이학박사 학위논문

DNA methylation dynamics and importance of its regulation during embryo development in *Arabidopsis*

애기장대 배 발달과정 동안 DNA 메틸레이션 역동성과 그 조절의 중요성

2023년 8월

서울대학교 대학원

생명과학부

이 재 훈

DNA methylation dynamics and importance of its regulation during embryo development in *Arabidopsis*

애기장대 배 발달과정 동안 DNA 메틸레이션 역동성과 그 조절의 중요성

지도 교수 최 연 희

이 논문을 이학박사 학위논문으로 제출함

2023년 8월

서울대학교 대학원

생명과학부

이 재 훈

이재훈의 이학박사 학위논문을 인준함

2023년 8월

위 원 장

부위원장

위원

위원

위원

DNA methylation dynamics and importance of its regulation during embryo development in Arabidopsis

A dissertation submitted in partial fulfillment of
the requirement for the degree of

DOCTOR OF PHILOSOPHY

to the faculties of

the School of Biological Sciences

at

Seoul National University

by
Jaehoon Lee

Date Approved	
Feb. 13, 2023	**Chair**
	Vice Chair
	Examiner
	Examiner
	Examiner

Abstract

DNA methylation dynamics and importance of its regulation during embryo development in *Arabidopsis*

Jaehoon Lee

School of Biological Sciences

The Graduate School

Seoul National University

DNA methylation refers to a modification at the C-5 of a cytosine base. DNA methylation is one of the epigenetic factors which can alter gene and transposon activity without changing DNA sequences. It also helps organisms for adapting to their environments. A double fertilization takes place in flowering plants; a sperm cell fused with an egg cell produces diploid embryo and the other sperm cell fused with diploid central cell produces triploid endosperm. Endosperm nourishes

developing embryo, thus it has a similar function to a placenta in mammals. Different from embryo, endosperm is a terminally differentiated companion tissue in that its genome is not transmitted to the next generation. Embryo and endosperm develop inside the seed coat. As seed matures, growing embryo takes the most of the seed volume at the later stage of embryogenesis.

A lot of studies reported the defects in DNA methylation mutants embryogenesis, thereby it has been well known that the proper DNA methylation is critical for normal embryo development. However, the causality of methylation difference between different Arabidopsis ecotypes during embryo development has not been fully understood.

Here, I present comprehensive pairwise comparisons for the methylation differences during seed ripening and germination between two different ecotypes. Using Columbia-0 (Col-0) and Cape Verde Island (Cvi) ecotypes, I chose three developmental seed stages; 1) freshly harvested (FH), 2) after ripening (AR), 3) germination stimulated (GS). I discovered overall dynamics of DNA methylation during seed ripening and germination. Among three cytosine contexts, CG context showed the most differently methylated as Col displayed always higher mCG levels than Cvi for three stages in both genic regions and transposable elements (TEs). In contrast, non-CG methylation levels were very similar between both ecotypes. Interestingly, I found Col-specific DNA

sequences (Col-own region, COR), where were considered as diverged regions, were more hypermethylated than conserved regions (CR) in both ecotypes. CORs were enriched with TEs and structural variations (SVs) including insertions, deletions. TEs in COR were longer in size, more pericentromeric, and more hypermethylated than those in the CRs, suggesting that the COR TEs are younger than TEs in CR. By using previously reported *ago4/6/9* and *drm1/2* methylome datasets, I demonstrated that RNA-directed DNA methylation (RdDM) pathway contributes to the hypermethylation in younger CORs.

Although the representative patterns of DNA methylation, specifically for a gradual increase in CHH methylation during embryo development have been already reported by other groups, by using different five embryo developmental stages, I found nine different clustered regions based on the similar methylation change during embryonic development. Those clusters showed different positional and chromatic characters. Taken all together, my results demonstrate the possibility of regulation on newly acquired regions by RdDM pathway. And I suggest a new analysis to reveal the biological meanings of CHH methylation changes during embryo development.

Keywords: Structural variant (SV), RNA-directed DNA methylation (RdDM), Methylome, Arabidopsis, embryo development

Student Number: 2016-20395

Table of Contents

List of Figures

List of Tables

Abbreviations

FH	Freshly harvested
AR	After ripening
GS	Germination stimulated
COR	Col-0 own region
CR	Conserved region
DMR	Differentially methylated region
eDMR	Differentially methylated region in ecotype comparison
sDMR	Differentially methylated region in stage comparison
RdDM	RNA-directed DNA methylation
gbM	Genebody methylation
SNP	Single nucleotide polymorphism
SV	Structural variant
TE	Transposable element
WGBS	Whole genome bisulfite sequencing

I. Background

1.1 Epigenetics

After sequencing technique has rapidly developed, whole genomic sequences of many model species have been reported from year 2000. At present, Next generation sequencing makes it possible to obtain the entire sequence information of an organism more easily. However, despite this current situation, genetics or the DNA sequences alone still cannot explain a lot of phenomena observed in living organisms. This is, at least partly, due to the epigenetic regulation on genes, TEs and chromatins. Epigenetics stand for the “above-genetics”, meaning the changes of genetic environment without changes in DNA sequences. Nucleosome density, histone modification, DNA methylation and small RNAs the representatives of the epigenetic regulation.

1.2 DNA methylation in *Arabidopsis*

DNA methylation found in various organisms refers to an attachment of methyl-group to the 5th carbon at cytosine base. Whereas histone modification is commonly conserved within eukaryotes, DNA methylation is not (Figure 1). This does not mean that DNA methylation is not essential, because many organisms having DNA methylation utilize it for an important regulatory mechanism. Arabidopsis, a representative of flowering plants and a model organism that have DNA methylation system. Arabidopsis has all cytosine methylation contexts, CG, CHG and CHH (H=A, T or C) as well as DNA methyltransferases such as METHYLTRANSFERASE 1(MET1), CHROMOMETHYLASE 3(CMT3), DOMAINS REARRANGED METHYLASE 1/2 (DRM1/2) and CHROMOMETHYLASE 2 (CMT2) depending on each cytosine context. In addition, Arabidopsis also has active DNA demethylases such as DEMETER (DME), Repressor of transgene silencing 1 (ROS1), DEMETER-LIKE 1 (DML1) and DML2. Methylation levels are increased by DNA methyl transferases activity and decreased by active demethylase. DNA methylation levels are also decreased by passive demethylation without DNA methyltransferases when cells are divided. Therefore, DNA methylation levels can be tightly regulated to control gene expression and TE silencing.

DNA methylation can be divided by two different genomic features.

First, genebody methylation (gbM) has known to affect gene expression [1]. Second, DNA methylation on TE has mainly believed to be a defense mechanism to silence the expression of external DNA [2]. Those DNA methylation on genomic features were regulated by different way. Compared to gbM which is mainly consisted of CG methylation, TE is methylated in all cytosine contexts. CG methylation levels are quite stable during all life cycle, whereas non-CG methylation levels are dynamically changed [3]. Additionally, non-CG methylation are differently established by cell-types and by genomic environments [3, 4].

Compared to plants, mammals show DNA methylation mainly at CG context. Two rounds of global demethylation were observed in primordial germ cells and in an early embryo after fertilization during life cycle of mammals [5] and its importance in embryo development were reported in mammals [5]. Intriguingly, similar to the reprogramming of mCG levels in mammals during reproduction, mCHH levels in Arabidopsis are reduced in early embryos and gradually increased by RNA-directed DNA methylation pathway (RdDM pathway) [6].

Figure 1

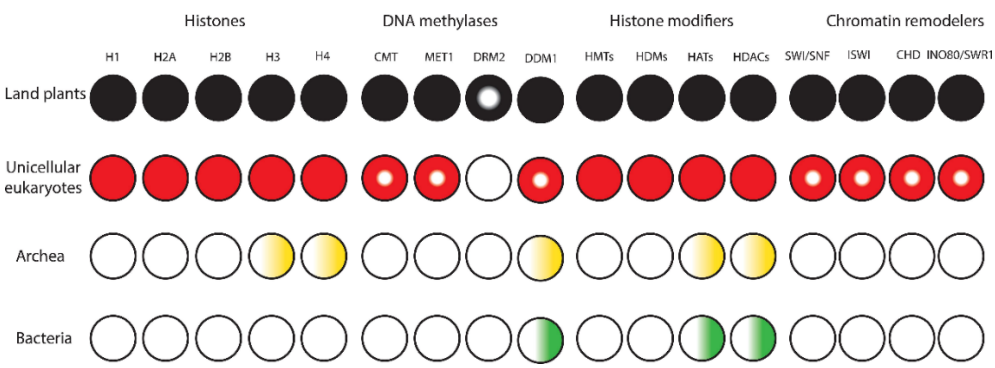

Figure 1. Presence of protein homologs.

Filled circle indicates presence of orthologs in all lineages. Semi-filled circle denotes partially conservation in lineages. Circle filled in gradient indicates presence of homologs [7].

II. Ecotype specific DNA methylation

2.1 Introduction

More than 1,000 ecotypes existing as natural variations are found in Arabidopsis thaliana [8]. Those ecotypes typically exhibit phenotypically and physiologically polymorphic. If genetic mutation or change is beneficial for living organism to adapt to a new environment, that specific mutation can be naturally selected and inherited stably. On the other hand, epigenetic change can alter chromatin states, resulting in changes in gene and TE expressions. Therefore, it can also increase survival rate of individuals if that change is beneficial. Compared to genetic changes, epigenetic changes are reversible and more flexible to an environment. Since epigenetic changes are rapidly responsive to a new environment, it can be a very good strategy for organisms to adapt to the changing environment. Such epialleles frequently associated with phenotypic changes are found in nature and and are inherited through generations [9]. Meanwhile, different ecotypes have shown to have different methylome states [10]. Multiple pathways can contribute to such different methylation states [10]. Taken together, phenotypic differences in different ecotypes are the product of collective differences in the genome and epigenome. However, the previous studies mainly

focused on epigenetic differences in the commonly existing genomic sequences conserved among ecotypes. Analyses for the epigenomes in ecotype-specific regions are relatively recently conducted.

It is important to know the epigenome differences in different ecotypes, but understanding of causality generating such differences is also important. CG methylation (mCG) is symmetrically maintained by maintenance methyltransferase, DNA METHYLTRANSFERASE 1 (MET1) during DNA replication [11]. CHG (mCHG) is also symmetrically maintained by CHROMOMETHYLASE 3 (CMT3) using its chromodomain that recognizes H3K9 methylation [12, 13]. Asymmetric CHH methylation (mCHH) is established by RNA-directed DNA methylation (RdDM) [2, 14, 15]. Heterochromatic regions, mainly TE or repeat sequence, are transcribed by plant-specific RNA polymerase IV (pol IV), and the transcripts are processed into 24-nucleotide(nt) small RNAs or 21-nt aberrant small RNAs by Dicers. These small RNAs that are bound to Argonaute (Ago) act together with pol V-derived transcripts and these complexes recruit de novo methyltransferase, DOMAIN REARRANGED METHYLTRANSFERASE 2 (DRM2), to the target DNA sequences to methylate at all three cytosine contexts. Also, chromodomain-containing CMT2 methylates heterochromatic mCHH sites [16, 17].

Here, I present a genome-wide comparison of the DNA methylome

during seed ripening and germination in Col-0 and Cvi. I classified Arabidopsis genome into two regions based on the mapping characteristic of whole-genome bisulfite sequencing (WGBS) reads to define the regions in reference Col-0 for comparing Col-0 with Cvi. Those two regions are 1) comparable conserved regions (CR), where the WGBS reads from both Col-0 and Cvi are well mapped to Col-0 reference genome; 2) Col-0 own regions (COR), where only the Col-0 WGBS reads are well mapped (See Methods and materials section for the detailed conditions). I demonstrate that CR exhibited ecotype-specific differentially methylated regions (eDMRs). Through a collaboration with Sang-Yoon Shin, it was found that eDMRs were strongly correlated with highly expressed 24-nt small RNA clusters in an ecotype-specific manner. I also discovered significantly higher methylation levels in the COR than CR. Then, I found that structural variants (SVs), especially newly acquired in Col-0, are responsible for the higher methylation levels in COR, providing evolutionary insight into how different methylation levels in different ecotypes are originated from natural variation within a species.

2.2 Materials and methods

2.2.1 Plant growth condition

Two accessions of Arabidopsis thaliana, Col-0 (Columbia-0) and Cvi (Cape Verde Islands) were used in this study. Seeds were sown at MS (Marashige and Skoog) medium, grown until 10 days after germination, and transferred to the soil at 22°C under long photoperiod (16 hours of light, 8 hours of dark) with cool white fluorescent light (100 μmole/m2/s). Fully matured green seeds from brownish siliques just after starting to dry were harvested from both ecotypes and used for FH seeds. Seeds that were harvested from fully dried Col-0 plants were used for Col-0 AR. Cvi seeds harvested from the same staged dried plants were further dried for 60 days and then used for Cvi AR. For GS samples, AR-staged seeds from both ecotypes were treated stratification in the dark at 4°C for three days, exposed to white light for two hours, and then incubated in the dark at 22°C for 24 hours.

2.2.2 DNA library construction

All samples were stored at -80°C after quick freezing with liquid nitrogen. DNA of all samples was extracted by grinding with mortar. All of the Whole-genome bisulfite sequencing (WGBS) libraries were constructed using the KAPA library preparation kit (Roche) and EqiTech

Bisulfite by Macrogen (Seoul, South Korea).

2.2.3 Sequencing data processing

All Sequencing procedure was performed with the HiSeq2000 platform by Macrogen (Seoul, South Korea). 101 bp paired-end reads were generated. All reads were trimmed sequentially by Trim Galore (options, --clip_R1 10 --three_prime_clip_R1 5) and Trimmomatic (options, SE -threads 16 SLIDINGWINDOW:2:20 MINLEN:40). Then, by using Bismark with bowtie2, WGBS reads of Col-0 and Cvi were mapped to the TAIR10 genome without allowing any mismatches. Mapped reads were deduplicated and then cytosine information was collected by using Bismark script (deduplicate_bismark and bismark_methylation extractor) under default conditions.

2.2.4 Measurement of DNA methylation levels

Factional methylation within 50 bp windows was calculated by the mean level of each cytosine without overlap. Informative windows including at least 3 cytosines with minimal 10 reads aligned per each cytosine context as a probe were used. For comparing methylation levels between ecotypes and during seed ripening and germination, windows valid within all of compared samples were used. Exceptionally,

a 5 kb window including at least 10 sites of cytosine methylation context with more than 10 reads was used for chromosomal view.

2.2.5 Identification of differentially methylated region

eDMRs only in conserved region (CR) by comparing two ecotypes at the same stage were defined with a methylation difference more than or less than a standard deviation from the mean of methylation difference using 50bp windows satisfying the above conditions which were used for collecting informative windows.

2.2.6 Identification of Conserved and Col-0 own region

Two groups of regions were defined, conserved region (CR) and Col-0 own region (COR), depending on the 101 bp mapped reads from bisulfite sequencing. If a 50 bp window probe satisfies the conditions of at least 3 cytosines with minimal 10 reads per each cytosine context in both Col-0 and Cvi, the probe is included in CR. If a probe satisfies the same conditions only in Col-0 but not in Cvi, the probe is included in COR.

2.2.7 Methylation from genebody and transposable element

Comparable genes and transposable elements (TEs), including at

least 5 cytosine sites with a minimum of 10 reads in both ecotypes were used for comparing. Four particular parts from the AR stage were sectioned; methylated in Col-0 but not Cvi (green), methylated in Cvi but not Col-0 (pink), middle mC level in both ecotypes (cyan), and high mC level in both ecotypes (purple).

2.2.8 Identification of variable regions between ecotypes

For identification of DNA sequence differences between Col-0 (TAIR10) and Cvi (v2.0), SNPs and structural variations were extracted using MUMmer4 (https://github.com/mummer4/mummer). The composition is counted using the number of overlapped windows. According to the number of SNPs, the fractional methylation level (50 bp) was calculated using only Col FH as a representative.

2.3 Results

2.3.1 CG eDMRs are strongly maintained in the genebody throughout development

To understand variation in the molecular dynamics during seed ripening and germination processes, Whole genome bisulfite sequencing (WGBS) was performed, on seeds sampled from two Arabidopsis accessions, Col-0 and Cvi. I chose Cvi to compare to the reference, Col-0, since Cvi is known to have the lowest CG gene body methylation (gbM) among over 1,000 Arabidopsis accessions [10]. Seeds were sampled from three stages: freshly harvested (FH) fully matured green seeds from green plants, after-ripening (AR) seeds harvested from fully dried plants, and germination-stimulated seeds (GS) (Figure 2).

For a global methylome view, I used comparable conserved regions (hereafter, CR) from both ecotypes that are well mapped to TAIR10 in both ecotypes (see Methods). Figure 3 shows the chromosomal heat map of mCG, mCHG, and mCHH during development. The most striking difference between the two ecotypes was observed in mCG at chromosomal arms (e.g. black sector), both ecotypes show a distinctly higher mCHH change at pericentromeric regions (e.g. red sector) in the AR stage (Figure 3).

Figure 2

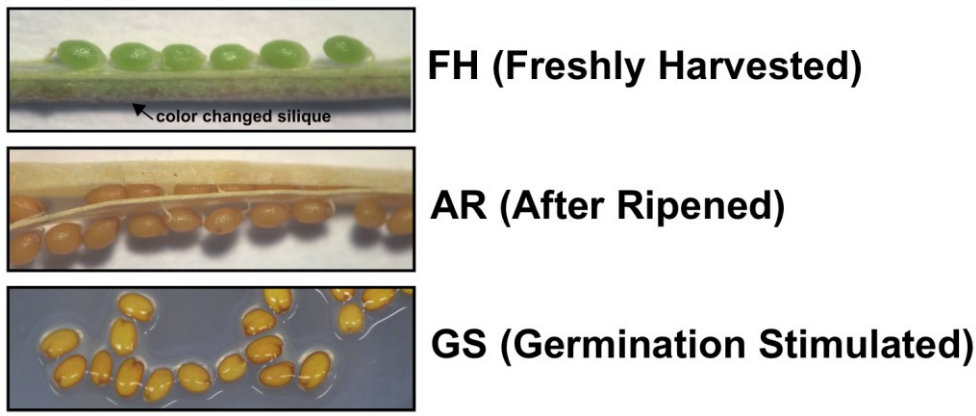

Figure 2. Images of sample stages used in this study.

3 stages during development from both ecotypes were used for this study

Figure 3

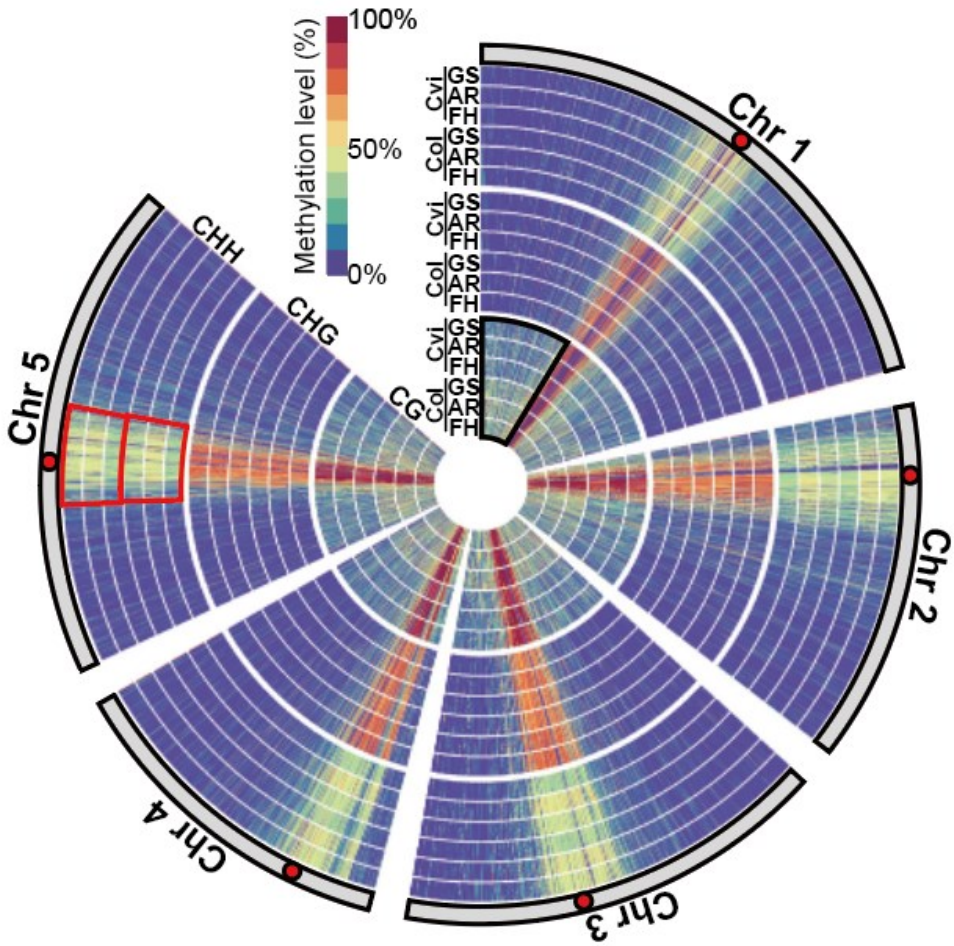

Figure 3. Genome view of DNA methylation during seed ripening and germination in both ecotypes.

Chromosomal DNA methylation view. Red-colored circle indicates centromere of each chromosome. Sections in black or red indicate each peri-centromere and chromosomal arm, respectively.

mCG did not globally change during the three developmental stages, maintaining the methylation difference between the two ecotypes (Figure 4A). In contrast, mCHG and mCHH were the highest in the AR stage, while their methylation levels were almost the same in the two ecotypes (Figure 4A). The dynamics of the genebody methylation (gbM) mirrored the global view (Figure 4). The non-CG methylation of TE showed the similar dynamic pattern to the non-CG genic regions, except for the higher methylation levels (Figure 4). On the other hand, the mCG pattern of TE was different from the gene or global view; AR seeds exhibited the highest methylation levels (Figure 4). These results suggested that Cvi, which is known to have the lowest mCG levels among all ecotypes [10], display the same mCHH dynamics during seeds maturation and germination as Col-0, at least for their conserved regions [18-20]. Therefore, cytosine methylation tends to reach maximal levels at the AR stage and then decrease during germination, although their levels in each accession are different.

Figure 4

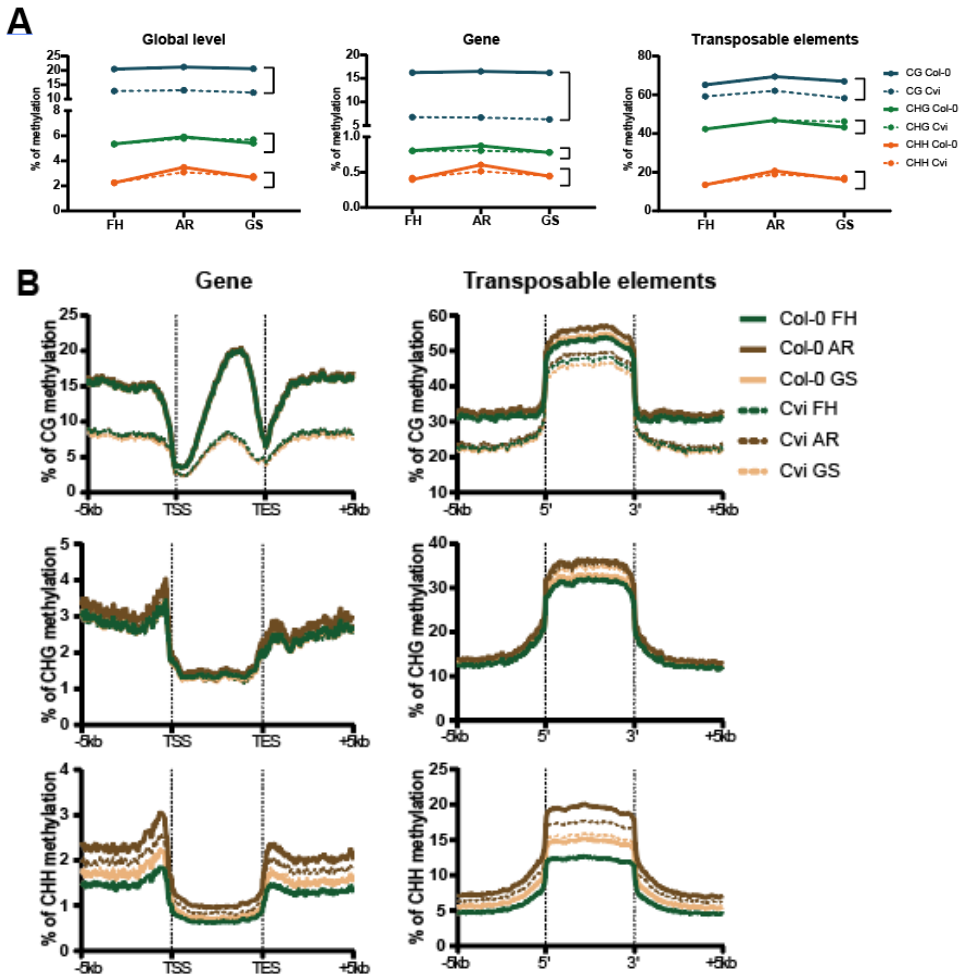

Figure 4. DNA methylation during seed ripening and germination in both ecotypes

A Average of DNA methylation level (%) of Col-0 (lined) and Cvi (dotted) was calculated according to developmental stages in global (left panel), gene (middle panel), and transposable elements (right panel). **B** The methylation level of gene, transposable elements, and surrounding regions were analyzed for each cytosine context.

Since the levels of non-CG methylation were very similar in both ecotypes and mCG were significantly different, the mCG difference is the primary source of the total methylation difference between Col-0 and Cvi.

Since the global mCG level ecotype differences, but not mCHG and mCHH levels, were maintained throughout seed ripening and germination processes, I asked if the global mCG level difference between both ecotypes is also maintained from local regions or not. First, I calculated the difference in methylation levels of each 50 bp window at each stage and each of the genomic features. Then, I defined eDMR from comparable CR where the DNA sequences exist in both ecotypes as the regions representing between-ecotype Δ(mC level) more than a standard deviation from the global mean of Δ(mC level) at each stage. Overall, CG eDMRs are more strongly maintained than CHG or CHH eDMRs during development (Table 1). Specifically, more than 89% of CG eDMRs (58,923) of a stage were maintained during development (Table 1). CHG eDMRs were slightly less maintained than CHH eDMRs in approximately half of conserved DMRs (22,647) during three developmental stages.

Table 1. The number of differentially methylated regions (eDMRs) during seed ripening and germination

The size of a window is 50 bp without overlap. R&G indicates continuous DMR during seed ripening and germination.

CONTEXT	STAGE	TOTAL	GENE	GENE&TE	TE	IGR
CG	R&G	58923	49205	464	4212	5042
	FH	66565	54651	558	5234	6122
	AR	65382	54182	535	4883	5782
	GS	65820	54319	532	5193	5776
CHG	R&G	22647	5933	452	9988	6274
	FH	40812	7775	788	22260	9989
	AR	38969	7366	732	21266	9605
	GS	45516	7592	810	26700	10414
CHH	R&G	77406	13981	1760	37089	24576
	FH	134631	20497	2740	73208	38186
	AR	127451	19341	2673	67778	37659
	GS	137852	19227	2750	77278	38597

Interestingly, genic regions were more robustly maintained than other genomic features in all cytosine contexts (Table 1). More than 90% of CG eDMRs (49,205) were well maintained in the genic regions. By contrast, eDMRs were the least maintained in TEs compared to other genomic features, especially for non-CG sites (Table 1).

To further confirm the highly maintained ecotype difference in the genic regions, the mC of each gbM between both ecotypes at each developmental stage was compared. Then, I sectioned the AR stage into four particular parts (copper metallic: methylated only in Col-0, pink: methylated only in Cvi, azure blue: middle mC levels in both ecotypes, purple: high mC levels in both ecotypes) to track the mC change of each gene in sections in FH and GS stages. Figure 5 shows the presence of more hyper mCG genes in Col-0 (X-axis) than Cvi (Y-axis). Furthermore, the position of sections was well maintained for genes even in other developmental stages, meaning that each mCG gbM was maintained. This result supported our previous finding that mCG gbM was static and strongly maintained differentially in the ecotypes throughout development. By contrast, the position and aspect of the sectioned parts from the AR stage in TEs were changed in FH and GS stages (Figure 5), suggesting mC of TE regions are dynamically changed.

Figure 5

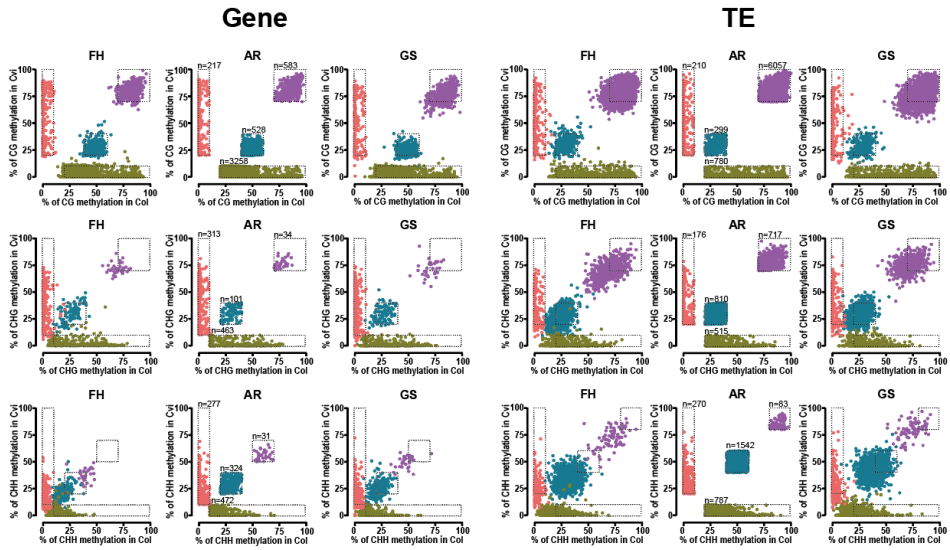

Figure 5. Scatterplot of genebody and transposable element methylation.

Stable difference of genebody methylation level between Col-0 and Cvi. Genes including more than 5 sites of analyzed cytosine context with at least 10 reads were analyzed. Four differentially colored sections were selected at the AR stage for tracing the change in other stages. Sectioned several parts from the AR stage, methylated in Col-0 but not Cvi (copper metallic), methylated in Cvi but not Col-0 (pink), middle level in both ecotypes (azure blue), and highly methylated in both ecotypes (purple). Selected four sections from highly methylated in Col-0 but not in Cvi, highly methylated in Cvi but not in Col-0, middle level in both ecotypes, and highly methylated in both ecotypes in the AR stage to track changes of methylation.

2.3.2 Methylation levels on eDMRs are positively correlated with differential expression of 24-nt small RNAs

I used transcriptome data prepared by JS Jeon laboratory and analyzed by SY Shin with same schemed sample, the data was analyzed by SY Shin, to determine whether the methylation difference between two ecotypes is caused by differential expression of a specific methylation regulating gene. As a result, any genes with dramatic difference was not found (Figure 6). However, some genes related to RdDM pathway were differentially expressed (marked with a cross in Figure 6). Therefore, to verify that the DNA methylation difference in both ecotypes is due to d activity of RdDM pathway, I checked the overlaps between differentially expressed 24-nt RNA regions and eDMRs to see whether methylation level differences are correlated. Differentially expressed small RNA cluster (e24sRC) in ecotype comparison was defined by SY Shin. As a result, a considerable number of overlaps were observed between Col-hypermethylated eDMRs (hyper eDMRs) and e24sRC-Col. Accordingly, a similar result was also observed between Col-hypomethylated eDMRs (hypo eDMRs) and e24sRC-Cvi (Figure 7).

Figure 6

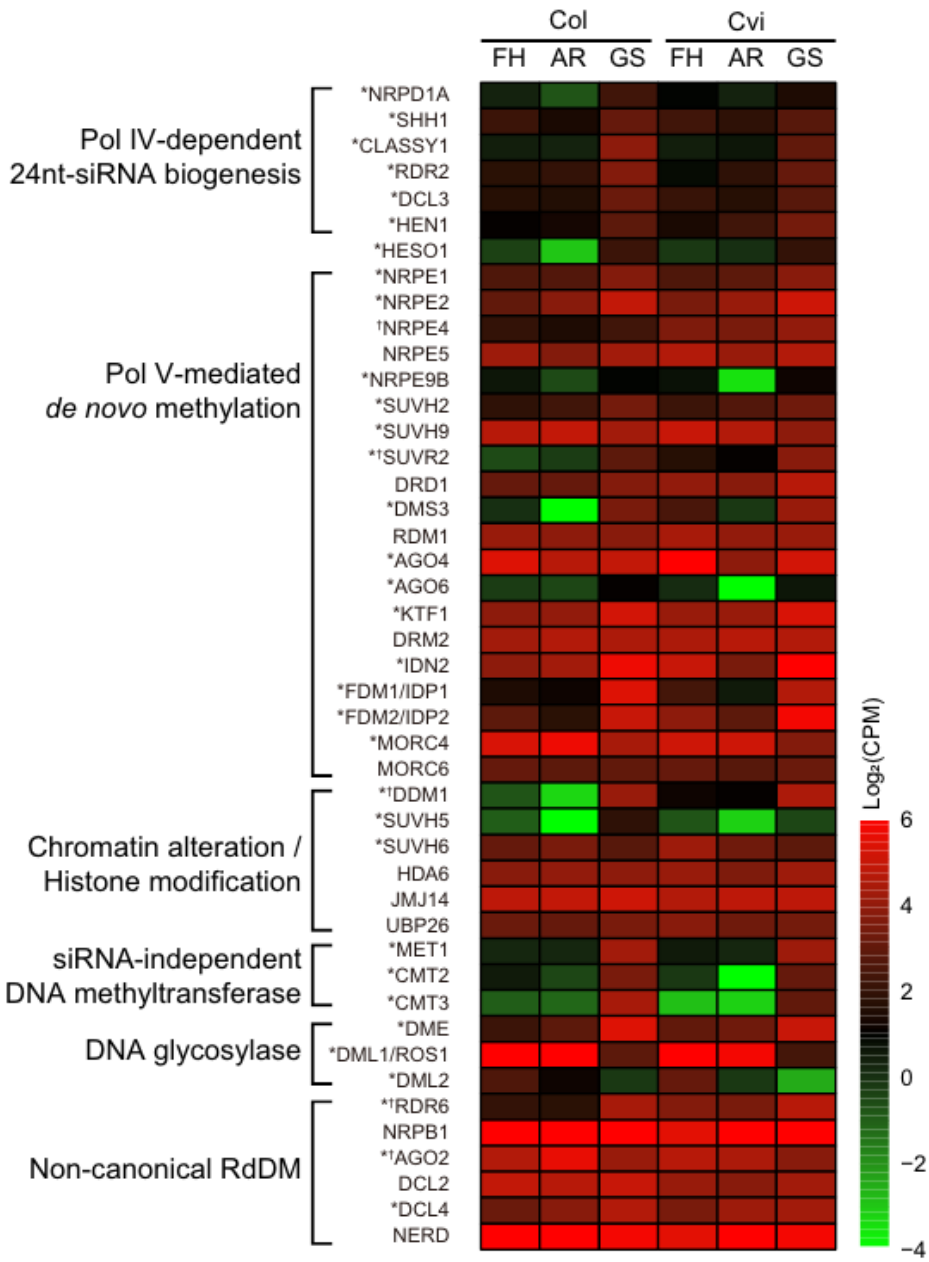
Col
Cvi
FH
AR
GS
FH
AR
GS
Pol IV-dependent 24nt-siRNA biogenesis
*NRPD1A
*SHH1
*CLASSY1
*RDR2
*DCL3
*HEN1
*HESO1
Pol V-mediated de novo methylation
*NRPE1
*NRPE2
†NRPE4
NRPE5
*NRPE9B
*SUVH2
*SUVH9
*†SUVR2
DRD1
*DMS3
RDM1
*AGO4
*AGO6
*KTF1
DRM2
*IDN2
*FDM1/IDP1
*FDM2/IDP2
*MORC4
MORC6
Chromatin alteration / Histone modification
*†DDM1
*SUVH5
*SUVH6
HDA6
JMJ14
UBP26
siRNA-independent DNA methyltransferase
*MET1
*CMT2
*CMT3
DNA glycosylase
*DME
*DML1/ROS1
*DML2
Non-canonical RdDM
*†RDR6
NRPB1
*†AGO2
DCL2
*DCL4
NERD
Log2(CPM)
6
4
2
0
−2
−4

Figure 6. Expression profiling of genes affecting DNA methylome and histone modification from two ecotypes of Arabidopsis.

The expression level of genes involved in DNA methylation, demethylation, histone modifications, and chromatin remodelers are represented as log2(CPM). Differentially expressed across stages are marked by an asterisk (*), and differentially expressed between ecotypes are marked by a crossmark (†).

Figure 7

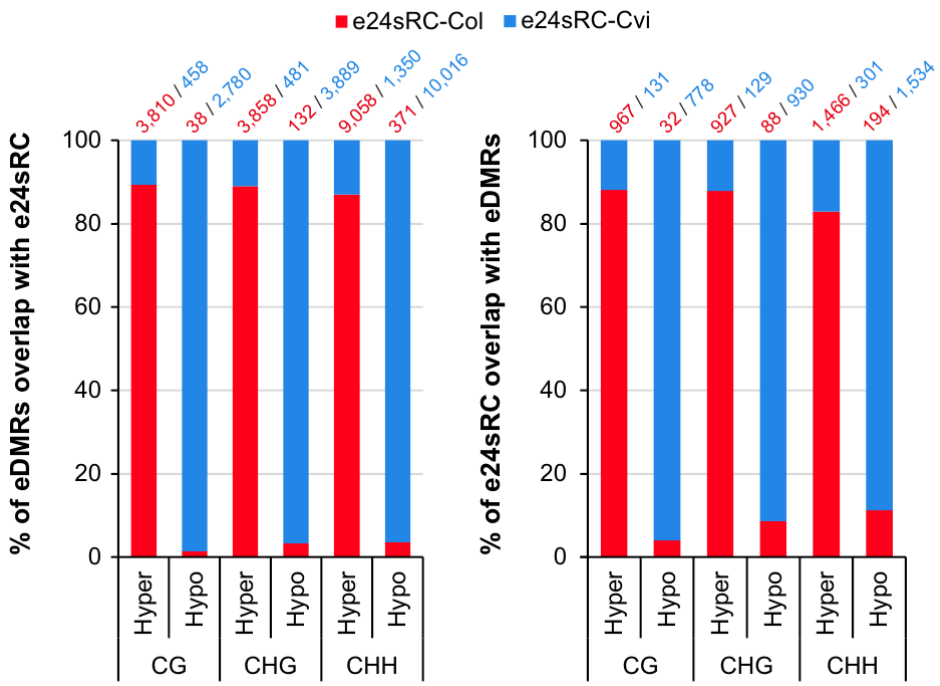

Figure 7. the relationship between eDMR and e24sRC.

Numbers and compositional ratio of ecotype-specific differentially methylated regions (eDMRs) overlapping e24sRCs (left panel), and that of e24sRCs overlapping eDMRs (right panel). Hyper: hypermethylated region in Col-0; Hypo: hypermethylated region in Cvi.

With positive correlation patterns between e24sRC expression levels and methylation on eDMRs, I speculated that differential methylation in eDMRs is mainly affected by RdDM pathway. To address the involvement of RdDM pathway on the formation of eDMR, I analyzed methylation difference using methylome dataset from ago4/6/9 and drm1/2 mutants sampled in mature embryos [21]. Results showed that the methylation level of eDMRs along with e24sRCs or Non-e24sRC (fold-change < 2-fold or FDR > 0.05 between ecotypes throughout all developmental stages) were significantly decreased compared to eDMRs that did not overlap with any 24sRCs (Figure 8).

In conclusion, these positive correlations between methylation and 24-nt small RNA expression in eDMRs and the mutant analysis results suggest the contribution of RdDM in regulating methylation levels in an ecotype-specific manner.

Figure 8

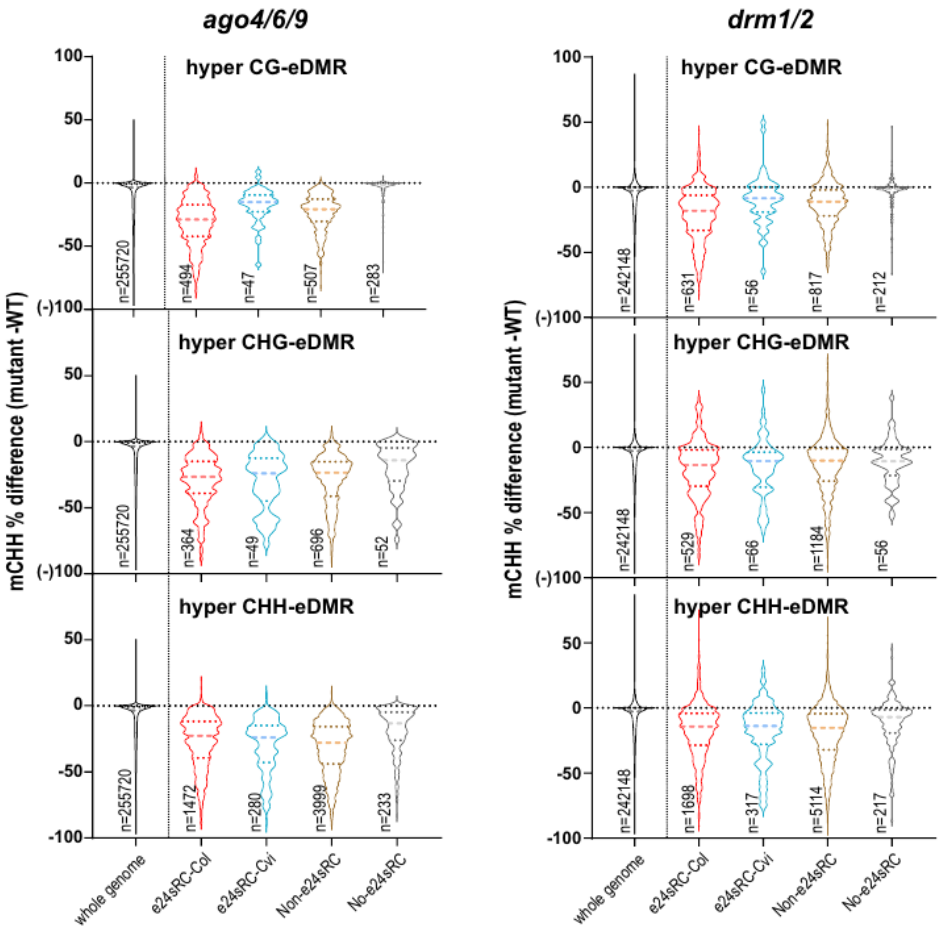

Figure 8. mCHH levels of eDMR in *ago4/6/9* and *drm1/2*.

DNA methylation change on eDMRs in RdDM machinery mutants. Labels in X-axis indicate categories of 24sRCs overlapping with eDMR (i.e., e24sRC-Col, e24sRC-Cvi, Non-e24sRC), or indicate eDMRs not overlapped with any of 24sRCs (No-e24sRC).

2.3.3 A large composition of TE and intrinsic features in COR contribute to hypermethylation

To compare methylome levels between ecotypes, I selected informative regions/probes passing the criteria of 50 bp windows, including at least 3 cytosine contexts with at least 10 reads, in all six samples analyzed. This is represented as an intersection part in Figure 9A or comparable conserved regions called “CR”, comprising more than 80% of total probes for each mC context. As seen in Figure 4A, no significant difference was found in average of non-CG methylation levels between ecotypes during ripening and germination processes in this CR (Figure 9B left). However, when I did the same global comparison from all measurable regions, which passed the criteria from at least one out of six samples (Figure 9A, union for each ecotype), Col-0 displayed significant higher average non-CG methylation level than Cvi across all three stages (Figure 9B right). A much bigger difference was found in mCG levels for the all measurable regions between two ecotypes compared to mCG differences in CR (Figure 9B). This indicated that the additionally included regions/probes, most of which were comprised of Col-0 own regions (hereafter, COR) in major, display higher non-CG methylation level compared to CR. Because I mapped the sequencing reads of both ecotypes to the Col-0 reference genome without allowing mismatch, COR would include many sequence differences between the ecotypes. My result suggests that COR is the site that contains higher

methylation levels in all cytosine contexts than in comparable intersection regions.

Therefore, I hypothesized the presence of a relationship between DNA sequence diversity and DNA methylation and analyzed the composition of the genomic features for those ecotype-specific regions. This analysis revealed that COR was composed of significantly more TEs than CR for all cytosine contexts (Figure 9C). Given that TEs demonstrate generally higher methylation levels than other genomic features, I supposed that the higher mC levels in COR were mainly due to the high composition of TE. Interestingly, however, when I compared the average mC levels of TE from CR and COR, I found TE of COR shows higher average mC levels than TE of CR, which is true for the genic regions as well (Figure 9D).

In summary, CR exhibited very similar levels of non-CG methylation between ecotypes, while relatively COR, where was supposed to be a diverse region, exhibited higher methylation level than CR in Col-0. In other words, the high mC levels in COR are not only from a larger TE composition but also from intrinsic properties retained in COR.

Figure 9

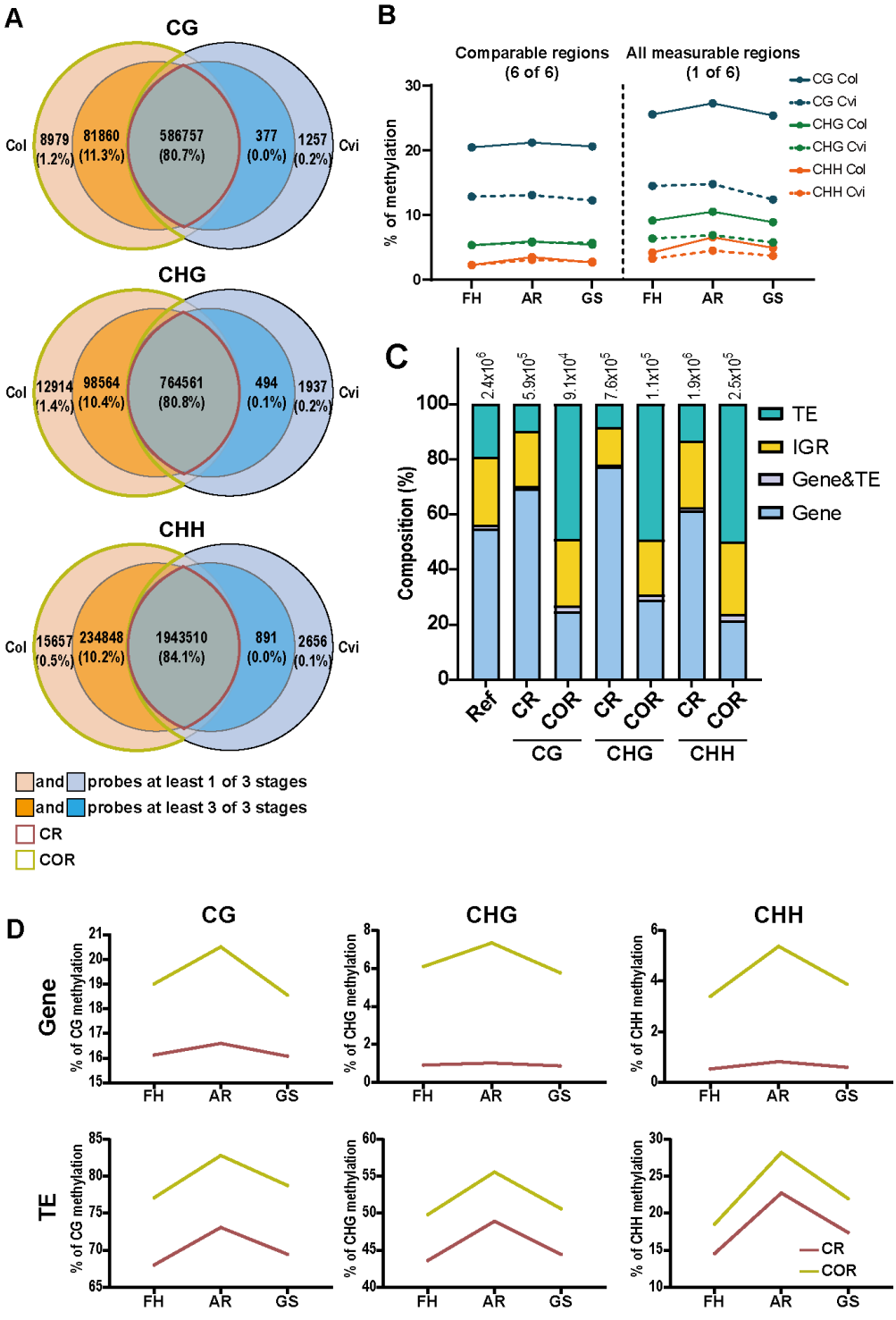
A
CG
Col
8979 (1.2%)
81860 (11.3%)
586757 (80.7%)
377 (0.0%)
1257 (0.2%)
Cvi
CHG
12914 (1.4%)
98564 (10.4%)
764561 (80.8%)
494 (0.1%)
1937 (0.2%)
CHH
15657 (0.5%)
234848 (10.2%)
1943510 (84.1%)
891 (0.0%)
2656 (0.1%)
and probes at least 1 of 3 stages
and probes at least 3 of 3 stages
CR
COR
B
Comparable regions (6 of 6)
All measurable regions (1 of 6)
% of methylation
CG Col
CG Cvi
CHG Col
CHG Cvi
CHH Col
CHH Cvi
FH
AR
GS
C
Composition (%)
2.4x10^6
5.9x10^5
9.1x10^4
7.6x10^5
1.1x10^5
1.9x10^6
2.5x10^5
TE
IGR
Gene&TE
Gene
Ref
CR
COR
CG
CHG
CHH
D
Gene
TE
% of CG methylation
% of CHG methylation
% of CHH methylation
CR
COR

Figure 9. Hypermethylated Col-0 own regions (COR) where was sufficiently mapped in Col-0 but not in Cvi.

A Venn diagrams for describing how many windows could be analyzed in each ecotype. All numbers indicate the number of 50 bp windows. The windows in the dark red colored intersection (CR) were read in 6 samples and regions, and the windows where only were read enough more than one stage in Col-0 but not in Cvi were colored in lime (COR); those two groups were compared in other analysis (b–d). **B** Average methylation according to developmental stage in CR (left side) and all measurable regions considered a probe in more than one stage. **C** The composition ratio of genomic features in both CR and COR was compared. The number of each group is written above each bar. Ref, the composition of TAIR10 genome. **D** Averages of CR and COR methylation level of Col-0 according to developmental stages were compared in gene and transposable elements.

2.3.4 The correlation between mC levels and SNPs in COR

Since the COR was estimated to have a difference sequence between Col-0 and Cvi, indirectly or directly resulting in higher methylation, I considered single nucleotide polymorphism (SNP) and structural variants as a candidate. However, when I divided CORs into two groups depending on the presence of SNPs or not, the number of windows with SNPs was much lower than those without SNPs in COR (Table 2), suggesting other factors beyond SNPs.

Despite fewer SNPs in COR, I analyzed the relationship between the number of SNPs and their DNA methylation levels, if any, in COR and in CR for a comparison. To do this, I used methylation levels of the Col-0 and the SNP information identified by a comparison between Col-0 (TAIR10) and Cvi (Ver.2) reference sequences [22]. The analysis clearly demonstrated that COR decreased in mC levels, whereas the CR exhibited an increase in mC levels as the number of SNP increased, regardless of cytosine contexts and stages (Figure 10). If the number of SNPs directly correlates with DNA methylation levels, it would show the same pattern in both regions. Since my results showed an anticorrelation of mC with the number of SNPs in COR but a positive correlation in CR, only the number of SNPs does not affect mC in COR and in CR.

Table 2. The number of windows including SNP per cytosine context in COR

Single Nucleotide Polymorphism (SNP) was decided by comparing Col-0 and Cvi genome sequence. The size of the windows is 50 bp without overlap. The percentage in the bracket is based on the total number of COR for each cytosine context.

CYTOSINE CONTEXT	CG	CHG	CHH
TOTAL NO. OF WINDOWS	90,839 (100%)	111,478 (100%)	250,505 (100%)
NO. OF WINDOWS WITH SNP	23,232 (25.6%)	25,137 (22.5%)	62,685 (25.0%)
NO. OF WINDOWS WITHOUT SNP	67,607 (74.4%)	86,341 (77.5%)	187,820 (75.0%)

Figure 10

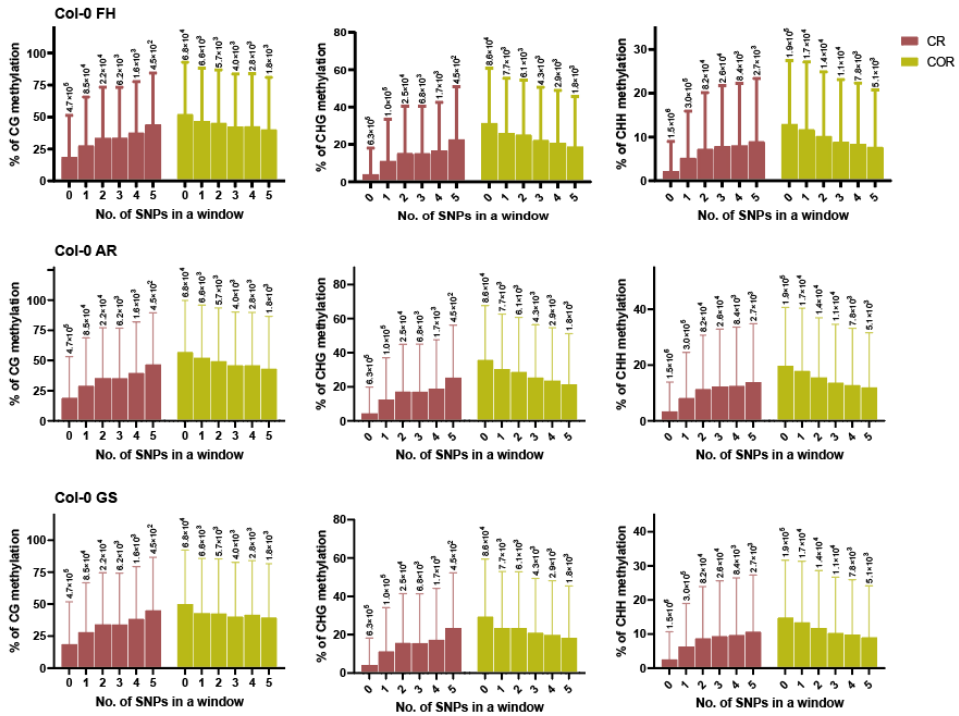

Figure 10. Correlation between SNPs and methylation level of CR and COR.

The correlation between the number of SNPs and DNA methylation is analyzed. Numbers in X-axis represent the number of SNPs in a 50-bp window. The height of the box, mean; the length of error bar, std. Numbers above an error bar indicate counted regions.

Next, I tested the correlation between mC and the number of SNPs separately for gene and TE only at FH stage. This revealed that the correlations shown in each group resulted from different genomic features. The positive correlation in CR was contributed not only from TE, but also from gene that comprises at least 60% of the CR genomic features in all cytosine contexts (Figure 11A and Figure 9C). It could be assumed that regardless of the type of genomic feature, the less changed sequence, the low methylated. However, the anticorrelation detected in COR was mainly from gene, although its portion in COR is less than 30% (Figure 11A and Figure 9C). On the contrary, although TE constitutes the most significant genomic portion (approximately 50%) in COR, it does not significantly contribute to the anticorrelation between mC levels and the number of SNPs. In conclusion, our analyses reveal that number of SNPs affects mC levels differently on CR and COR.

Further, I investigated whether any possible link exists between mC level and a specific substitution. I defined 50 bp windows, including any SNP in CR and COR, and then normalized methylation levels respectively with the average and standard deviation. My analysis revealed that the windows, including one or more of the four transition mutations (AG, CT, GA, and TC), are more methylated than other substitutions (Figure 11B). Notably, the mC levels of four transition mutations are very similar in CR, but only AG and TC transition mutations are more methylated in COR, among other transitions or transversions (Figure

11B). Therefore, it is tempting to speculate that AG and TC mutations might be either a cause or consequence of the high methylation levels in this region. Alternatively, COR might reside in a more heterochromatic region of a chromosome than CR, thereby tending to be more methylated through AG and TC mutations over time. Indeed, I found that CR regions are located throughout the chromosome except in the pericentromeric region, whereas COR is located not only in the chromosomal arm but also even more in the pericentromeric region (Figure 11C).

Figure 11

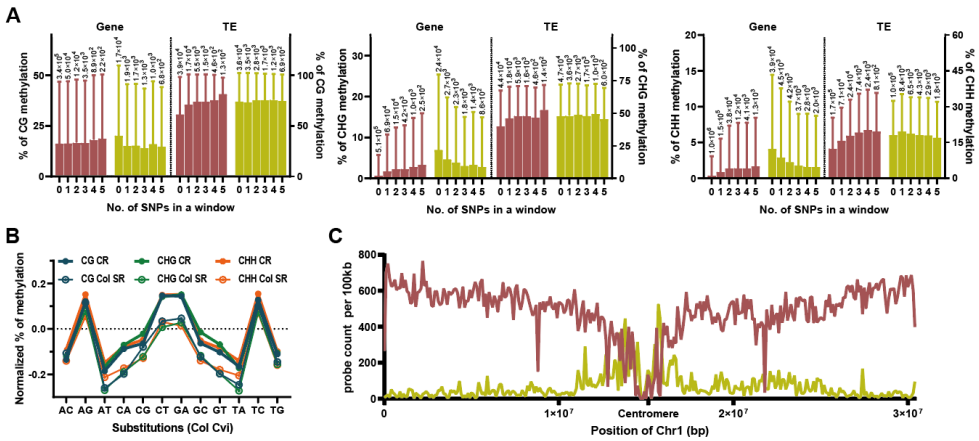

Figure 11. Correlation between SNPs and methylation level on gene and TE and position of CR and COR.

A The windows overlapped in gene and transposable element (TE) from each CR, and COR was analyzed. The positive correlation of CR was strongly demonstrated in TE, and the negative correlation of COR was not shown in the windows overlapped with TE, but it was shown in the windows overlapped with gene. The height of the bar, mean; the length of error bar, std. **B** Methylation level according to each substitution, obtained by comparing Col-0 and Cvi (the first letter is from Col), was normalized with mean and standard deviation ($\frac{x-m}{\sigma}$). **A**-**B** Methylation level of Col-0 FH was used as a representative. **C** The position of CR and COR was analyzed in chromosomal view.

2.3.5 Ecotype-specific structural variations in COR are hypermethylated

Since the number of SNPs does not explain the high mC levels in COR, I analyzed structural variances (hereafter, SVs) within COR by comparing TAIR10 and the de novo Cvi genome using a whole-genome sequence comparison tool (MUMmer, see Methods). SVs were categorized and described in an aspect of Col-0 into six groups and the other unclassified SVs (Figure 12A left); a deletion (DEL) which is a type of gap, an insertion (INS) which is a type of gap, a duplicated sequence (DUP), an insertion (BRK) which is not found in Cvi, an inverted sequence (INV), and a translocation event (SEQ). For Col-0 whole-genome categorization into the 6 groups above, 89% belongs to the Unclassified followed by DUP, comprising 5.5% of the genome (Figure 12A right 3 panels). As expected, CR contains very few structural variances, especially for the regions where SNPs are low (Figure 12B). In contrast, distinct structural features such as DUP, BRK, and INS are more enriched in COR (Figure 12B). Interestingly, more structural variants are included in COR when no SNP exists, and then, as SNPs increases, only the DUP and Unclassified classes are found (Figure 12B).

Figure 12

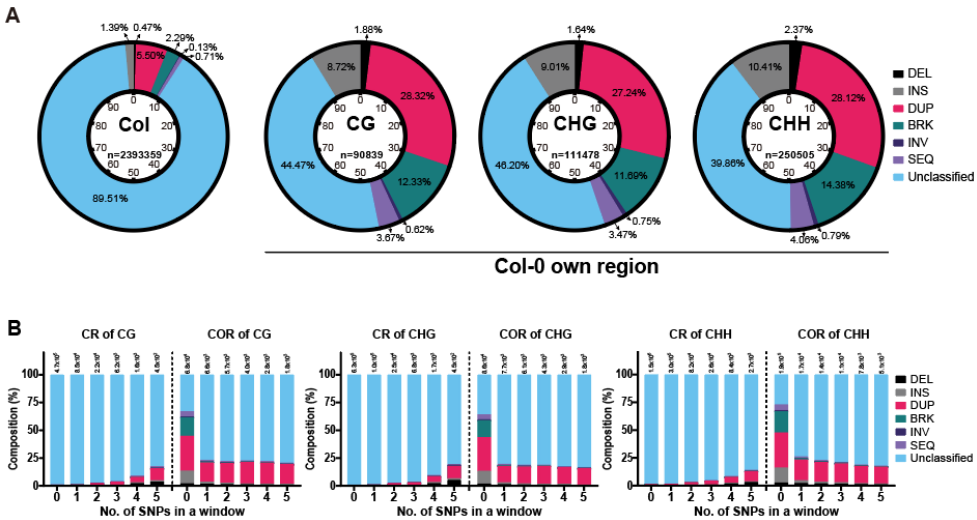

Figure 12. Enriched structural variants (SV) in COR.

A Structural variants (SVs) were analyzed for COR and compared the composition of SVs in COR with that in reference genome. The proportion of each SV was indicated as %. Unclassified are the regions not included in other SVs. "Col" indicates the composition of Col-0 ref genome. **B** The composition of each SV from CR and COR according to the number of SNPs was analyzed.

To confirm that these structural variants could contribute to the high mC levels in COR, I confirmed the methylation levels in SVs. DUP showed strikingly higher mC levels than other variances (Figure 13). BRK, SEQ, and INV generally showed high mC levels. Interestingly, INS was in COR with the 3rd higher composition among SVs, but its mC level was even lower than the mean of mC levels of COR. Notably, the mC levels of SEQ contribute to a high mC level in COR, and even its composition ratio was lower than INS (Figure 13).

Since COR included many SVs with high DNA methylation levels, especially DUP and BRK which were additional sequence compared to Cvi showed hypermethylation, and it was known that those gained sequence induces hypermethylation [23, 24], I speculated RdDM regulation underlying methylation difference in COR. To confirme RdDM dependency on COR contributing high mCHH levels, publicly available ago4/6/9 and drm1/2 mutant methylome at mature embryos were used [21]. These results proved that between CR and COR where both regions displayed RdDM-dependent methylation lost for all cytosine contexts by these mutants, COR was more affected than CR (Figure 14). In detail, four types of SVs shown hypermethylation in COR (DUP, BRK, INV and SEQ in Figure 13) were most affected.

Taken together, along with a previous implication on the relationship between sequence variation and differential methylation [10], our

analyses concordantly suggest that hypermethylation in COR compared to CR results from, at least partially, the sequence and structural differences from Cvi and is regulated by RdDM pathway.

Figure 13

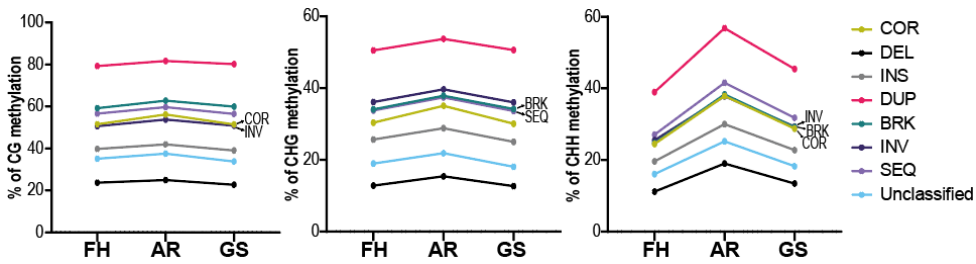

Figure 13. The DNA methylation levels and changes of SVs.

DNA methylation dynamics of each SV during seed ripening and germination were analyzed. The average level of methylation is represented as dots and linked with a line.

Figure 14

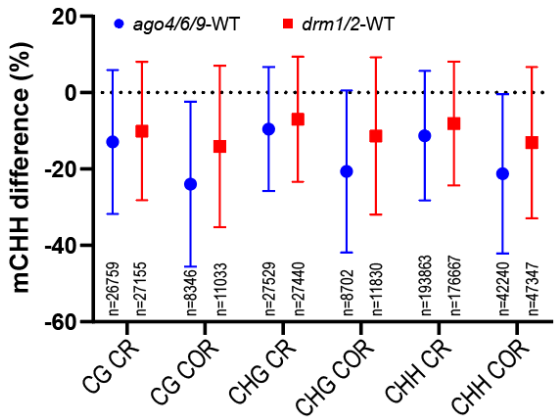

Figure 14. Changes of DNA methylation on CR and COR in RdDM mutants.

CHH methylation of *ago4/6/9* and *drm1/2* compared to wild type. Methylation levels of CORs were more affected than those of CR by RdDM mutants.

Figure 15

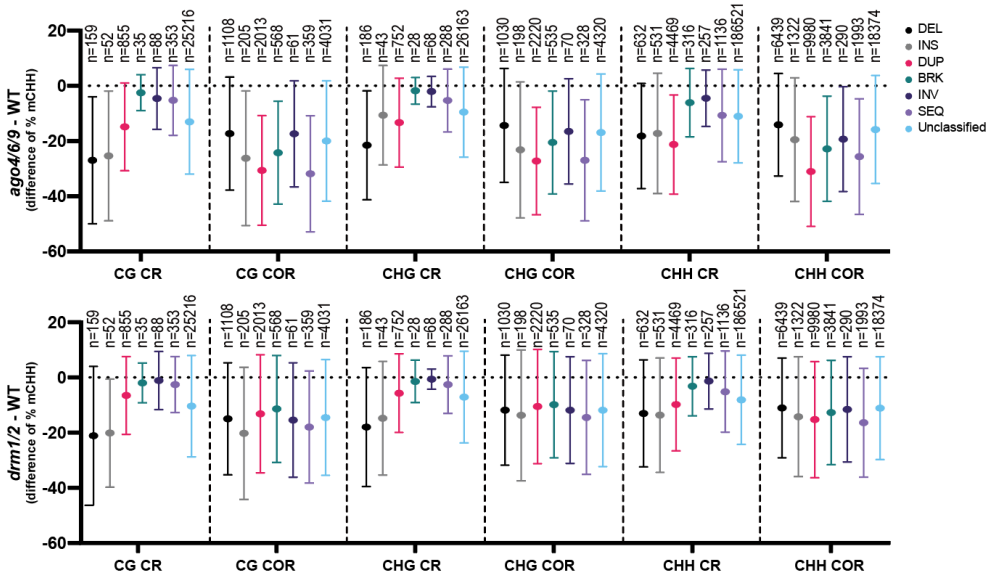

Figure 15. Changes of DNA methylation on CR and COR overlapped with SVs in RdDM mutants.

CHH methylation differences between wild type and RdDM machinery mutants in SV-overlapping CR and COR. Sequence-gain type SVs in COR, DUP and BRK, were more hypomethylated than those in CR in mutants.

2.4 Discussion

2.4.1 DNA methylation during seed ripening and germination.

During double fertilization most of CHH methylation are lost substantially, and re-gained in the embryo and reaches to the highest levels at seed maturation and desiccation, which corresponds to my AR stage (Figure 4) [18-20]. In my results, eDMRs for all methylation contexts overlap with each other, implying that de novo mCHH levels and 24-nt small RNAs levels are inter-connected with the previously methylated mCG and mCHG levels. Since symmetric methylation levels, especially CG context, remain constant, unlike mCHH methylation during reproduction, it will be interesting if the previously methylated symmetric sites play a role in generating and guiding 24-nt small RNAs-mediated de novo methylation.

2.4.2 Ambiguous cause and function of CG methylation levels.

I showed that global non-CG methylation levels in the CR are almost the same between two ecotypes for all three stages (Figure 9B). In contrast, an approximately 8% difference in mCG levels between the two ecotypes is maintained throughout development (Figure 4 and 9B). Genes were highly enriched (more than 65%) in the CR as compared to genes in whole reference genome (56%) or even in COR (20%) (Figure

9C), suggesting that CG gbM levels are substantially controlled in an ecotype-specific manner and, therefore, it seems to be important for different ecotypes living in different niches. It has been known that thousands of genes have gbM in the CG context preferentially in the exons with very low levels of non-CG methylation [1, 25-27]. However, the function of gbM is still bizarre. Mutation in mCG maintenance methyltransferase1, MET1, does not alter the major steady-state gene expressions in Arabidopsis [25, 26]. Moreover, variation in gbM in natural Arabidopsis populations is not likely correlated with gene expression [10]. Nonetheless, gbM is most frequently found in constitutively expressed genes, such as housekeeping genes, and least present in variably expressed genes [1]. Consistent with other reports [10, 28], gbM, in this study, remained constant depending upon ecotypes across the developmental process. The most plausible function of gbM is thought to be homeostasis; gbM possibly prevents aberrant transcription from internal cryptic promoters inside genes, thereby stabilizing gene expression from its own promoter. Or, gbM might enhance splicing efficiency [29]. The cause-and-effect of gbM is yet to be further demonstrated.

2.4.3 COR was not lost in Cvi, but acquired in Col-0.

TEs introduced into an organism are silenced epigenetically as part of

a genome defense system. However, TEs undergo accumulation of genetic mutations over time on top of epigenetic modifications and lose their transposition activity. Therefore, older TEs tend to be relatively short, and lose mobile activity, and have less intact sequences, compared to younger TEs. Additionally, based on a previous study, mobile, young TEs are more hypermethylated than non-mobile TEs [30]. DNA sequences of CR were relatively conserved in both ecotypes compared to those of variable COR. From the evolutionary point of view, the conserved CR were assumed to be inherited from a common ancestor, leading me to hypothesize that TEs in CR are older than COR TEs. To test this, I compared TE size and methylation levels in two regions. Despite the large variance, the average TE size of CR is significantly shorter than that of COR (Figure 16A). Furthermore, mC levels of CR are significantly lower than those of COR, regardless of cytosine contexts (Figure 16B). This supports my hypothesis that TEs in ecotype-specific regions could be introduced relatively more recently than TEs in CR. However, the chromosomal location of the two groups of TE can also explain, at least in part, the size and methylation levels. I provided that TEs from 'Common' regions are away from the centromere and are located more in chromosomal arms, whereas 'COR' are highly enriched in pericentromeric regions (Figure 11C). Perhaps, all these aspects, such as age, size, location, and mC levels, are interconnected elaborately to silence TE expression and transposition

for genome integrity.

Figure 16

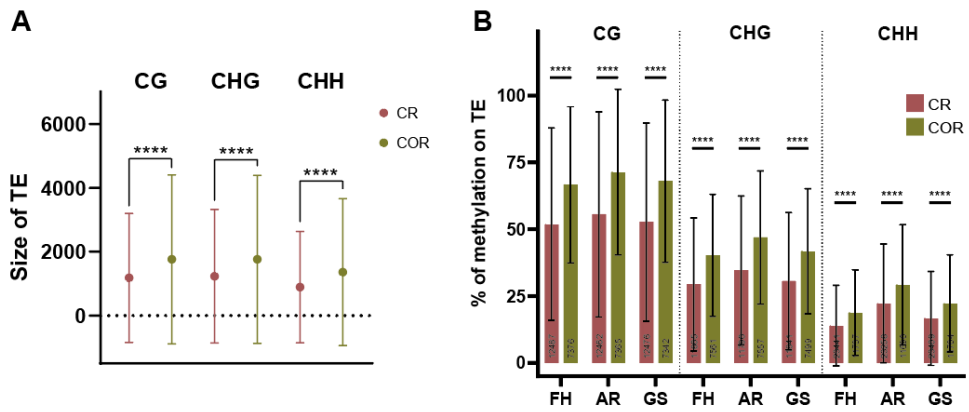

Figure 16. Changes of DNA methylation on CR and COR overlapped with SVs in RdDM mutants.

CHH methylation differences between wild type and RdDM machinery mutants in SV-overlapping CR and COR. Sequence-gain type SVs in COR, DUP and BRK, were more hypomethylated than those in CR in mutants.

2.4.4 DNA methylation of newly acquired gene.

Intriguingly, I found a positive and negative correlation between mC levels and the number of SNPs in CR and COR, respectively (Figure 10 and Figure 11). It is tempting to speculate that a selective pressure might be applied in this evolutionary conserved CR, so that as SNPs increase, mC levels also increase, regardless of gene or TE. Given that the amino acid sequences of methylated genes likely evolve more slowly than those of unmethylated genes , it is plausible that increasing mC levels help to keep the amino acid sequences unchanged, as SNPs and mutations accumulate during evolution in genes of CR. A recent interesting report suggested that old long terminal repeat (LTR) TEs with an accumulation of mutations over the time could be transcribed again, leading to a secondary RdDM pathway . Thus, mC levels of old TEs in CR may increase to silence TE transcription generating again from the accumulated mutations. By contrast, COR shows anticorrelation of mC levels with the number of SNP, and this anticorrelation is mainly from gene (Figure 10 and Figure 11A). This is interesting because genes in CR and COR show opposite patterns for mC levels to the number of SNP. Specifically, the mC level decrease is striking from SNP 0 to SNP 1 in COR (Figure 11A). This result prompts me to think that genes with newly obtained SNPs in COR lose methylation rapidly, which can alter gene transcription, and this might contribute to a rapid evolution of genes in COR, at least in part, compared to genes in CR.

III. DNA methylation during embryo development

3.1 Introduction

Seed plants have comprehensive DNA methylation system with multiple associated enzymes. Arabidopsis methylation has several representative and unique features such as robust transgenerational maintenance of CG methylation particularly on gene body [26, 31, 32], global demethylation and imprinting in central cell and endosperm [33-35]}, and RNA directed DNA methylation (RdDM) [36], and these have been extensively studied. Especially, RdDM pathway is known to be closely connected with TE silencing [36]. Interestingly, de novo methylation is targeting not only the heterochromatic TEs but also the euchromatic TEs with separate regulatory pathways. While CHROMOMETHYLASE2 (CMT2)-dependent DNA methylation is concentrated on heterochromatic TEs with its binding activity to H3K9 dimethylation via chromo domain [17], DOMAINS REARRANGED METHYLTRANSFERASE 1/2 (DRM1/2)-mediated de novo methylation is targeting euchromatic TEs [36]. Since CHH methylation has to be regulated by de novo pathway in every cell division cycle due to the absence of maintenance mechanism [26] and CHG [2], mCHH levels can be actively changing by expression of associating enzymes and small

RNAs of TE. Indeed, in contrast to the constant levels of CG methylation, CHH methylation is cell-type specific and tissue specific, thereby, dynamically changed as plant develops [6, 10, 19, 20, 31, 35, 37]. Therefore, the spatiotemporal methylomes are strongly demanded to understand the dynamics of TE activity with corresponding silencing mechanism. Indeed, recent studies of methylome during seed maturation and germination show gain and loss of CHH methylation respectively [19, 20]. Furthermore, Papareddy et al. showed the gradual CHH methylation and corresponding small RNAs of TE simultaneously during the early embryogenesis [6].

Although the overall patterns of methylation change during embryogenesis have been reported, in-depth analyses for the regions where methylation levels are changing have yet to be examined. In this study, I report the changes in global DNA methylation during five stages of embryogenesis. By using optimized methods to extract DNAs from each stage, whole genome bisulfite sequencing (WGBS) libraries were generated with at least two replicates. Global methylation levels of CG and CHG were quite stable. However, the majority of change in DNA methylation occurred in CHH context, and the most significant increase was observed particularly in TE. Then, I clustered genomic regions into nine groups based on the methylation patterns that were changed similarly during embryo development. Genomic features of those nine clusters during embryo development for each cytosine methylation

(mC) context were also examined. Using predicted intact LTR, I discovered and suggest that the frequencies of mC contexts on TE likely converge at a certain point as time goes on.

3.2 Materials and methods

3.2.1 Plant growth condition

Arabidopsis thaliana, Col-gl (Columbia-glabrous) ecotype was used. Plants were grown on soil in an environmentally controlled room at 22°C under long photoperiods (16-h light/8-h dark) with cool white fluorescent light (100 μmole/m2/s). After 24 h of emasculation (Park et al., 2016), fully matured stamens were picked from open flowers with tweezers. Pollen was rubbed onto the emasculated stigma for fertilization, under a dissection microscope. Pollinated plants were incubated in the growth room until they reached a suitable stage for sampling; 4 days after pollination (DAP4) for globular, DAP5 for heart, DAP7 for torpedo, DAP9 for bending torpedo, and DAP12 for mature green stage embryos.

3.2.2 Construction of bisulfite sequencing libraries

All procedures from DNA extraction to construction of WGBS library

were conducted as described in previous paper [38].

3.2.3 WGBS data processing and DNA methylation level

All Sequencing procedures were performed using the HiSeqXten platform (Macrogen, Korea). Paired end reads (150 bp) were generated. All reads were trimmed (10 bp for 5'end and 5 bp for 3'end) using Trim galore (options, --clip_R1 10 --three_prime_clip_R1 10), and low quality and short reads (<70 bp) were removed using Trimmomatic (options, SE -threads 16 SLIDINGWINDOW:2:20 MINLEN:40). Reads were mapped to the Arabidopsis TAIR10 genome by Bismark with hisat2 under the option –hisat2 –local. PCR duplicates were removed and methylation levels were extracted using the Bismark scripts with default option (deduplicate_bismark and bismark_methylation_extractor, respectively). Methylation levels were basically calculated by dividing the counts of methylated cytosine by the number of cytosine (meC + C) per single cytosine, then methylation levels included in a 50-bp window were averaged. Furthermore, the windows including more than 3 cytosine sites for each mC context were only analyzed.

3.2.4 clustering

The methylation levels of a window were changed to Z-scored value with mean and standard deviation ($\frac{x-m}{\sigma}$; x, mean; m, average; σ, standard deviation). Whole windows with methylation levels in all stages and with standard deviation (>0) were used for estimating a model using mclust which is an R package for normal mixture modeling. Clustering was conducted with best model and the number of optimal clusters.

3.2.5 genomic features of clusters

All windows were annotated with overlapping gene or TE. If a window was annotated with gene and TE, the window was classified as 'Gene&TE'. In addition, if a window does not belong to any genomic feature, it is classified as 'IGR (intergenic region)'.

3.2.6 Distributions of each cluster on chromosome

Relative frequencies were calculated with window counts within a 100kb bin by a cluster. Therefore, the sum of Y-axis value is 100.

3.2.7 cytosine frequency of predicted de novo LTR

Using previously reported predicted intact LTRs [39], the number of

mC contexts on a predicted intact LTR were counted. Frequencies of context were calculated by dividing counted number of mC context with the LTR length.

3.3 Results

3.3.1 Dynamic change of CHH methylation during embryo development.

To reveal DNA methylation change during embryo development, my collaborator KH park developed the optimizing method of methylome construction for Arabidopsis embryos of five developmental stages as minimizing the input material by adapting two previously reported protocols of mammalian single-cell methylome construction [38, 40, 41]. The five stages are; globular, heart, torpedo, bending torpedo and mature green stages which correspond to 4, 5, 7, 9, and 12 DAP, respectively (Figure 17). These stage-specific bisulfite sequencing libraries were constructed by using the optimized protocol [38] (Figure 17). All the libraries were made up with more than 2 biological replicates for each stage with robust CT conversion (>99%) and genome coverage (>40X) and this sufficiently satisfies the recommended standard of methylation data of epigenetic consortium (Figure 17) [42].

I found that gradual increase of global CHH methylation particularly at pericentromeric region was clearly seen during embryo development and this was largely due to the mCHH increase in TE (Figure 18 and Figure 19). This is also fairly corresponding to the expression dynamics of small RNAs of TE during embryo development [6]. Accompanying expression profile of related genes for non-CG methylation for TE [6]

and accumulation of repressive histone marks, H3K9me2 [21], and my data imply the global change of epigenetic landscape during embryogenesis. In contrast, all cytosine contexts of gbM was static with only slight increase of CHH methylation. The fact of that genebody usually has mainly CG methylation and the strong maintenance mechanism of mCG might lead to static methylome dynamics during embryogenesis [26, 32, 43]. Notable aspect of embryo methylome dynamics of TE was that the increase of methylation was shown mainly at CHH context without corresponding change of CG and CHG methylation. On the other hand, CHH methylation has its own specific pathway involving various enzymes of small RNA biogenesis and processing with DNA polymerase variants [36, 44]. Furthermore, it has relatively weak feedback mechanism with those factors comparing to CG and CHG methylation [36, 44]. Therefore, the CHH methylation pathway could be operated by their own spatiotemporal manner separately with CG and CHG methylation. In addition, it was shown that the mCHG levels on TE decreased after heart stage [45] and my data recapitulated that pattern (Figure 19), which is quite similar to the expression patterns of CMT3, IBM1 and KYP [45]. Therefore, this implies that there would be changes in chromatin before and after the heart stage.

Figure 17

Sample	Globular (DAP4)	Heart (DAP5)	Torpedo (DAP7)	Bending torpedo (DAP9)	Mature green (DAP12)
Number of libraries	7	3	6	2	5
Total coverage (X)	41	44	61	44	70
Average of CT convertion (%)	99.2	99.5	99.5	99.1	99.4

Figure 17. Information of whole genome bisulfite sequencing (WGBS) libraries used in this study.

The day pollinated was counted as 1 days after pollination (DAP1). Coverage was calculated by dividing the number of cytosine counts from mapped reads after PCR deduplication with the number of cytosine counts in Arabidopsis TAIR10 genome. Plastid methylation levels were considered as unconverted CT level.

Figure 18

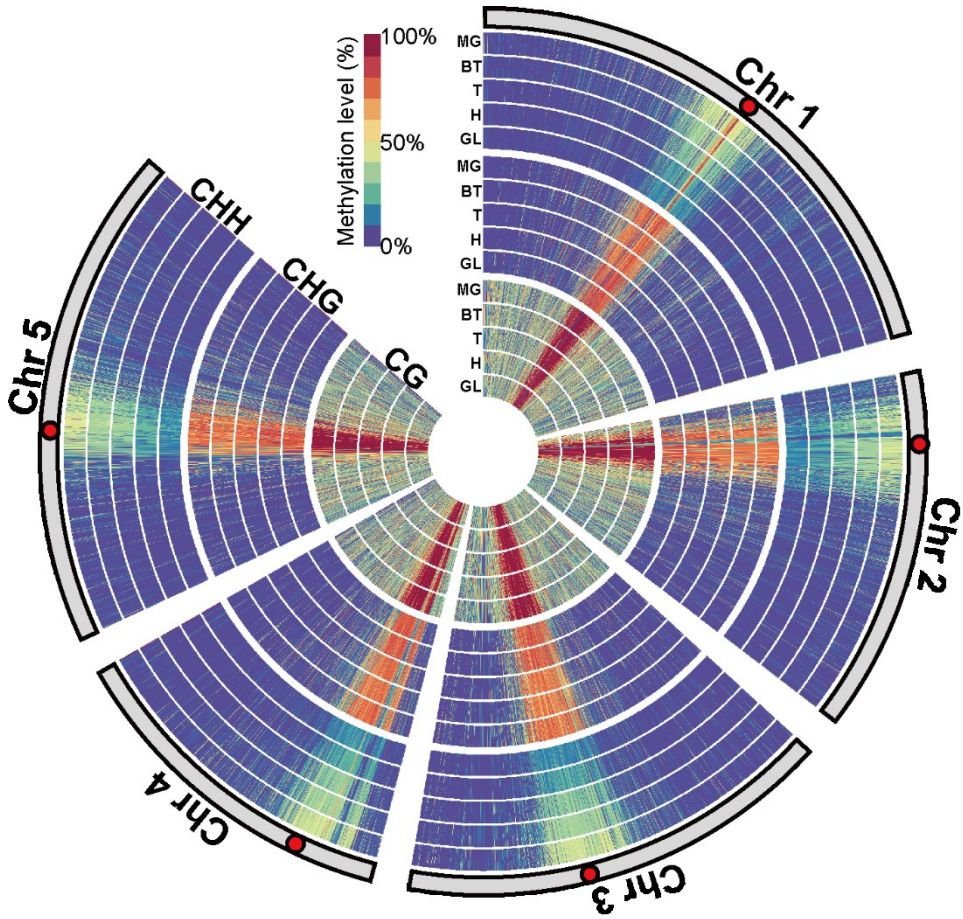

Figure 18. Heatmap of DNA methylation during embryo development.

Five stages were used for confirming the changes of DNA methylation. It is filled from the inner circle in order of development. GL, globular; H, heart; T, torpedo; BT, bending torpedo; MG, mature green.

Figure 19

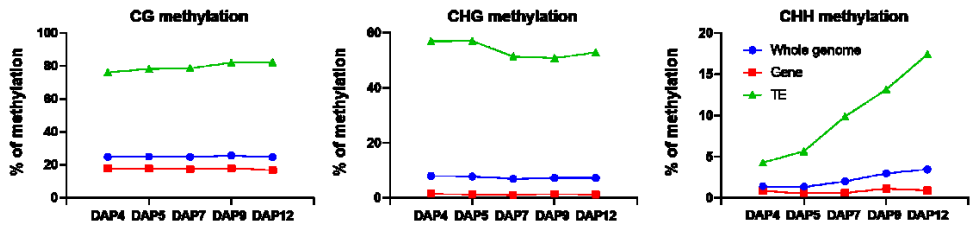

Figure 19. Global DNA methylation levels of genomic features.

Averages of DNA methylation levels of windows were shown by each genomic feature.

3.3.2 The regions with different patterns of change during embryo development

My results demonstrated the overall change in DNA methylation during embryo development in TE and genes (Figure 18 and Figure 19). However, genome is not covered by genes and TEs only, but it contains more than those including intergenic sequences controlled by DNA methylation. Therefore, it needs to be analyzed in more detail.

To divide genome in detail, I assumed that the regions with similar properties would be regulated in a similar way. Accordingly, I aimed to categorize 50-bp windows with the changes in DNA methylation with similar changing patterns. First of all, DNA methylation levels of a window was converted to Z-score of the window (see details in Materials and methods). After that, the windows were classified according to their Z-scored value by using mclust (normal mixture modeling tool of R package). All of mC contexts were to be classified into 9 classes (Figure 20). Based on those results, I obtained 9 classes with different dynamics of DNA methylation. As results, unexpectedly, there are groups with different patterns of change in all contexts (Figure 21). This result suggests that various regulatory mechanisms for the methylation change during embryogenesis are operating in a spatiotemporal manner. In addition, the most common pattern of mCG change was the class 1 where the DNA methylation levels are

continuously increasing (Figure 21). This cannot be seen using analysis of the global methylation pattern (Figure 19). Therefore, my analysis could highlight the discovery of the new findings.

Figure 20

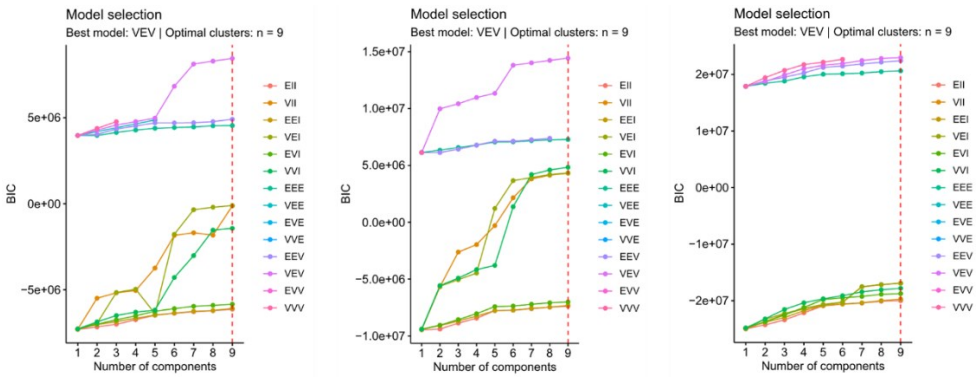

Figure 20. The result of normal mixture modeling.

Normal mixture modeling was conducted with mclust (R package). The model and the number of clusters were recommended.

Figure 21

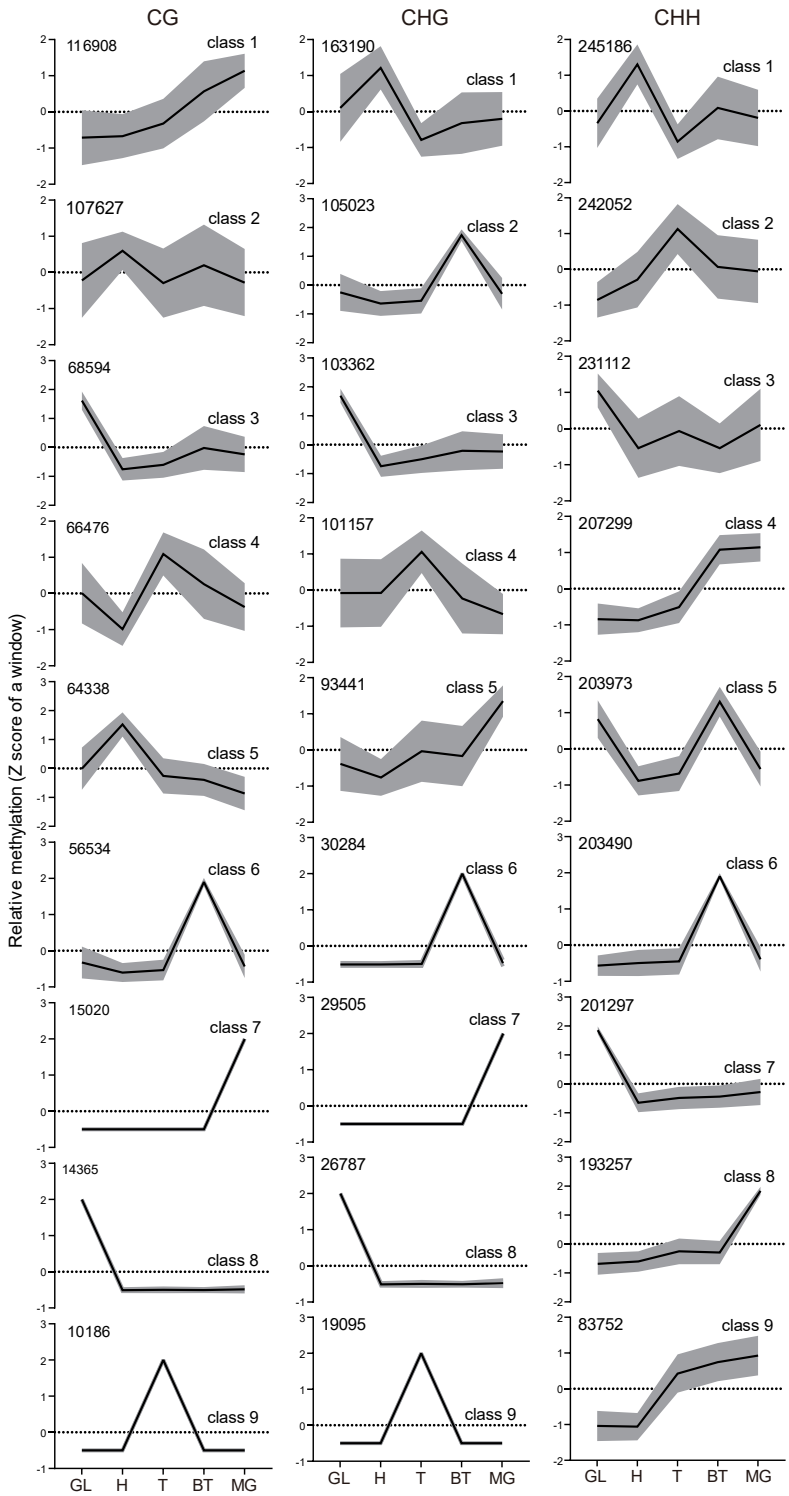
CG
CHG
CHH
Relative methylation (Z score of a window)
116908
class 1
163190
class 1
245186
class 1
107627
class 2
105023
class 2
242052
class 2
68594
class 3
103362
class 3
231112
class 3
66476
class 4
101157
class 4
207299
class 4
64338
class 5
93441
class 5
203973
class 5
56534
class 6
30284
class 6
203490
class 6
15020
class 7
29505
class 7
201297
class 7
14365
class 8
26787
class 8
193257
class 8
10186
class 9
19095
class 9
83752
class 9
GL
H
T
BT
MG

Figure 21. Changing pattern by class during embryo development.

Using the result of classification in Figure 20, the patterns of classes were traced by Z-score, not by DNA methylation levels. Black line and gray shade indicate the average and standard deviation of Z-score in a stage, respectively. The numbers in the upper left of each graph indicate the number of windows included in a class. GL, globular; H, heart; T, torpedo; BT, bending torpedo; MG, mature green.

3.3.3 The composition of genomic features and distribution on chromosome by a class

During embryo development, my analysis discovered the various classes that are changed in DNA methylation differently. In order to find out what kinds of genomic characteristics are associated with those classes regulated differently, the composition of the genomic feature and the location on chromosome for each class were analyzed.

I expected that a class with dynamically changing in DNA methylation would include a lot of TEs relative to genes. However, genes were more included in all classes except for CHH class 9 that shows gradual increase during embryogenesis (Figure 22). That result suggested that genic methylations dynamically changed in specific tissue of embryo, because I judged the changes with a Z-score value. In addition, classes having a lot of TEs such as class 1 of CG and class4/8/9 in CHH showed the increase of DNA methylation (Figure 22 and Figure 21). DNA methylation levels in class 1 of CG was differently increased from those in class 4/8/9 of CHH. In other words, there is a possibility of regulation by different pathway. Class4/8/9 of CHH showed a rapid increase of DNA methylation at a different stage (Figure 21). Therefore, it suggests that change of DNA methylation is affected not only by genomic features, but also by other properties.

Figure 22

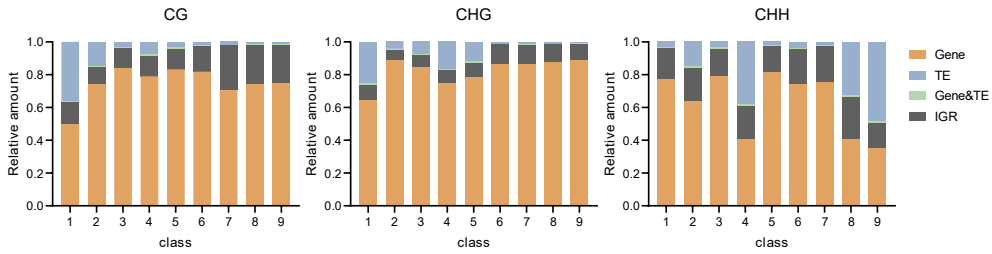

Figure 22. The composition of each genomic feature by a class.

Each composition of genomic features was calculated with the counts of windows.

I hypothesized that the position on chromosome was one of those properties. To examine the position of each class on chromosome, I analyzed frequencies of distribution of each class by using windows in chromosome 1 as a representative. As a result, it seemed to be largely divided into two categories, one was concentrated in pericentromeric region and the other was spread throughout the chromosomal arms (Figure 23). This result was based on the result of composition of TE.

In summary, TE composition was helpful to understand the properties of each class. However, this result is insufficient to identify the characteristics of each class. Thus, further in-depth analyses are needed to reveal the more detailed dynamics of methylation change during embryogenesis.

Figure 23

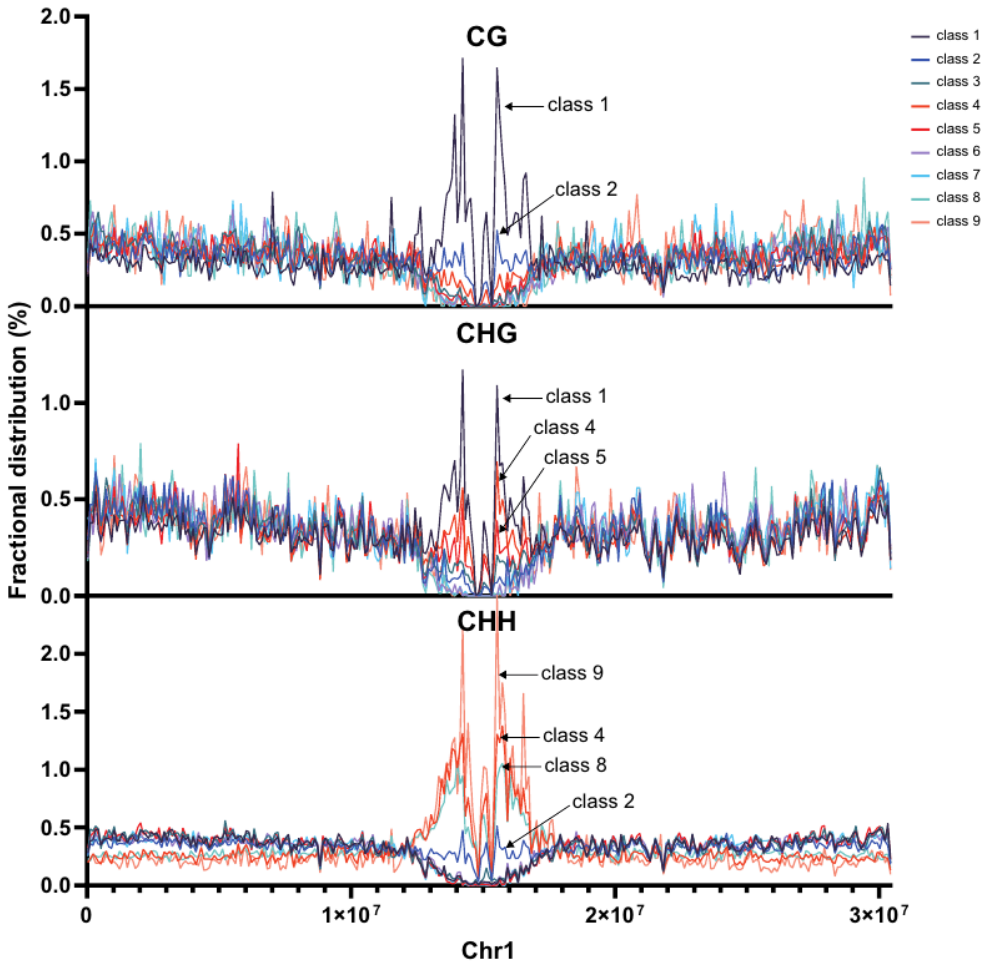

Figure 23. Positions of each class on chromosome 1.

Distribution of windows included in a class were analyzed. The count of windows in a 100 bin was converted to relative frequency (%).

3.4 Discussion

I observed gradual resetting of CHH methylation during early embryo development. This is not accompanying the corresponding resetting of CG and CHG methylation. Instead, CG and CHG methylation levels are rather high at early embryos (Figure 19). Small RNA-dependent feature and distinct associating enzymes of CHH methylation might result in such independent resetting. Indeed, expression of small RNAs and associating enzymes are correlated with the resetting of CHH methylation [6]. CHH methylation of TEs are regulated by bipartite pathways according to their location which are strongly related to various features such CG and CHG methylation, H3K9me2 and nucleosome density [6, 46, 47]. Actually, there is no clear border of heterochromatin. However, the facts that heterochromatin is loosened in early embryo was revealed recently [6]. This implies that those regions should be established as normal heterochromatin. I speculate that there is a direct correlation with such a gradual resetting of CHH methylation during embryo development. Additionally, I guess that those switch between hetero- and euchromatin in early embryo may be correlated with evolution throughout epigenetics. Previous data showed that heterochromatic TEs are usually longer and less mutated than euchromatic TE [48, 49]. I assume that highly methylated fraction of a TE may be more protected than non-methylated, a decrease in the

number of cytosine in TE at heterochromatic region makes it more attacked by transposition event and has more chance to transpose to euchromatic region. Furthermore, recent data showed that young and long terminal repeat (LTR) transposons in euchromatin are regulated by CMTs in contrast to that euchromatic TEs that are usually regulated by RdDM [39]}. Longer and younger LTR transposon has more CHG methylation in their analysis [39]. Interesting correlation among length, location, cytosine ratio, DNA methylation and its regulatory pathway of TEs are suggesting their trajectory of evolution. The study of the process and mechanism of each correlation will decipher the evolution of TE inside of the host plant.

Based on these, I also assumed that a TE whose transposition activity is low should have been changed in their DNA sequences. In particular, genetic mutation is more frequently generated by DNA methylation [50]. To confirm this, I obtained predicted de novo LTR and their LTR sequence similarity [39]. And then, frequencies of mC contexts of predicted de novo LTR were calculated. The low LTR similarity is considered old since genetic mutation accumulated. Unexpectedly, intact LTRs expected to be recently introduced have diverse mC frequencies, but old LTRs which have low LTR similarity have similar frequencies of mC (Figure 24). It seems that the frequency of mC of LTR converge to a certain point over time. It is unknown whether this phenomenon is caused by DNA methylation, but it might contribute to

understanding of the change in TE sequences after introduction. Overall, my data indicate that DNA methylation levels are somehow correlated with the changed in DNA sequences over the time and an organism might use for adapting to the environment.

Figure 24

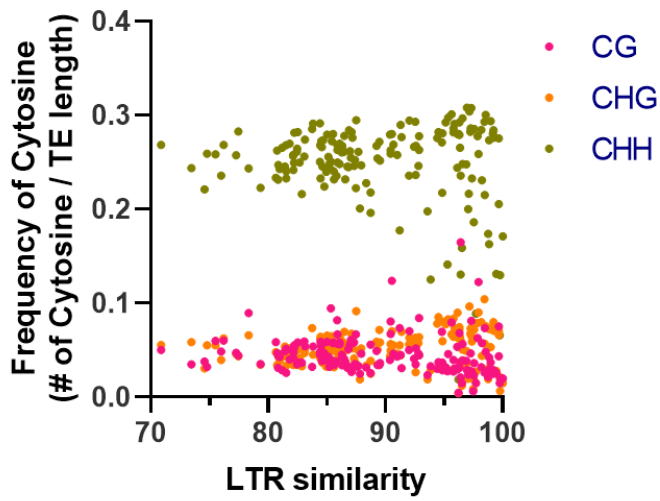

Figure 24. Frequencies of cytosine methylation contexts in predicted intact LTR.

The frequency of a mC contexts on an intact LTR was shown. Frequency was calculated by dividing the counts of each mC context on an intact LTR with its length. Intact LTR and its similarity was used from previous report [39].

3.5 References

1. Kim MY, Zilberman D: **DNA methylation as a system of plant genomic immunity.** *Trends in Plant Science* 2014, **19:**320-326.
2. Law JA, Jacobsen SE: **Establishing, maintaining and modifying DNA methylation patterns in plants and animals.** *Nature Reviews Genetics* 2010, **11:**204-220.
3. Han Q, Bartels A, Cheng X, Meyer A, An YQC, Hsieh TF, Xiao WY: **Epigenetics Regulates Reproductive Development in Plants.** *Plants-Basel* 2019, **8**.
4. Kawakatsu T, Stuart T, Valdes M, Breakfield N, Schmitz RJ, Nery JR, Urich MA, Han XW, Lister R, Benfey PN, Ecker JR: **Unique cell-type-specific patterns of DNA methylation in the root meristem.** *Nature Plants* 2016, **2**.
5. Greenberg MVC, Bourc'his D: **The diverse roles of DNA methylation in mammalian development and disease.** *Nature Reviews Molecular Cell Biology* 2019, **20:**590-607.
6. Papareddy RK, Paldi K, Paulraj S, Kao P, Lutzmayer S, Nodine MD: **Chromatin regulates expression of small RNAs to help maintain transposon methylome homeostasis in Arabidopsis.** *Genome Biology* 2020, **21**.
7. Hajheidari M, Koncz C, Bucher M: **Chromatin Evolution-Key Innovations Underpinning Morphological Complexity.** *Front Plant Sci* 2019, **10:**454.
8. Weigel D, Mott R: **The 1001 Genomes Project for Arabidopsis thaliana.** *Genome Biology* 2009, **10**.
9. Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K,

Weigel D: **Spontaneous epigenetic variation in the Arabidopsis thaliana methylome.** *Nature* 2011, **480:**245-249.

10. Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, et al: **Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions.** *Cell* 2016, **166:**492-505.
11. Kankel MW, Ramsey DE, Stokes TL, Flowers SK, Haag JR, Jeddeloh JA, Riddle NC, Verbsky ML, Richards EJ: **Arabidopsis MET1 cytosine methyltransferase mutants.** *Genetics* 2003, **163:**1109-1122.
12. Lindroth AM, Cao XF, Jackson JP, Zilberman D, McCallum CM, Henikoff S, Jacobsen SE: **Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation.** *Science* 2001, **292:**2077-2080.
13. Bartee L, Malagnac F, Bender J: **Arabidopsis cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene.** *Genes & Development* 2001, **15:**1753-1758.
14. Kawashima T, Berger F: **Epigenetic reprogramming in plant sexual reproduction.** *Nature Reviews Genetics* 2014, **15:**613-624.
15. Nuthikattu S, McCue AD, Panda K, Fultz D, DeFraia C, Thomas EN, Slotkin RK: **The Initiation of Epigenetic Silencing of Active Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs.** *Plant Physiology* 2013, **162:**116-131.
16. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D: **The Arabidopsis Nucleosome**

Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. *Cell* 2013, **153:**193-205.

17. Stroud H, Do T, Du JM, Zhong XH, Feng SH, Johnson L, Patel DJ, Jacobsen SE: **Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis.** *Nature Structural & Molecular Biology* 2014, **21:**64-+.
18. Lin JY, Le BH, Chen M, Henry KF, Hur J, Hsieh TF, Chen PY, Pelletier JM, Pellegrini M, Fischer RL, et al: **Similarity between soybean and Arabidopsis seed methylomes and loss of non-CG methylation does not affect seed development.** *Proceedings of the National Academy of Sciences of the United States of America* 2017, **114:**E9730-E9739.
19. Kawakatsu T, Nery JR, Castanon R, Ecker JR: **Dynamic DNA methylation reconfiguration during seed development and germination.** *Genome Biology* 2017, **18**.
20. Bouyer D, Kramdi A, Kassam M, Heese M, Schnittger A, Roudier F, Colot V: **DNA methylation dynamics during early plant life.** *Genome Biology* 2017, **18**.
21. Parent JS, Cahn J, Herridge RP, Grimanelli D, Martienssen RA: **Small RNAs guide histone methylation in Arabidopsis embryos.** *Genes & Development* 2021, **35:**841-846.
22. Jiao WB, Schneeberger K: **Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics.** *Nature Communications* 2020, **11**.
23. Keller TE, Yi SV: **DNA methylation and evolution of duplicate genes.** *Proc Natl Acad Sci U S A* 2014, **111:**5932-5937.
24. Remus R, Kammer C, Heller H, Schmitz B, Schell G, Doerfler W:

Insertion of foreign DNA into an established mammalian genome can alter the methylation of cellular DNA sequences. *Journal of Virology* 1999, **73:**1010-1022.

25. Roudier F, Teixeira FK, Colot V: **Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more.** *Trends in Genetics* 2009, **25:**511-517.
26. Bewick AJ, Schmitz RJ: **Gene body DNA methylation in plants.** *Current Opinion in Plant Biology* 2017, **36:**103-110.
27. Takuno S, Gaut BS: **Gene body methylation is conserved between plant orthologs and is of evolutionary consequence.** *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110:**1797-1802.
28. Kawakatsu T, Huang SS, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, et al: **Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions.** *Cell* 2016, **166:**492-505.
29. Zilberman D: **An evolutionary case for functional gene body methylation in plants and animals.** *Genome Biology* 2017, **18**.
30. Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, Colot V: **The Arabidopsis thaliana mobilome and its impact at the species level.** *Elife* 2016, **5**.
31. Hsieh PH, He SB, Buttress T, Gao HB, Couchman M, Fischer RL, Zilberman D, Feng XQ: **Arabidopsis male sexual lineage exhibits more robust maintenance of CG methylation than somatic tissues.** *Proceedings of the National Academy of Sciences of the United States of America* 2016, **113:**15132-15137.
32. Picard CL, Gehring M: **Proximal methylation features associated with nonrandom changes in gene body**

methylation. *Genome Biology* 2017, **18**.

33. Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D: **Genome-wide demethylation of Arabidopsis endosperm.** *Science* 2009, **324:**1451-1454.
34. Ibarra CA, Feng X, Schoft VK, Hsieh TF, Uzawa R, Rodrigues JA, Zemach A, Chumak N, Machlicova A, Nishimura T, et al: **Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes.** *Science* 2012, **337:**1360-1364.
35. Park K, Kim MY, Vickers M, Park JS, Hyun Y, Okamoto T, Zilberman D, Fischer RL, Feng X, Choi Y, Scholten S: **DNA demethylation is initiated in the central cells of Arabidopsis and rice.** *Proc Natl Acad Sci U S A* 2016, **113:**15138-15143.
36. Matzke MA, Mosher RA: **RNA-directed DNA methylation: an epigenetic pathway of increasing complexity (vol 15, 394, 2014).** *Nature Reviews Genetics* 2014, **15**.
37. Gutzat R, Rembart K, Nussbaumer T, Hofmann F, Pisupati R, Bradamante G, Daubel N, Gaidora A, Lettner N, Dona M, et al: **Arabidopsis shoot stem cells display dynamic transcription and DNA methylation patterns.** *Embo Journal* 2020, **39**.
38. Yoo H, Park K, Lee J, Lee S, Choi Y: **An Optimized Method for the Construction of a DNA Methylome from Small Quantities of Tissue or Purified DNA from Arabidopsis Embryo.** *Molecules and Cells* 2021, **44:**602-612.
39. Wang ZM, Baulcombe DC: **Transposon age and non-CG methylation.** *Nature Communications* 2020, **11**.
40. Luo CY, Rivkin A, Zhou JT, Sandoval JP, Kurihara L, Lucero J, Castanon R, Nery JR, Pinto-Duarte A, Bui B, et al: **Robust single-**

cell DNA methylome profiling with snmC-seq2. *Nature Communications* 2018, **9**.

41. Yu B, Dong X, Gravina S, Kartal O, Schimmel T, Cohen J, Tortoriello D, Zody R, Hawkins RD, Vijg J: **Genome-wide, Single-Cell DNA Methylomics Reveals Increased Non-CpG Methylation during Human Oocyte Maturation.** *Stem Cell Reports* 2017, **9:**397-407.
42. Ziller MJ, Hansen KD, Meissner A, Aryee MJ: **Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing.** *Nature Methods* 2015, **12:**230-+.
43. Hofmeister BT, Lee K, Rohr NA, Hall DW, Schmitz RJ: **Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation.** *Genome Biology* 2017, **18**.
44. Raju SKK, Ritter EJ, Niederhuth CE: **Establishment, maintenance, and biological roles of non-CG methylation in plants.** *DNA Methylation* 2019, **63:**743-755.
45. Papareddy RK, Paldi K, Smolka AD, Huther P, Becker C, Nodine MD: **Repression of CHROMOMETHYLASE 3 prevents epigenetic collateral damage in Arabidopsis.** *Elife* 2021, **10**.
46. Zhao YY, Chen XM: **Non-coding RNAs and DNA methylation in plants.** *National Science Review* 2014, **1:**219-229.
47. Zhang Y, Harris CJ, Liu QK, Liu WL, Ausin I, Long YP, Xiao LD, Feng L, Chen X, Xie YB, et al: **Large-scale comparative epigenomics reveals hierarchical regulation of non-CG methylation in Arabidopsis.** *Proceedings of the National Academy of Sciences of the United States of America* 2018, **115:**E1069-E1074.
48. Quesneville H: **Twenty years of transposable element**

analysis in the Arabidopsis thaliana genome. *Mob DNA* 2020, **11:**28.

49. Ahmed I, Sarazin A, Bowler C, Colot V, Quesneville H: **Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis.** *Nucleic Acids Research* 2011, **39:**6919-6931.

50. Poulos RC, Olivier J, Wong JWH: **The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes.** *Nucleic Acids Research* 2017, **45:**7786-7795.

V. Abstract in Korean

국문초록

DNA 메틸화는 사이토신 염기의 5번 탄소에 생긴 변형을 의미한다. DNA 메틸화는 DNA 서열을 바꾸지 않고 유전자와 전위인자 활성을 변화시킬 수 있는 에피유전학적 요인 중 하나이다. 또한 유기체가 그들의 환경에 적응하는 것을 돕는다. 난세포와 융합된 수배우자는 이배체 배를 형성하고, 다른 수배우자는 이배체 중심세포와 융합되어 삼배체 배유를 형성한다. 배유는 발달 중인 배에 영양을 공급하기 때문에 포유류의 태반과 비슷한 기능을 한다. 배와 달리 배유는 유전체가 다음 세대로 전달되지 않는다는 점에서 분화된 동반 조직이다. 배와 배유는 종자껍질 안에서 발달한다. 종자가 성숙함에 따라, 성장하는 배는 발생 후기 단계에서 종자 부피의 대부분을 차지한다.

많은 연구들이 DNA 메틸화 돌연변이 발생시 결함을 보고했고, 따라서 적절한 DNA 메틸화가 정상적인 배 발달에 중요하다는 것은 잘 알려져 있다. 그러나 배 발달 중 다른 애기장대 생태형 간 메틸화 차이의 인과관계는 완전히 이해가 되지는 않았다.

이 논문에서 종자가 성숙하고 발아하는 동안 두 생태형 사이의 메틸화 차이를 포괄적인 상호비교를 제시한다. Columbia-0(Col-0)와 Cape Verde Island(Cvi) 생태형을 사용하여 3 단계의 종자

발달 단계를 정했다; 1) 숙성 전(FH), 2) 숙성 후(AR), 3) 발아 자극처리(GS). 종자가 성숙하고 발아하는 동안 DNA 메틸화의 전반적인 변화를 발견했다. 3개의 사이토신 종류 중 가장 다르게 메틸화된 점은 유전자 영역과 전위인자에서 3단계 모두 Col-0가 항상 Cvi보다 높은 mCG 수준을 보인다는 것이다. 대조적으로, CG이외의 메틸화 수준은 두 생태형이 매우 유사했다. 흥미롭게도, 분화된 영역으로 간주되는 Col-0 특이적 DNA 서열(COR)이 두 생태형 모두에서 보존 서열(CR) 보다 더 메틸화가 되었다는 것을 발견됐다. COR은 전위인자를 다수 포함하며 삽입, 삭제를 포함한 구조변이(SV)가 많았다. COR의 전위인자가 CR의 전위인자보다 애기장대 유전체에 늦게 도입됐다는 점을 시사한다. 이전에 보고된 ago4/6/9 및 drm1/2 메틸롬 데이터를 사용하여 RdDM pathway가 COR의 높은 메틸화 수준에 기여한다는 것을 입증했다.

비록 DNA 메틸화의 대표적인 패턴, 특히 배 발달 동안 CHH 메틸화의 점진적인 증가가 다른 그룹을 통해 이미 보고 되었지만, 5 단계의 배 발달 과정단계의 배를 사용하여 메틸화 변화를 기준으로 9개의 서로 다른 클러스터 지역을 발견했다. 그 클러스터들은 유전체상 위치와 염색질의 특성이 다르게 나타났다.

요약하면, 본 연구는 RdDM pathway에 의해 새로 획득된 DNA 서열에 대한 조절의 가능성을 시사한다. 그리고 배 발달 중 CHH 메틸화 변화의 생물학적 의미를 밝히기 위한 새로운 분석을 제안

한다.

주요어: 구조 변이, RdDM pathway, 메틸롬, 애기장대, 배 발달

학번: 2016-20395