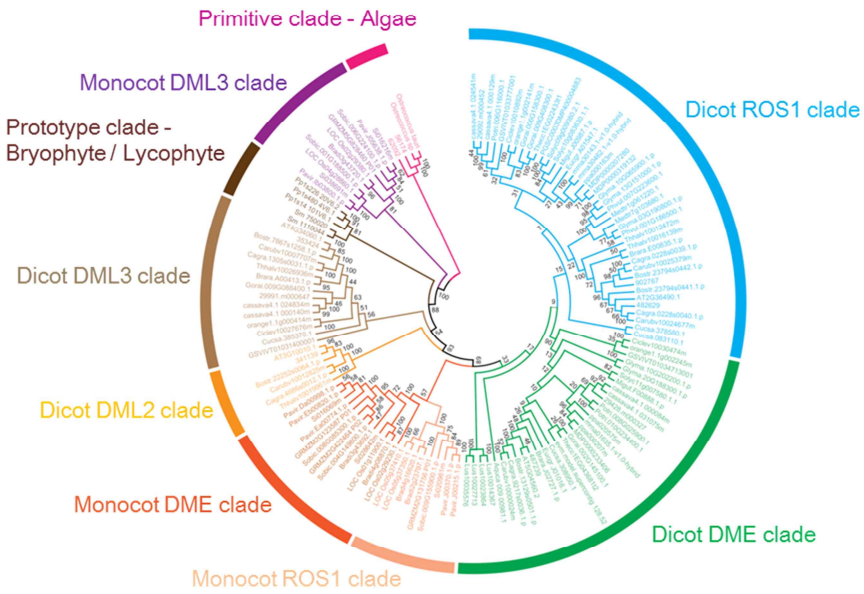
Figure 3. Amino acid alignment in domain B from algae to angiosperm. Arrowheads indicate the conserved cysteine residues. Two, three or four conserved cysteine residues were conserved in gymnosperm DME homologs (red box).

Classification of DME homologs by the phylogenetic analysis

Although DME homologs in plants are highly conserved, it is reported that enzyme activities and biological functions among DME family proteins of *A. thaliana* are different (Gehring et al., 2006; Gong et al., 2002; Huh et al. 2008; Orte-Galisteo et al., 2007). Therefore, sequence differences between DME homologs may reflect different functions. It suggests that the DME family proteins are necessary for plants to survive and the divergence of DME family proteins might result during the evolution. Therefore, it is interesting to speculate how gene duplication and subfunctionalization DME family genes have differentiated. To classify DME homologs into subclasses, the alignment data were used to construct the phylogenetic tree (Figure 4). The phylogenetic analysis was performed by neighbor-joining method on a set of 135 DME homologs excluding six inaccurate sequences. Since these sequences are derived from dicots, monocots, lycophyte, bryophyte, and algae, the phylogenetic tree will reflect the evolution process of plant-specific DNA demethylases involving gene duplication and subsequent functional divergence. Nine major subclasses are identified in the phylogenetic tree, including primitive algal, prototype bryophyte/lycophyte, monocot DML3, dicot DML3, dicot DML2, monocot ROS1, dicot ROS1, dicot ROS1, monocot DME, and dicot DME clades (Figure 4). For easier classification, all clades were named after AtDME homologs based upon the homology to each member. The monocot DME clade includes previously characterized OsROS1a (Loc_os01g11900.1). The monocot DML3 clade was named because of its dissimilarity compared to other DME clades like the AtDML3 clade. The clade including bryophyte *P. patens* and lycophyte *S. moellendorffii* is named as the prototype clade. Because these basal

land plants have homologs with a high similarity to other land plant members, unlike algae. The primitive clade includes DME homologs from non-land plant species such as *Ostreococcus sp RCC809*, *Ostreococcus tauri*, *Micromonas sp. RCC209* and *Coccomyxa subellipsoidea*. This phylogenetic tree displays dissimilarities among DME homologs (Figure 4B). Therefore, this classification reveals the relationship among divergent DME family proteins.

A


Primitive clade - Algae
Monocot DML3 clade
Prototype clade -
Bryophyte / Lycophyte
Dicot DML3 clade
Dicot DML2 clade
Monocot DME clade
Monocot ROS1 clade
Dicot ROS1 clade
Dicot DME clade

B

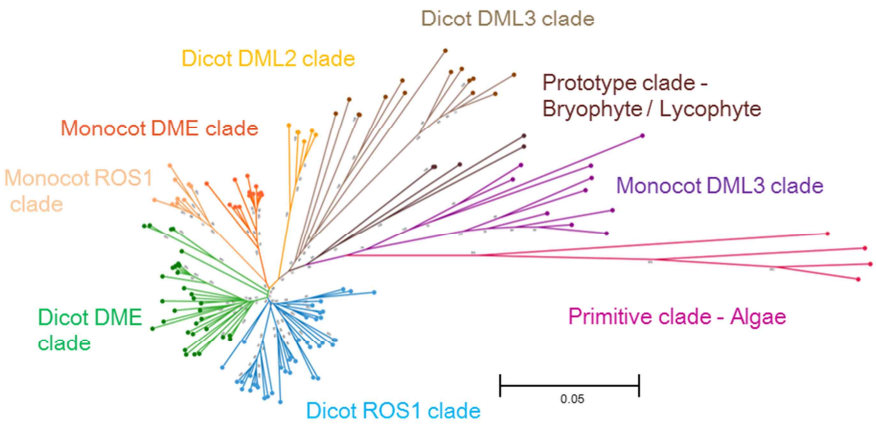

Dicot DML3 clade
Dicot DML2 clade
Prototype clade -
Bryophyte / Lycophyte
Monocot DME clade
Monocot ROS1
clade
Monocot DML3 clade
Dicot DME
clade
Primitive clade - Algae
Dicot ROS1 clade
0.05

Figure 4. The phylogenetic trees showing the evolutionary relationship among DNA demethylases in plant lineage. (A) The phylogenetic tree was constructed by the neighbor joining method using 135 DME homologs derived from 42 plants and algae. Numbers at the node indicate the statistical support obtained from 2000 bootstrap replicates. Plant DNA demethylases are classified into 9 subgroups – primitive, prototype, monocot DML3, dicot DML3, dicot DML2, monocot ROS1, dicot ROS1, monocot DME and dicot DME clades. (B) The phylogenetic tree with a genetic distance was represented. The bar indicates amino acid substitution rates per site.

Sequence diversity observed in DML3 clade

In the phylogenetic tree, the DME and ROS1 clades are most populated. They account for approximately 67.9% of whole homologs and also the genetic distance between the DME and ROS1 clades is highly close compared to the DML3 clade (Figure4, Table 3). This means that the DML3 clade has a higher rate of amino acid substitution rates than DME or ROS1 clades, and these sequence variations on the DML3 clade have been more emerged. Thus it suggests that homologous genes in DME and ROS1 clades would be more pressured by the natural selection compared with homologs in DML3 clade.

Table 3. Classification of DME homologs from 33 sequenced angiosperms.

DME	ROS1	DML2	DML3	Species name	Family	Clade
1	2	1	1	*Arabidopsis lyrata*	Brassicaceae	Eudicots
1	1	1	1	*Arabidopsis thaliana*	Brassicaceae	Eudicots
1	2	1	1	*Capsella grandiflora*	Brassicaceae	Eudicots
1	2	1	1	*Capsella rubella*	Brassicaceae	Eudicots
1	2	1	1	*Boechera stricta*	Brassicaceae	Eudicots
0	2	1	1	*Eutrema salsugineum*	Brassicaceae	Eudicots
1	1	0	1	*Brassica rapa*	Brassicaceae	Eudicots
1	2	0	1	*Cucumis sativus*	Cucurbitaceae	Eudicots
2	3	0	0	*Glycine max*	Fabaceae	Eudicots
0	2	0	0	*Phaseolus vulgaris*	Fabaceae	Eudicots
0	2	0	1	*Medicago truncatula*	Fabaceae	Eudicots
2	2	0	2	*Manihot esculenta*	Euphorbiaceae	Eudicots
1	1	0	1	*Ricinus communis*	Euphorbiaceae	Eudicots
1	1	0	1	*Linum usitatissimum*	Linaceae	Eudicots
2	1	0	1	*Populus trichocarpa*	Salicaceae	Eudicots
1	2	0	1	*Gossypium raimondii*	Malvaceae	Eudicots
1	1	0	0	*Theobroma cacao*	Malvaceae	Eudicots
1	1	0	1	*Eucalyptus grandis*	Myrtaceae	Eudicots
1	2	0	0	*Malus domestica*	Rosaceae	Eudicots
2	1	0	1	*Prunus persica*	Rosaceae	Eudicots
1	2	0	2	*Fragaria vesca*	Rosaceae	Eudicots
1	1	0	1	*Citrus clementina*	Rutaceae	Eudicots
1	1	0	1	*Citrus sinensis*	Rutaceae	Eudicots
1	1	0	1	*Vitis vinifera*	Vitaceae	Eudicots
1	1	0	2	*Mimulus guttatus*	Scrophulariaceae	Eudicots
0	1	0	2	*Solanum tuberosum*	Solanaceae	Eudicots
1	2	0	1	*Solanum lycopersicum*	Solanaceae	Eudicots
2	1	0	2	*Sorghum bicolor*	Poaceae	Monocots
2	1	0	1	*Zea mays*	Poaceae	Monocots
3	2	0	2	*Panicum virgatum*	Poaceae	Monocots
2	1	0	2	*Setaria italica*	Poaceae	Monocots
2	2	0	1	*Brachypodium distachyon*	Poaceae	Monocots
2	2	0	2	*Oryza sativa*	Poaceae	Monocots

Estimation of differentiation of DME, ROS1 and DML3 homologous genes

It is evident that expansion of the DME family genes has occurred after the colonization of early bryophytes ~450 million years ago (MYA) (Figure 1) (Rensing et al., 2008). This expansion event may be critical to control gene expression by locus-specific DNA demethylation in plants under new environments (Orte-Galisteo et al., 2007; Penterman et al., 2007). The DME, ROS1 and DML3 clades are predominantly occupied in the phylogenetic tree (Figure 4). In the view of typical plant lineages, it is reasonable to speculate that DME homologs in these clades might have been derived from a prototype of basal land plants. The DML3 group in monocots and dicots are distantly related to the DME and ROS1 clades (Figure 4). It suggests that the DML3 group has diverged both DME and ROS1 clades groups exist in monocots and dicots (Figure 4). These topologies suggest that the DME clade and the ROS1 clade in dicots and monocots are diverged after the speciation of dicots and monocots.

Interestingly, the DML2 clade genes are present in Brassicaceae (Figure 4, Table 3). This infers that the DML2 clade genes might have diverged recently in dicots with unknown functions specific to the Brassicaceae family.

Catalytic activity of removing 5mC is conserved in all subclasses of land plants

The phylogenetic tree and sequence alignments (Figures 1, 4) clearly show that highly conserved residues in three functional domains are present in most DME homologs from nine major subclasses. However, it is needs to be verified whether those homologs have an actual catalytic activity to remove 5mC. Thus, I carried out

the enzyme activity test of 5mC excision by several DME homologs from each subclass, to speculate when the plant-specific demthylation system has been established. According to recent studies on plant specific DNA demethylases, DME homologs in the monocot DME and ROS1 clades (LOC_Os01g11900; ROS1a), (LOC_Os05g37350; DNG701) and AtDME family proteins (AtDME, AtROS1, AtDML2 and AtDML3) have the 5mC excision activity (Gehring et al., 2006; Gong et al., 2002; Ortega-Galisteo et al., 2007; Ono et al., 2012; La et al., 2011). I chose DME homologs from *Solanum lycopersicum* (Solyc09g0090801.1, Solyc10g083630.1.1 and Solyc11g007580.1.1), *Oryza sativa* (Loc_os02g29238), *P. patens* (Pp1s14 101v6.1) and *Ostreococcus tauri* (gi|Ostta4|32774|0500010139), for *in vitro* 5mC excision activity test. Reactions were performed using N-terminus truncated enzymes except the full-length algal DME (OtDME) and each enzyme was incubated with 35-mer oligonucleotides including 5mC. (Figure 5B). As shown in Figure 5A, β- and δ-elimination products were produced by all land plants' DME homologs, whereas the 5mC excision activity was not observed in OtDME (Figure 5A). These results indicate that homologous proteins in all angiosperm clades have the canonical 5mC excision activity. Interestingly, PpDME from moss also showed the 5mC excision activity like other DME homologs from angiosperms. It suggests that the 5mC excision activity is conserved from basal land plants to angiosperms.

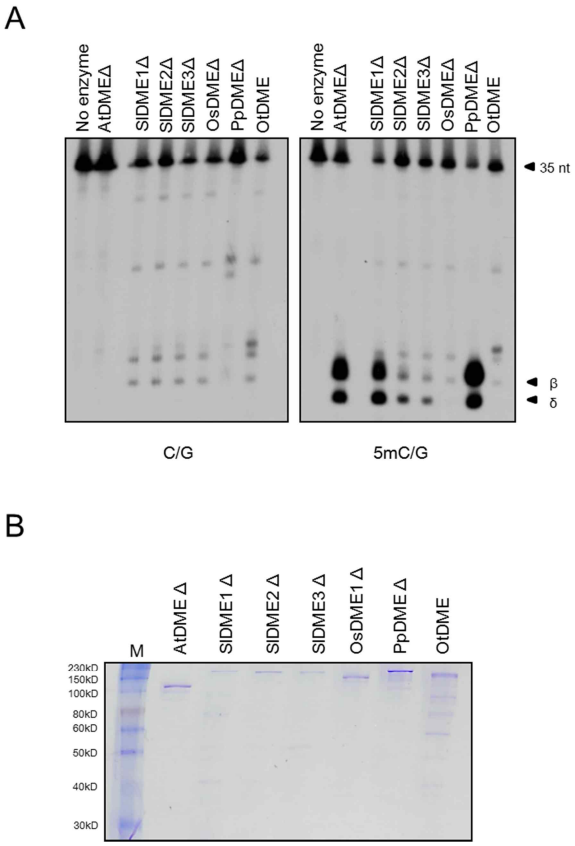


Figure 5. The 5mC excision activity of DME homologs isolated from diverse plant species. (A) The 5mC excision activity was performed with OtDME, PpDME, SlDME1, SlDME2, SlDME3, AtDME and OsDME from algae (*Ostreococcus tauri*), moss (*Physcomitrella patens*), dicot (*Solanum lycopersicum* and *Arabidopsis thaliana*), and monocot (*Oryza sativa*) species. Enzymes were reacted with radiolabeled 35-mer oligonucleotide substrate (25 nM) containing 5mC or C at position 18. Enzyme reactions were done at 37 ^{o}C for 2 hr. AtDMEΔ was used as a reaction control. (B) Affinity-Purified enzymes used for 5mC excision assay. Enzymes were electrophoresed on a 10% polyacrylamide gel and visualized by coomassie brilliant blue staining. M is a size marker.

The 5mC excision activity is not detected in DME homolog in algae

It was previously reported that the HhH, GPD and 4Fe-4S cluster in the glycosylase domain are necessary for DME to excise 5mC (Guan et al., 1998; Kuo et al., 1992; Mok et al., 2010). OtDME also contains significantly conserved HhH, GPD and 4Fe-4S cluster motifs, but 5mC excision activity was observed (Figures 1, 5). The 5mC excision activity test was performed with *E.coli* cell extract expressing OspDME (jgi|OstRCC809|36503|fgenesh1_pg.C_scaffold_5000084, *Ostreococcus sp*. *RCC809*). The 5mC excision activity was detected with AtDME cell extract, producing 3'-OH by AP endonuclease and AtDME activities in the cell extract (Figure 6A). However, the 5mC excision activity was not detected in OspDME cell extract albeit proper expression of OspDME was achieved (Figures 6A, 6C). Also the 5mC excision activity was not detected in the cell extract expressing OtDME, OspDME or OspDMEΔ (Figures 6B, 7). These data indicate that the plant-specific DNA demethylation system might be absent in green algae. These results also suggest that after the emergence of land plants, the DME family proteins might have acquired the 5mC excision activity.

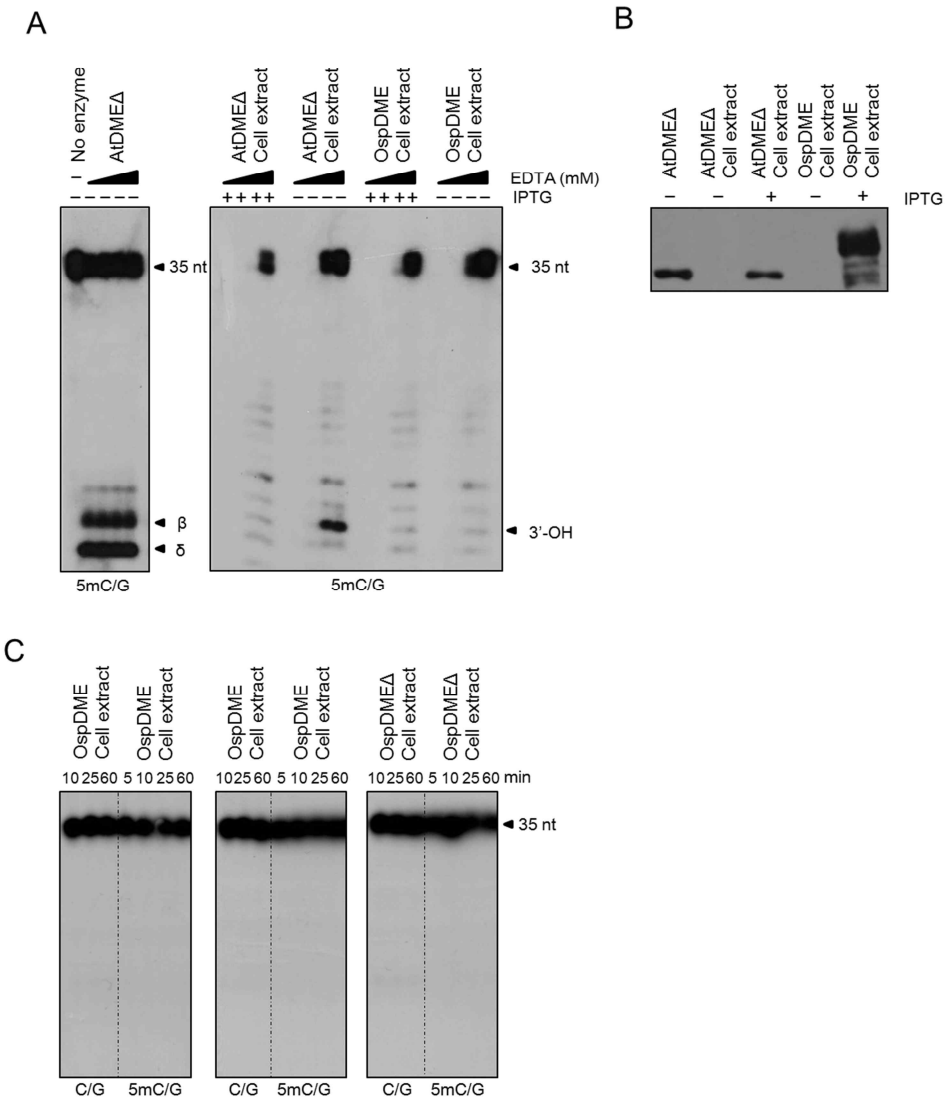

A
No enzyme
AtDMEΔ
AtDMEΔ Cell extract
AtDMEΔ Cell extract
OspDME Cell extract
OspDME Cell extract
EDTA (mM)
IPTG
35 nt
β
δ
3'-OH
5mC/G
5mC/G
B
AtDMEΔ
AtDMEΔ Cell extract
AtDMEΔ Cell extract
OspDME Cell extract
OspDME Cell extract
IPTG
C
OspDME Cell extract
OspDME Cell extract
OspDME Cell extract
OspDME Cell extract
OspDMEΔ Cell extract
OspDMEΔ Cell extract
10 25 60 5 10 25 60 min
35 nt
C/G
5mC/G
C/G
5mC/G
C/G
5mC/G

Figure 6. The 5mC excision activity of algal DME homologs. (A) The 5mC excision activity test with cell extract of *E. coli* expressing OspDME. AtDMEΔ, cell were used as reaction controls. The reaction was performed with increasing amounts of EDTA (0.1, 0.5, 1 or 2 mM) and IPTG. Reactions were done at 37 ^{o}C for 2 hr. (B) Western blot analysis of proteins from cell extracts by anti His antibody. Purified AtDMEΔ was used as s positive control. (C) The 5mC excision activity test with cell extract of *E. coli* expressing OtDME, OspDME and OspDMEΔ. The reaction was performed with various time courses (5, 10, 25, 60 min) with 2 mM of EDTA.

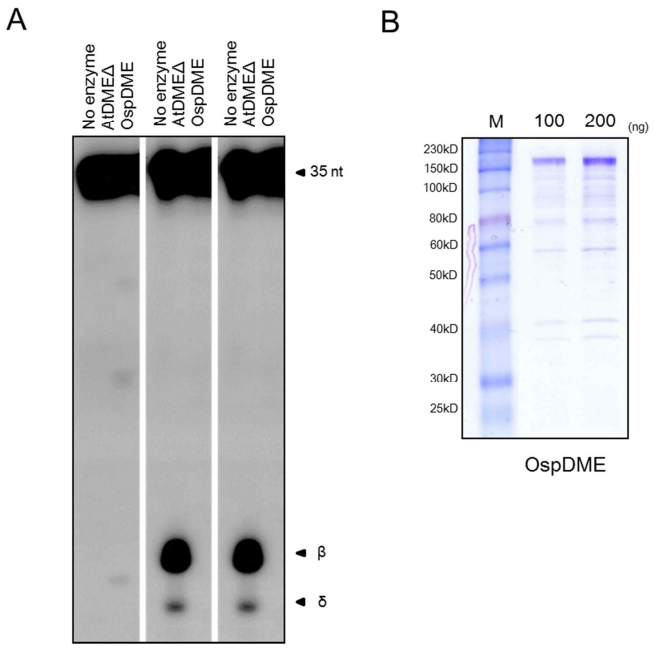


Figure 7. Base excision activity test of OspDME. (A) Base excision activity was performed with OspDME. Enzymes were reacted with radiolabeled 35-mer oligonucleotide substrate containing C, 5mC or T at position 18 at 37 ^{o}C for 2 hr. AtDMEΔ was used as a reaction control. (B) OspDME was electrophoresed on a 10% polyacrylamide gel and visualized by coomassie brilliant blue staining. M is a size marker.

Glycosylase domain of algal DME homologs cannot excise 5mC with domain A and B of AtDME

We showed that glycosylase domains of algal DME homologs have conserved HhH, GPD and 4Fe-4S cluster motifs. But relatively little sequence similarity of domains A and B compared with those of other DME homologs was present (Figure 1). This implies that the absence of 5mC excision activity in algal DME homologs might be due to divergent sequences in domains A and B. Since the chimeric proteins produced by domain swapping between AtDME and other family members showed 5mC excision activity (Jang et al., unpublished), I investigated whether the glycosylase domain of algal DME homolog contributes to the 5mC excision activity or not. The chimeric enzyme in which the glycosylase domain of OtDME was flanked with domains A and B of AtDME, (DOD) was produced. As shown in Figure 8, the DOD chimeric proteins did not produce any β- or δ-elimination products (Figure 8). This result implies that algal DME homolog has the glycosylase domain that cannot process 5mC, and might have other unknown functions comparing with DME homologs in land plants.

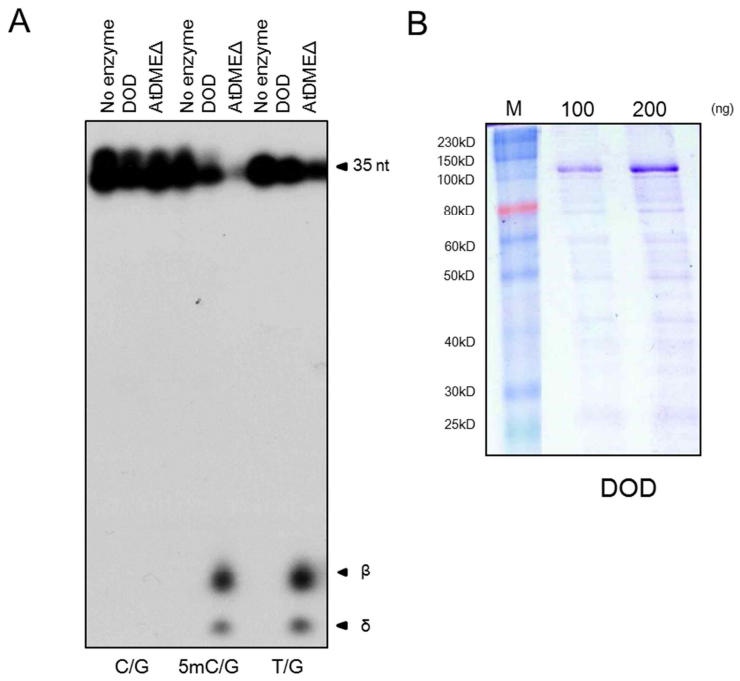


Figure 8. Base excision activity of chimeric enzymes. The chimeric enzyme (DOD) was constructed with domains A and B from AtDME with the glycosylase domain of OtDME. (A) The 5mC excision activity of DOD. AtDMEΔ was used as a reaction control. Enzymes were reacted with radiolabeled 35-mer oligonucleotide substrate containing C, 5mC or T at position 18. Reactions were done at 37 ℃ for 2 h. (B) DOD enzyme was electrophoresed on a 10% polyacrylamide gel and visualized by Coomassie brilliant blue staining. M is a size marker.

DISCUSSION

Plant specific DNA demethylation system was established after the emergence of land plants

The DME family proteins in *A. thaliana* have three functional domains interspersed with poorly conserved regions. My study shows that these distinct domains are highly conserved in all plant lineages including *Ostreococcus, Micromonas* and *Physcomitrella* (Figure 1). The phylogenetic and biochemical analyses on DME family proteins from various species provide clues to the evolutionary path and differentiation of plant-specific DNA demethylases. Algal species have a single conserved DME homolog in the annotated genomes but these ancient enzymes have no 5mC excision activity (Figures 5, 6, 7). On the other hand, two or three highly conserved DME homologs are present in *P. patens* and *S. moellendorffii*, and PpDME has a canonical 5mC excision activity (Figure 5). These results led me to hypothesize that the sequences of conserved DME family domains were created first in algae, from which the sequences of plant-specific DNA demethylases present in land plants have originated.

In this study, algal DME homologs exhibit no 5mC excision activity (Figures 5, 6, 7). In general, DNA methylation in higher multicellular eukaryotes regulates tissue-specific gene expression during cellular and developmental processes (Bender, 2004; Bird, 2002). The reason of absence of 5mC excision activity in algae is speculated that because algae used here for biochemical analyses are unicellular organisms, active DNA demethylation might be unnecessary to tissue-specific gene regulation. It is plausible that 5mC excision

activity in algae is dispensable, and alternatively, the primitive DME homologs in these organisms might have other functions. Therefore, it suggests that ancestral sequences of DME homologs were created in algae and after the emergence of land plants, these genes may have acquired the functions of DNA demethylase, providing essential functions to regulate plant development.

In addition to 5mC excision, excision activity of oxidative damaged DNA bases (5-hydroxyl-deoxycytosine, 5-hydroxyl-deoxyuracil and thymidine glycol) was identified in AtDME (Data not shown). Besides AtDME, it was reported that ROS1, directly or indirectly, participates in oxidative DNA damage repair in *A. thaliana* (Questo et al. 2013). Thus, these findings imply that when ancient algae colonized the land, primal DME homologs of land plants were probably used to cope with increased DNA oxidative damages caused by direct exposure to high-energy radiation from sunlight. And after the adaptation to land, it is considered that the active DNA demethylation system was brought in established as being necessary to dynamically regulate DNA methylation.

This study suggests that the DNA demethylation system is likely absent in green algae or significantly different from that of land plants. Therefore, I speculate that DNA demethylases have evolved from an ancestral algal protein which has similar structures but might have a different function(s). After the emergence of land plants, the ancestral genes may have diverged and acquired DNA demethylase activity, providing essential functions to regulate physiological and developmental processes required for new land plant era.

References

Bender, J. (2004). DNA methylation and epigenetics. Annu Rev Plant Biol *55*, 41-68.

Bestor, T.H. (2000). The DNA methyltransferases of mammals. Hum Mol Genet *9*, 2395-2402.

Bhutani, N., Brady, J.J., Damian, M., Sacco, A., Corbel, S.Y., and Blau, H.M. (2010). Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature *463*, 1042-1047.

Bhutani, N., Burns, D.M., and Blau, H.M. (2011). DNA demethylation dynamics. Cell *146*, 866-872.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev *16*, 6-21.

Bruner, S.D., Norman, D.P., and Verdine, G.L. (2000). Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. Nature *403*, 859-866.

Chan, S.W., Henderson, I.R., and Jacobsen, S.E. (2005). Gardening the genome: DNA methylation in Arabidopsis thaliana. Nat Rev Genet 6, 351-360.

Choi, Y., Gehring, M., Johnson, L., Hannon, M., Harada, J.J., Goldberg, R.B., Jacobsen, S.E., and Fischer, R.L. (2002). DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in arabidopsis. Cell *110*, 33-42.

Choi, Y., Harada, J.J., Goldberg, R.B., and Fischer, R.L. (2004). An invariant aspartic acid in the DNA glycosylase domain of DEMETER is necessary for transcriptional activation of the imprinted MEDEA gene. Proc Natl Acad Sci U S A *101*, 7481-7486.

Cortellino, S., Xu, J., Sannai, M., Moore, R., Caretti, E., Cigliano, A., Le Coz, M., Devarajan, K., Wessels, A., Soprano, D., et al. (2011). Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. Cell *146*, 67-79.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

David, S.S., O'Shea, V.L., and Kundu, S. (2007). Base-excision repair of oxidative DNA damage. Nature *447*, 941-950.

Doerfler, W. (1983). DNA methylation and gene activity. Annu Rev Biochem *52*, 93-124.

Gehring, M., Huh, J.H., Hsieh, T.F., Penterman, J., Choi, Y., Harada, J.J., Goldberg, R.B., and Fischer, R.L. (2006). DEMETER DNA glycosylase establishes MEDEA polycomb gene self-imprinting by allele-specific demethylation. Cell *124*, 495-506.

Gehring, M., Reik, W., and Henikoff, S. (2009). DNA demethylation by DNA repair. Trends Genet *25*, 82-90.

Gong, Z., Morales-Ruiz, T., Ariza, R.R., Roldan-Arjona, T., David, L., and Zhu, J.K. (2002). ROS1, a repressor of transcriptional gene silencing in

Arabidopsis, encodes a DNA glycosylase/lyase. Cell *111*, 803-814.

Guan, Y., Manuel, R.C., Arvai, A.S., Parikh, S.S., Mol, C.D., Miller, J.H., Lloyd, S., and Tainer, J.A. (1998). MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily. Nat Struct Biol *5*, 1058-1064.

Guillard, R.R., and Ryther, J.H. (1962). Studies of marine planktonic diatoms. I. Cyclotella nana Hustedt, and Detonula confervacea (cleve) Gran. Can J Microbiol *8*, 229-239.

Guo, J.U., Su, Y., Zhong, C., Ming, G.L., and Song, H. (2011). Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell *145*, 423-434.

He, X.J., Chen, T., and Zhu, J.K. (2011). Regulation and function of DNA methylation in plants and animals. Cell Res *21*, 442-465.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., et al. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science *333*, 1303-1307.

Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics *22*, 2971-2972.

Huh, J.H., Bauer, M.J., Hsieh, T.F., and Fischer, R.L. (2008). Cellular programming of plant gene imprinting. Cell *132*, 735-744.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y.

(2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129-1133.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science *333*, 1300-1303.

Iyer, L.M., Abhiman, S., and Aravind, L. (2011). Natural history of eukaryotic DNA methylation systems. Prog Mol Biol Transl Sci *101*, 25-104.

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

Kapazoglou, A., Drosou, V., Argiriou, A., and Tsaftaris, A.S. (2013). The study of a barley epigenetic regulator, HvDME, in seed development and under drought. BMC Plant Biol *13*, 172.

Kim, M.Y., and Zilberman, D. (2014). DNA methylation as a system of plant genomic immunity. Trends Plant Sci *19*, 320-326.

Kriaucionis, S., and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science *324*, 929-930.

Kuo, C.F., McRee, D.E., Fisher, C.L., O'Handley, S.F., Cunningham, R.P., and Tainer, J.A. (1992). Atomic structure of the DNA repair [4Fe-4S] enzyme endonuclease III. Science *258*, 434-440.

La, H., Ding, B., Mishra, G.P., Zhou, B., Yang, H., Bellizzi Mdel, R., Chen, S., Meyers, B.C., Peng, Z., Zhu, J.K., et al. (2011). A 5-methylcytosine DNA glycosylase/lyase demethylates the retrotransposon Tos17 and promotes its transposition in rice. Proc Natl Acad Sci U S A *108*, 15498-15503.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet *11*, 204-220.

Long, H.K., Blackledge, N.P., and Klose, R.J. (2013). ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. Biochem Soc Trans *41*, 727-740.

Matzke, M.A., and Birchler, J.A. (2005). RNAi-mediated pathways in the nucleus. Nat Rev Genet *6*, 24-35.

Mok, Y.G., Uzawa, R., Lee, J., Weiner, G.M., Eichman, B.F., Fischer, R.L., and Huh, J.H. (2010). Domain structure of the DEMETER 5-methylcytosine DNA glycosylase. Proc Natl Acad Sci U S A *107*, 19225-19230.

Morales-Ruiz, T., Ortega-Galisteo, A.P., Ponferrada-Marin, M.I., Martinez-Macias, M.I., Ariza, R.R., and Roldan-Arjona, T. (2006). DEMETER and REPRESSOR OF SILENCING 1 encode 5-methylcytosine DNA

glycosylases. Proc Natl Acad Sci U S A *103*, 6853-6858.

Newell-Price, J., Clark, A.J., and King, P. (2000). DNA methylation and silencing of gene expression. Trends Endocrinol Metab *11*, 142-148.

Norman, D.P., Chung, S.J., and Verdine, G.L. (2003). Structural and biochemical exploration of a critical amino acid in human 8-oxoguanine glycosylase. Biochemistry *42*, 1564-1572.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature *497*, 579-584.

Ono, A., Yamaguchi, K., Fukada-Tanaka, S., Terada, R., Mitsui, T., and Iida, S. (2012). A null mutation of ROS1a for DNA demethylation in rice is not transmittable to progeny. Plant J *71*, 564-574.

Ortega-Galisteo, A.P., Morales-Ruiz, T., Ariza, R.R., and Roldan-Arjona, T. (2008). Arabidopsis DEMETER-LIKE proteins DML2 and DML3 are required for appropriate distribution of DNA methylation marks. Plant Mol Biol *67*, 671-681.

Penterman, J., Zilberman, D., Huh, J.H., Ballinger, T., Henikoff, S., and Fischer, R.L. (2007). DNA demethylation in the Arabidopsis genome. Proc Natl Acad Sci U S A *104*, 6752-6757.

Qian, W., Miki, D., Zhang, H., Liu, Y., Zhang, X., Tang, K., Kan, Y., La, H., Li, X., Li, S., et al. (2012). A histone acetyltransferase regulates active DNA

demethylation in Arabidopsis. Science *336*, 1445-1448.

Questa, J.I., Fina, J.P., and Casati, P. (2013). DDM1 and ROS1 have a role in UV-B induced- and oxidative DNA damage in A. thaliana. Front Plant Sci *4*, 420.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y., et al. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science *319*, 64-69.

Scharer, O.D., and Jiricny, J. (2001). Recent progress in the biology, chemistry and structural biology of DNA glycosylases. Bioessays *23*, 270-281.

Stroud, H., Greenberg, M.V., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E. (2013). Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell *152*, 352-364.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet *9*, 465-476.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol *30*, 2725-2729.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature *473*, 343-348.

Wilson, G.G. (1988). Type II restriction modification systems. Trends Genet *4*,

314-318.

Worden, A.Z., Lee, J.H., Mock, T., Rouze, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V., et al. (2009). Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science *324*, 268-272.

Wu, S.C., and Zhang, Y. (2010). Active DNA demethylation: many roads lead to Rome. Nat Rev Mol Cell Biol *11*, 607-620.

Zemach, A., Kim, M.Y., Silva, P., Rodrigues, J.A., Dotson, B., Brooks, M.D., and Zilberman, D. (2010). Local DNA hypomethylation activates genes in rice endosperm. Proc Natl Acad Sci U S A *107*, 18729-18734.

Zemach, A., and Zilberman, D. (2010). Evolution of eukaryotic DNA methylation and the pursuit of safer sex. Curr Biol *20*, R780-785.

Zhang, H., and Zhu, J.K. (2012). Active DNA demethylation in plants and animals. Cold Spring Harb Symp Quant Biol *77*, 161-173.

Zhu, J.K. (2009). Active DNA demethylation mediated by DNA glycosylases. Annu Rev Genet *43*, 143-166.

초 록

DNA 메틸화 현상은 다양한 진핵 생물에서 후생유전학적인 유전자 발현 조절 주요 표지 중 하나로 알려져 있다. 애기장대의 DEMETER (DME)와 DME의 상동 단백질들은 5-메틸시토신을 직접적으로 인지하여 제거하는 능동적인 DNA 탈 메틸화 현상의 역할을 수행한다. 이러한 *DME* 유전자 군은 식물에서만 특이적으로 발견되며 동물에서 그 대응관계에 있는 유전자는 보고된 바가 없다. 그러므로 DME 단백질의 기원과 식물의 진화과정에서 DME 단백질의 변화를 유추해 보는 것은 흥미 있는 일이다. 이를 추정하기 위하여, 조류와 이끼, 종자식물을 포함한 42개의 종의 135개의 DME 상동단백질에 대한 계통분석이 실시되었다. 탈 메틸화 과정에 중요한 기능을 가진 DME 단백질의 3개의 도매인이 모든 육상식물에서 공통적으로 발견되었고 매우 보존되어있다는 것이 확인되었다. 반면에 조류의 상동단백질에서는 그 보존 정도가 떨어지는 것으로 나타났다. 재조합 단백질을 이용한

생화학적 분석 결과 이끼류에서 유래한 상동단백질에서는 DNA 탈 메틸화 활성을 확인하였지만 조류의 상동단백질에서는 그 활성을 확인하지 못하였다. 이러한 발견은 DNA 탈 메틸화 효소의 기능이 이끼류를 비롯한 육상식물과 조류가 다르다는 것을 의미한다. 이 연구를 통하여 DNA 탈 메틸화 과정의 성립과 그 생물학적인 중요성에 대한 진화적인 관점을 제시한다.