Similarity between soybean and *Arabidopsis* seed methylomes and loss of non-CG methylation does not affect seed development

Jer-Young Lin[a,1], Brandon H. Le[a,1,2], Min Chen[a,1], Kelli F. Henry[a], Jungim Hur[a], Tzung-Fu Hsieh[b,3], Pao-Yang Chen[a,4], Julie M. Pelletier[c], Matteo Pellegrini[a], Robert L. Fischer[b], John J. Harada[c], and Robert B. Goldberg[a,5]

[a]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA 90095; [b]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720; and [c]Section of Plant Biology, Division of Biological Sciences, University of California, Davis, CA 95616

Contributed by Robert B. Goldberg, October 3, 2017 (sent for review September 25, 2017; reviewed by Z. Jeffrey Chen, Brian A. Larkins, and Lila O. Vodkin)

We profiled soybean and *Arabidopsis* methylomes from the globular stage through dormancy and germination to understand the role of methylation in seed formation. CHH methylation increases significantly during development throughout the entire seed, targets primarily transposable elements (TEs), is maintained during endoreduplication, and drops precipitously within the germinating seedling. By contrast, no significant global changes in CG- and CHG-context methylation occur during the same developmental period. An *Arabidopsis ddcc* mutant lacking CHH and CHG methylation does not affect seed development, germination, or major patterns of gene expression, implying that CHH and CHG methylation does not play a significant role in seed development or in regulating seed gene activity. By contrast, over 100 TEs are transcriptionally de-repressed in *ddcc* seeds, suggesting that the increase in CHH-context methylation may be a failsafe mechanism to reinforce transposon silencing. Many genes encoding important classes of seed proteins, such as storage proteins, oil biosynthesis enzymes, and transcription factors, reside in genomic regions devoid of methylation at any stage of seed development. Many other genes in these classes have similar methylation patterns, whether the genes are active or repressed. Our results suggest that methylation does not play a significant role in regulating large numbers of genes important for programming seed development in both soybean and *Arabidopsis*. We conclude that understanding the mechanisms controlling seed development will require determining how *cis*-regulatory elements and their cognate transcription factors are organized in genetic regulatory networks.

seed development | DNA methylation | soybean | *Arabidopsis* | transposon

Seeds are derived from a double-fertilization process that leads to the differentiation of the seed coat (SC), endosperm, and embryo (EMB), the major regions of the seed that have distinct genetic origins and functions (1–3). The maternally derived SC differentiates from the ovule integuments that surround the embryo sac, transfers nutrients from the maternal plant to the developing EMB, and protects the seed during development and dormancy. The EMB and endosperm, by contrast, are descendants of the fertilized egg and central cell, respectively. The endosperm nourishes the EMB early in development, and in dicots, such as soybean and *Arabidopsis*, is absorbed and only present as a vestigial cell layer, the aleurone, in the mature seed. The EMB forms the two major embryonic organs: (*i*) an axis (AX), containing shoot and root meristems, which will give rise to the mature plant after seed germination; and (*ii*) the cotyledon (COTL), a terminally differentiated organ, which specializes in storage reserve production and senesces following germination (Fig. 1). Seeds shift into a maturation program following AX and COTL differentiation, that includes: (*i*) cessation of cell division, (*ii*) accumulation of storage reserves, and (*iii*) preparation for desiccation and dormancy (1, 2, 4) (Fig. 1). During this period, COTL cells enlarge and undergo a unique endoreduplication process that may facilitate the synthesis of highly prevalent seed-storage proteins that are utilized as an energy source during germination and early seedling (sdlg) growth (5, 6) (Fig. 1). At the end of maturation, programed water loss (i.e., desiccation) occurs, metabolic and developmental processes are suspended, a dormancy period begins that can last for several millennia (7), and the quiescent seed awaits an optimum environment for germination and sdlg growth (1) (Fig. 1).

DNA methylation plays a critical role in endosperm development (3, 8). *DEMETER* (*DME*) encodes a 5-methylcytosine glycosylase, and is expressed specifically in the central cell of the EMB sac (3, 9). DME removes methylated bases from maternal

Significance

We describe the spatial and temporal profiles of soybean and *Arabidopsis* seed methylomes during development. CHH methylation increases globally from fertilization through dormancy in all seed parts, decreases following germination, and targets primarily transposons. By contrast, CG- and CHG-context methylation remains constant throughout seed development. Mutant seeds lacking non-CG methylation develop normally, but have a set of up-regulated transposon RNAs suggesting that the CHH methylation increase may be a failsafe mechanism to reinforce transposon silencing. Major classes of seed genes have similar methylation profiles, whether they are active or not. Our results suggest that soybean and *Arabidopsis* seed methylomes are similar, and that DNA methylation does not play a significant role in regulating many genes important for seed development.

Author contributions: J.-Y.L., B.H.L., M.C., T.-F.H., P.-Y.C., M.P., R.L.F., J.J.H., and R.B.G. designed research; J.-Y.L., B.H.L., M.C., K.F.H., J.H., and J.M.P. performed research; J.-Y.L., B.H.L., M.C., and R.B.G. analyzed data; and J.-Y.L., B.H.L., M.C., and R.B.G. wrote the paper.

Reviewers: Z.J.C., University of Texas at Austin; B.A.L., University of Nebraska; and L.O.V., University of Illinois.

The authors declare no conflict of interest.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo [soybean methylome (accession nos. GSE34637, GSE37893, GSE37895, GSE41061, and GSE57762); *Arabidopsis* methylome (accession nos. GSE57755, GSE68131, and GSE68132); soybean transcriptome (accession nos. GSE29134, GSE29163, and GSE37895); and *Arabidopsis* transcriptome (accession no. GSE76447)].

[1]J.-Y.L., B.H.L., and M.C. contributed equally to this work.

[2]Present address: Department of Botany and Plant Sciences, University of California, Riverside, CA 92521.

[3]Present address: Department of Plant and Microbial Biology & Plants for Human Health Institute, North Carolina State University, Kannapolis, NC 28081.

[4]Present address: Institute of Plant and Microbial Biology, Academia Sinica, Taipei, 11529, Taiwan.

[5]To whom correspondence should be addressed. Email: bobg@ucla.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1716758114/-/DCSupplemental.

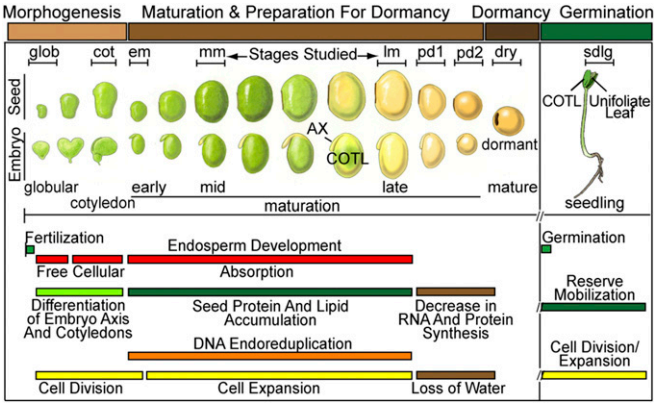

Fig. 1. Schematic representation of soybean seed stages and major developmental events. Adapted from refs. 38 and 48. Seed and EMB images are not drawn to scale. Brackets indicate stages investigated. Abbreviations are defined in Table 1.

genes within the central cell that are then expressed preferentially in the endosperm following fertilization, in comparison with their sperm-derived-methylated paternal counterparts. In addition, because *DME* is not expressed in the egg, the EMB is hypermethylated relative to the endosperm (10). Mutations in *DME* that disrupt these parent-of-origin (i.e., imprinting) methylation events result in abnormal endosperm development and seed abortion (3). By contrast, imprinting caused by differential methylation of paternal and maternal alleles does not appear to occur within the EMB, although there is conflicting evidence for the preferential activity of maternal and paternal genomes during early embryogenesis (3, 11–13). The extent to which DNA methylation events play a role in seed formation at all stages of development, and within different seed regions and tissue layers, remains largely unexplored.

We used soybean and *Arabidopsis* seeds to address the following questions: (*i*) Are there global DNA methylation changes during seed development from fertilization through dormancy and germination? (*ii*) Do seed regions and tissues have different methylation patterns? (*iii*) Is DNA methylation maintained during COTL cell endoreduplication? (*iv*) Are DNA methylation events in seed development conserved in different plants?

We applied whole-genome bisulfite (BS) sequencing (BS-Seq) and laser capture microdissection (LCM) to profile the DNA methylation landscape during seed development. We observed that DNA methylation profiles are similar in soybean and *Arabidopsis* seeds, which diverged ∼90 Mya (14). Global CHH methylation increases throughout the entire seed from differentiation to dormancy, targets all classes of transposable elements (TEs), and decreases in postgermination COTLs and sdlg. In addition, DNA methylation patterns in all sequence contexts are maintained during endoreduplication. Mutant *Arabidopsis* seeds lacking CHG and CHH methylation (15) develop and germinate normally, and have gene expression profiles that are mostly congruent with wild-type seeds. By contrast, 106 transposons are de-repressed transcriptionally in mutant seeds, suggesting that the increase in CHH methylation during seed development may be a failsafe mechanism to reinforce TE silencing. Finally, no significant DNA methylation changes occur around many genes known to be important for seed formation—including storage protein genes, fatty acid biosynthesis genes, and several major transcription factor (TF) genes—and many of these genes are in genomic regions devoid of DNA methylation at any stage of development. We conclude that the next major challenge to understanding seed development is to determine how *cis*-regulatory elements encoded in the genome and their cognate TFs that activate and repress gene activity are organized into regulatory networks that are required to "make a seed."

Results

Single-Base Resolution Soybean Seed Methylomes Throughout Development. We profiled the soybean DNA methylation landscape at single-base resolution from nine seed stages using BS-Seq to obtain a comprehensive methylation profile from postfertilization through dormancy and germination (Fig. 1 and Table 1). To compare the methylomes of different seed regions, subregions, and tissues at distinct developmental stages, we hand-dissected the AX COTL and SC from early-maturation (em) and midmaturation (mm) seeds, and used LCM to isolate: (*i*) AX, COTL, and SC from cotyledon (cot) seeds; (*ii*) SC tissue layers [parenchyma (PY) and palisade (PA)] from em seeds; and (*iii*) AX regions [plumule (PL), PY, vascular (VS), and root tip (RT)] from em seeds. Finally, we used LCM to obtain two em tissues showing differential endoreduplication (Fig. 1) within the same embryonic organ [COTL abaxial (ABPY) and adaxial (ADPY) PY tissues]. Collectively, our BS-Seq datasets provide a comprehensive spatial and temporal profile of the DNA methylation landscape across the entire soybean genome throughout all of seed development (Dataset S1).

In total, we generated ∼8 billion Illumina BS-Seq reads from all seed stages, regions, organs, and tissues, obtaining in each case 11–27× coverage of the ∼1-Gb soybean genome (Dataset S1). We assayed 273–287 million cytosines, representing 94–98% of all cytosines, at an average sequence depth of 5–13× per cytosine (*SI Materials and Methods* and Dataset S2). We checked the conversion efficiency of the BS treatment by examining the conversion of C-to-T in both the unmethylated chloroplast genome and a λ genome that was added to our samples as an internal control

Table 1. Development stage, region, and tissue abbreviations

Abbreviation	Description
Soybean seed stage and postgermination	
glob	Globular
cot	Cotyledon
em	Early maturation
mm	Midmaturation
lm	Late maturation
pd1	Early predormancy
pd2	Late predormancy
dry	Dry seed
sdlg	Whole seedling
Arabidopsis seed stage	
glob	Globular
lcot	Linear cotyledon
mg	Mature green
pmg	Postmature green
dry	Dry seed
Seed regions, subregions, and tissues	
ABPY	Abaxial parenchyma
ADPY	Adaxial parenchyma
AL	Aleurone
AX	Axis
COTL	Cotyledon
EMB	Embryo
HG	Hourglass
PA	Palisade
PL	Plumule
PY	Parenchyma
RT	Root tip
SC	Seed coat
sdlg-COTL	Seedling cotyledon
VS	Vascular

(*SI Materials and Methods*). We observed an average BS conversion efficiency of unmethylated C-to-T greater than 99.5% for both the chloroplast and λ genomes, indicating high conversion efficiency for our BS treatment (Dataset S1). The BS-Seq data from biological replicates of whole seeds and seed parts were in excellent agreement with each other (correlation coefficients > 0.99) (Fig. S1*A*). Additionally, we observed that 9%, 12%, and 79% of the seed methylomes were present in CG, CHG, and CHH contexts (where H = A, C, T), which was similar to the proportion of CG, CHG, and CHH sites in the soybean genome (Fig. S1*B*). We calculated the bulk methylation levels to determine the extent to which the soybean seed genome was methylated (10) (*SI Materials and Methods*), and observed: (*i*) an average methylation level of 12% for all detected cytosines, and (*ii*) average methylation levels of 57%, 36%, and 2% in the CG, CHG, and CHH contexts across all samples, respectively (Dataset S2), values similar to those obtained in other soybean methylome studies (16–18). Taken together, these results indicate that our datasets represent unbiased, deep representation, and highly reproducible profiles of soybean seed methylomes.

CHH Methylation Levels Increase During Soybean Seed Development. To determine whether global DNA methylation changes occurred during soybean seed development, we calculated the bulk methylation levels (*SI Materials and Methods*) for CG, CHG, and CHH contexts in 500-kb windows across the genome from whole seeds at the globular (glob), cot, em, mm, late maturation (lm), early predormancy (pd1), late predormancy (pd2), and dry stages, representing the differentiation, maturation, and dormancy phases (Fig. 1). The box plots show that there were no significant global DNA methylation changes in either CG or CHG contexts between: (*i*) adjacent seed stages or (*ii*) postfertilization glob and dry seeds (*t* test, $P < 0.001$ and >1.5-fold increase), suggesting that CG and CHG sites were either methylated at fertilization when seed development begins (i.e., before the glob stage) or before (Fig. 2*A* and Dataset S3). By contrast, global CHH methylation levels increased more than threefold from postfertilization (glob, cot) through maturation (em, mm, lm), and then plateaued from lm through desiccation (pd1 and pd2) and dormancy (dry) (Fig. 2*A* and Dataset S3). Most of the increase occurred after the cessation of cell division between the em and lm stages when the seed had ~3×10^6 cells and was enlarging due to cell expansion (19). This is consistent with the mechanism of CHH methylation via the RNA-directed DNA methylation (RdDM) pathway that can occur "de novo" and independently of DNA replication (20).

It was possible that the elevation in CHH bulk methylation level during seed development was due to: (*i*) accumulated methylation of the same cytosine sites in different cells, (*ii*) methylation of new cytosine sites, or (*iii*) both. To distinguish between these possibilities, we determined the absolute number of cytosine sites in the CHH context that were methylated at each developmental stage (*SI Materials and Methods*), and compared these values with the bulk CHH methylation levels (Fig. S2 *A* and *B*). The number of methylated CHH sites increased significantly (Fisher's exact test, $P < 0.001$ and >1.5-fold increase) from the cot to em stage, and then leveled off in subsequent stages (Fig. S2 *A* and *B*). By comparison, the bulk CHH level increased during the same developmental period, but continued to increase through the lm stage (Fig. 2 and Fig. S2*B*). These results indicate that the increase in CHH methylation during seed development is due to the addition of new methylated CHH sites across the genome, and the accumulated methylation of the same CHH sites in different seed cells as shown by the heuristic model (Fig. S2*B*).

CHH Methylation Levels Drop During Soybean Seed Germination. We determined whether the seed CHH methylation level was maintained after germination, by profiling sdlg and sdlg-COTL methylomes, representing the (*i*) transitional state from a dormant seed

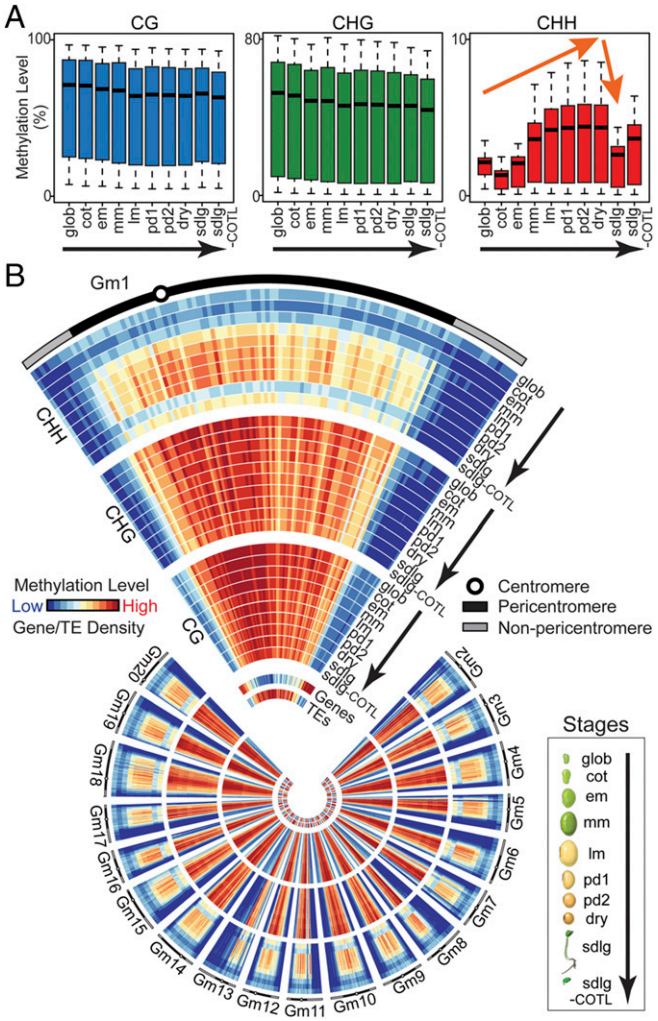

Fig. 2. Genome-wide methylation changes during soybean seed development and germination. DNA methylation levels in 500-kb windows across the genome are represented as box plots (*A*) and chromosome heat maps (*B*). The highest methylation levels (%) for heat maps tracks are 96, 80, and 9 for CG, CHG, and CHH contexts, respectively. Gene and TE tracks represent gene and TE densities in 500-kb windows across the genome. Gm, *Glycine max*. See Table 1 for abbreviations.

to a rapidly growing sdlg and (*ii*) COTL in different functional states (i.e., dormant seeds and metabolically active seed leaves), respectively (Fig. 1). The sdlg we assayed contained the developing root, elongating hypocotyl, postgermination COTL, and emerging leaves (Fig. 1 and *SI Materials and Methods*). In comparison with the dry seed, both the bulk CHH methylation levels and the number of methylated CHH sites dropped significantly in the sdlg during germination (*t* test and Fisher's exact test, $P < 0.001$ and >1.5-fold change) and in germinating sdlg-COTL, although to a lesser extent in the latter (*t* test and Fisher's exact test, $P < 0.001$ and <1.5-fold change) (Fig. 2*A* and Dataset S3). By contrast, CG and CHG methylation levels in the postgermination sdlg and sdlg-COTL were maintained and did not change significantly (Fig. 2*A* and Dataset S3). These data suggest that CG and CHG sites are methylated in seed AX tissues that give rise to the sdlg (e.g., PL, PY, RT) and are maintained following germination. By contrast, the hypomethylation of CHH sites in the sdlg compared with the dormant seed might indicate that either: (*i*) methylated CHH sites within the AX become diluted as they divide, increase in number, and differentiate within the germinating sdlg; (*ii*) CHH sites within

the AX were hypomethylated within the dry seed before germination (i.e., not methylated during seed development); or (*iii*) both. The different results obtained between CG- and CHG-context methylation, and CHH-context methylation in the sdlg following seed germination might also reflect the distinct mechanisms by which CG, CHG, and CHH sites are methylated: the former (CG and CHG) utilizing hemimethylated cytosines or replication-dependent histone variants as guides during replication, whereas the latter (CHH) occur de novo without the use of a methylated cytosine on one DNA strand (21, 22).

CHH Methylation Levels Change Within All Soybean TE Classes During Seed Development and Germination. The developmental changes in CHH-context methylation levels occurred predominantly in the pericentromeric regions of each soybean chromosome, where the majority of TEs were located (Fig. 2*B* and Fig. S3*A*). We divided the TEs by chromosome location, class, and length and determined that the genome-wide changes in seed and sdlg CHH-context methylation occurred within TEs, irrespective of location, size, or type in the genome (Fig. S3 *B–D*). For example, the temporal increases and decreases in CHH-context methylation occurred in DNA transposons (e.g., Mutator), retrotransposons (e.g., Gypsy and Copia), and all transposon size classes (Fig. S3 *B–D*). In addition, these changes also occurred in TEs that were located in the TE-rich pericentromeric region and the gene-rich chromosomal arms (Fig. S3 *B* and *C*). These results suggest that the mechanisms responsible for the developmental changes in CHH methylation levels are coordinated within TEs across the genome, and occur in parallel with developmental events that take place during seed development and germination (Figs. 1 and 2).

CHH Methylation Changes Occur Throughout the Entire Soybean Seed. We isolated seed AX, COTL, and SC regions at different developmental stages using both LCM (cot stage) and manual dissection (em and mm stages) to determine whether CHH-context methylation changes occurred throughout the entire seed (Fig. 3 *A* and *B*). It was necessary to use LCM for cot-stage seeds because the AX, COTL, and SC regions were too small to be dissected by hand (Fig. 3*A*).

The CG- and CHG-context bulk methylation levels were similar in all seed parts and did not change during seed development (Fig. 3*C*), analogous to the results obtained with whole seeds (Fig. 2). By contrast, the CHH-context methylation levels within the AX, COTL, and SC increased significantly (*t* test, $P < 0.001$ and >1.5 fold change) (Dataset S3) from the cot to mm stages (Fig. 3*C*), paralleling changes that were observed with the seed as a whole (Fig. 2*A*). In addition, the CHH methylation levels differed significantly (*t* test, $P < 0.001$, >1.5-fold change) between the AX, COTL, and SC regions at all developmental stages (Fig. 3*C* and Dataset S3). In general, the SC had the lowest level of CHH methylation, whereas the AX had the highest (Fig. 3*C*). The temporal and spatial differences in AX, COTL, and SC CHH-context methylation levels were also reflected within major TE classes scattered across the genome (Fig. S3*E*). Together, these results indicate that the biological events responsible for the temporal increase in CHH-context methylation during seed development are coordinated spatially within all major seed regions, and that the maternally derived SC layer is hypomethylated in comparison with the embryonic AX and COTL regions.

Individual Soybean SC Tissue Layers Are Hypomethylated. The SC consists of different tissue layers—hourglass, PA, and PY—of which the PY is the most prominent and constitutes most of the SC (Fig. 3 *D* and *E*). We used LCM to capture em-stage SC-PA and SC-PY tissue layers to determine whether the CHH-context hypomethylation occurred throughout the entire SC or was unique to a given layer (e.g., major PY layer). As a control, we used LCM to capture the entire SC and COTL from em-stage seeds (Fig. 3*D*). No bulk

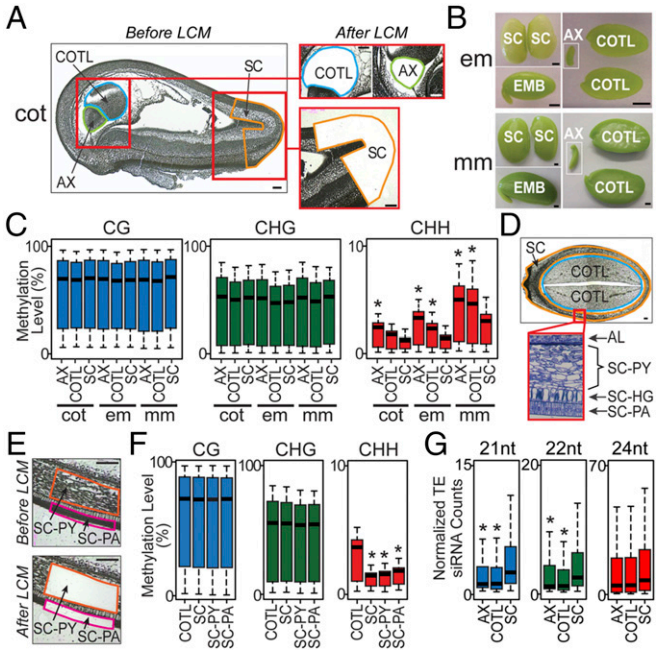

Fig. 3. Comparison of methylomes between soybean seed parts and SC layers. (*A*) Paraffin sections of cot-stage SC, embryonic AX, and embryonic COTL before and after capture by LCM. (*B*) Whole-mount pictures of SC, AX, and COTL from em- and mm-stage EMB and seeds. (*C*) Box plots of DNA methylation levels in 500-kb windows across the genome in different seed parts. Asterisks indicate significant comparisons between SC and other seed parts at the same stage (*t* test, $P < 0.001$ and fold change > 1.5). (*D*) Paraffin cross-section of an em-stage seed (*Upper*), and a plastic section of SC layers (*Lower*), which is the red boxed region shown in the whole-seed section. (*E*) Paraffin cross-sections of em-stage SC-PA and SC-PY layers before and after capture by LCM. (*F*) Box plots of LCM-captured em-stage COTL, SC, SC-PA, and SC-PY DNA methylation levels in 500-kb windows across the genome. Asterisks indicate significant comparisons between SC, SC layers, and the COTL (*t* test, $P < 0.001$ and fold change > 1.5). (*G*) Box plots of TE siRNA levels (reads per million mapped reads) in em-stage SC, AX, and COTL. Asterisks indicate statistically significant comparisons between SC and other seed parts (*t* test, $P < 0.001$ and fold change > 1.5). [Scale bars, 100 μm (*A*, *D*, and *E*) and 1 mm (*B*).] See Table 1 for abbreviations.

methylation differences were observed between the SC-PA and SC-PY layers in any cytosine context (CG, CHG, and CHH) (Fig. 3*F*). By contrast, the SC-PA and SC-PY CHH-context methylation levels were both significantly lower relative to the COTL (*t* test, $P < 0.001$ and >1.5-fold change) (Dataset S3), and similar to results obtained with the SC as a whole (Fig. 3 *C* and *F*). Taken together, these results indicate that the methylation levels of the entire SC reflect those within individual tissue layers, including CHH-context hypomethylation relative to the AX and COTL regions of the seed.

The Soybean SC Layer Contains Elevated Levels of siRNAs. We isolated small RNAs from em-stage AX, COTL, and SC, and used RNA-Seq to determine whether the CHH-context hypomethylation of the SC relative to other regions of the seed resulted in an elevated level of siRNAs derived from TEs (*SI Materials and Methods*). Twenty-four–nucleotide siRNA levels were similar in all seed regions (Fig. 3*G*). By contrast, both 21-nt and 22-nt siRNAs were elevated significantly (*t* test, $P < 0.001$ and >1.5-fold change) in the SC compared with the AX and COTL regions. These results suggest that: (*i*) CHH-context hypomethylation of TEs within the SC (Fig. S3*E*) might result in reduced TE silencing and TE movement, which would have little effect on subsequent development as the SC does not contribute to the postgerminating sdlg; (*ii*) elevated 21-nt and 22-nt siRNAs could mitigate this possibility by posttranscriptional silencing of SC TEs (23); and (*iii*) 21-nt and

22-nt SC siRNAs might move to other parts of the seed, such as the AX and COTL, which have higher CHH-context methylation levels (Fig. 3 *C* and *F*), and reinforce TE silencing (10) analogous to what occurs between vegetative and sperm cells within the pollen grain (23, 24).

CHH-Methylation Levels Within Soybean AX Subregions and Tissues Differ. We used LCM to capture tissues of the em AX that give rise to sdlg following germination. These included: the (*i*) PL, (*ii*) PY and VS, and (*iii*) RT that participate in forming the seed leaves, hypocotyl, and root of the germinating sdlg, respectively (Fig. S4*A*). The CG- and CHG-context bulk methylation levels for the AX-PL, AX-PY, AX-RT, and AX-VS tissues were similar and not significantly different from those observed within the germinating sdlg (Fig. 2*A* and Fig. S4*B*). By contrast, the bulk CHH-context methylation level of the RT was significantly higher than that of the AX-PL, AX-PY, or AX-VS tissues, indicating that the AX as a whole represents the average of different CHH-context methylation levels in specific AX tissues. This situation differs from that observed with the SC (Fig. 3*F*), and suggests that the decrease in CHH-context methylation observed in the postgermination sdlg might be a consequence of (*i*) hypomethylated CHH sites within seed AX tissues remaining unmethylated following germination (e.g., AX-PL, AX-PA, and AX-VS), and (*ii*) preexisting hypermethylated CHH sites being diluted as sdlg cells divide (e.g., AX-RT), both of which were predicted by our seed development and germination results (Fig. 2*A*). By contrast, the similar levels of CG- and CHG-context methylation in the sdlg compared with the developing seed probably result from preexisting methylated sites in seed AX tissue layers remaining methylated in the postgermination sdlg.

DNA Methylation and Endoreduplication Are Coupled During Soybean Seed Development. Because the increase in CHH-context DNA methylation coincided with the onset of COTL endopolyploidization (Figs. 1 and 2), we asked whether the DNA methylation landscape was maintained following endoreduplication in soybean seeds. We used LCM to capture em-stage COTL ABPY and ADPY tissues (Fig. 4*A*), taking advantage of the observation that COTL ABPY and ADPY tissues differ in endoreduplication timing: ABPY undergoes endoreduplication at the em stage, while ADPY does not (25). No significant differences were observed in the bulk methylation levels between endoreduplicating ABPY and nonendoreduplicating ADPY tissues in all three cytosine contexts (Fig. 4*B* and Dataset S3). In addition, >96% of cytosine sites across the genome retained their methylation status in endoreduplicating ABPY and nonendoreduplicating ADPY tissues. Confirming these observations, both the (*i*) bulk CG-, CHG-, and CHH-context methylation levels and (*ii*) cytosine site methylation status did not differ between mm-stage endoreduplicated COTL and nonendoreduplicated AX regions (5) (Fig. 3*C* and Dataset S3). Taken together, these data indicate that the methylation landscape is maintained following endoreduplication in COTL cells.

We compared the DNA sequence coverage along the entire soybean genome for: (*i*) em-stage COTL ABPY and ADPY tissues, (*ii*) em-stage COTL and AX regions, and (*iii*) mm-stage COTL and AX regions, and did not observe any major differences in genome coverage indicating that there was uniform DNA replication along the genome in endoreduplicating cells (Fig. 4*C*). That is, all DNA sequences in the genome were replicated to the same extent with no selective amplification. These results indicate that DNA methylation is maintained in all sequence contexts during endoreduplication and is highly coordinated with DNA synthesis in the absence of cell division.

Many Genes Important for Soybean Seed Development Are Present in Genomic Regions That Are Not Methylated. We scanned the seed methylomes in 5-kb sliding windows from postfertilization through

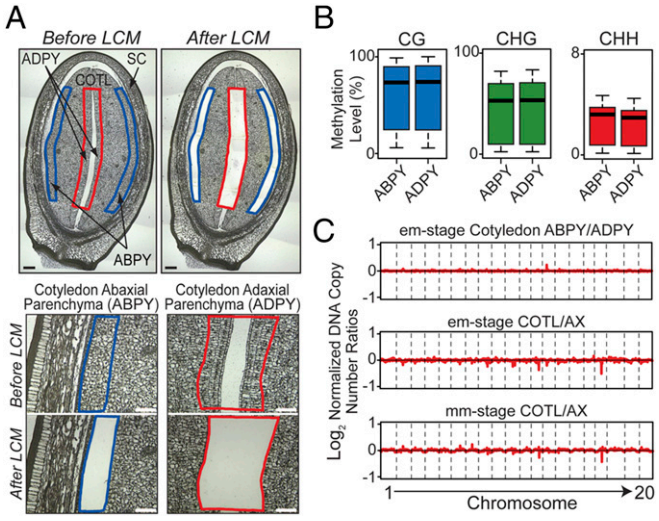

Fig. 4. DNA methylation levels in endoreduplicated and nonendoreduplicated soybean COTL regions. (*A*) Paraffin cross-sections of em-stage COTL endoreduplicated ABPY and nonendoreduplicated ADPY before and after capture by LCM. (Scale bars, 100 μm.) (*B*) Box plots of DNA methylation levels in 500-kb windows across the genome in em-stage COTL ABPY and ADPY regions. (*C*) Log_2 ratios of normalized DNA reads in 500-kb windows across all 20 chromosomes from: (*i*) em-stage ABPY and ADPY COTL regions, and (*ii*) em and mm seed parts (COTL and AX). See Table 1 for abbreviations.

dormancy and germination to characterize the methylation landscape surrounding ∼75 genes known to be important for seed development, and determined whether the regulation of these genes was associated with localized methylation changes in any cytosine context (Dataset S4). These included genes encoding: (*i*) storage proteins (e.g., glycinin and β-conglycinin), (*ii*) oil biosynthesis proteins (e.g., stearoyl-acyl–carrier-protein desaturase), (*iii*) transcription factors [e.g., LEAFY COTYLEDON1 (LEC1) and FUSCA3 (FUS3)], and (*iv*) germination-enhanced proteins (e.g., chlorophyll A/B binding protein and isocitrate lyase). Many laboratories, including our own, demonstrated that these genes are highly regulated and under transcriptional control (1, 26, 27), as suggested by the RNA-Seq data presented here (Fig. 5*A*, Fig. S5 *A*, *C*, and *D*, and Dataset S4).

Surprisingly, almost half of the seed and germination genes we investigated were localized within genomic regions designated as demethylated valleys (DMVs) (28), that averaged <5% bulk methylation level in any cytosine sequence context across all stages (Fig. 5, Fig. S5, and Dataset S4). Some of these regions extended for >50 kb but, on average, were 16 kb (Dataset S4). The methylation status of these regions did not change from fertilization through dormancy and germination, whereas the seed and germination genes in these regions were highly regulated (Fig. 5, Fig. S5, and Dataset S4). The remainder of genes we investigated fell into three categories: (*i*) genes with methylated TEs in their 5′ and 3′ flanking regions, (*ii*) genes with gene body methylation, and (*iii*) genes with gene body methylation and TEs in their 5′ and 3′ flanking regions (Fig. 5, Fig. S5, Dataset S4). In each case, however, no significant CG-, CHG-, or CHH-context methylation changes were observed within or flanking the seed and germination genes, although these genes were highly regulated during development (Fig. 5 and Fig. S5). In addition, previous *cis*-element analysis of the *GmLe1* and *GmKTi1* genes that have methylated flanking TEs (Fig. 5*B* and Fig. S5*E*) showed that these TEs were not required for regulation during seed development (29–31). Taken together, these data suggest that regulation of many important seed and germination genes is primarily due to transcriptional events that are independent of DNA methylation changes, in

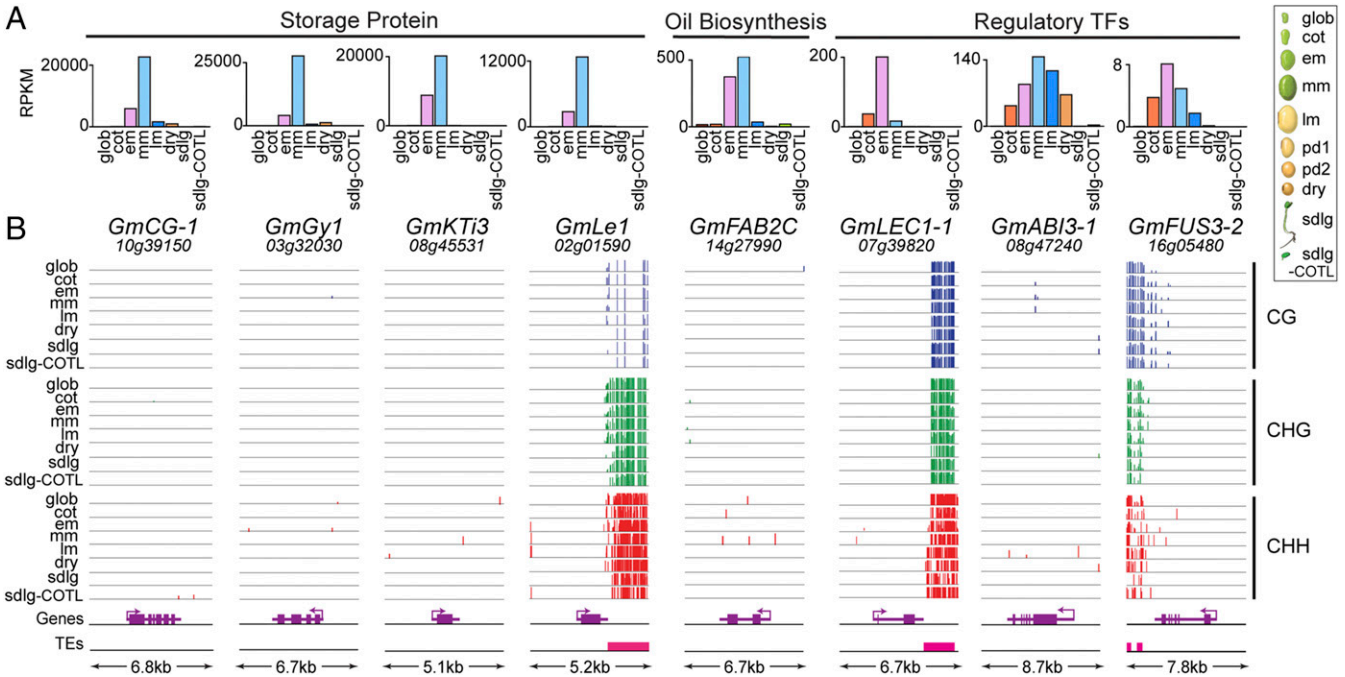

Fig. 5. Methylation levels and mRNA accumulation patterns of major soybean seed-specific gene classes during seed development and germination. (*A*) mRNA accumulation patterns. RPKM represents reads per kilobase per million sequences, and were taken from the Goldberg-Harada soybean (*i*) whole seed RNA-Seq dataset, GEO accession no. GSE29163 (37), and (*ii*) COTL-specific RNA-Seq dataset GSE29134 (sdlg-COTL). (*B*) Methylation levels of CG-, CHG-, and CHH-context sites are shown in genome browser view (vertical lines). Gene structures, transcription directions (arrows), and TEs are shown below each genome browser view. Adjacent genes are not shown. The size of each genomic region, including 2 kb of 5′ and 3′ flanking regions, is shown at the bottom. *GmABI3-1*, abscisic acid insensitive3-1; *GmCG-1*,β-conglycinin-1; *GmFAB2C*, stearoyl-ACP desaturase 2C; *GmFUS3-2*, FUSCA 3-2; *GmGy1*, glycinin 1; *GmKTi3*, Kunitz trypsin inhibitor 3; *GmLEC1-1*, Leafy Cotyledon 1-1; *GmLe1*, lectin 1. See Table 1 for developmental stage abbreviations. Gm, *Glycine max*.

agreement with our observations three decades ago using more primitive technology (26).

The Methylation Landscape of Soybean Seeds Is Conserved in *Arabidopsis*. We carried out a series of BS-Seq experiments with *Arabidopsis* seeds at stages comparable with those studied in soybean to determine whether our methylation observations were specific for soybean seeds or found generally in higher plants (Datasets S1–S3). Stages investigated included those undergoing: (*i*) morphogenesis and differentiation [glob and linear cot (lcot) stages], (*ii*) maturation [mature green (mg) and postmature green (pmg) stages], (*iii*) dormancy (dry seeds), and (*iv*) postgermination (leaves from 3-wk-old plants) (Fig. 6*A*). We obtained 45–100× coverage of the 120-Mb *Arabidopsis* genome for each seed stage and part investigated (Dataset S1), and observed that: (*i*) the base composition of our reads reflected that of the *Arabidopsis* genome (Fig. S1*C*), and (*ii*) on average, the bulk methylation levels of cytosines in the CG, CHG, and CHH contexts were 24.6%, 7.7%, and 1.6%, respectively, (Dataset S2),values similar to those obtained in other *Arabidopsis* methylome studies (32).

Surprisingly, the methylation events observed during *Arabidopsis* seed development were indistinguishable from those observed in soybean seeds. These included: (*i*) no significant changes in CG- and CHG-context bulk methylation levels during seed development and germination (Fig. 6 *B* and *C* and Dataset S3); (*ii*) a significant increase and decrease in both CHH-context bulk methylation levels and methylated sites (*t* test and Fisher's exact test, P value < 0.001 and >1.5-fold change) within pericentromeric region TEs during maturation and germination, respectively (Fig. 6 *B* and *C*, Fig. S2 *C* and *D*, and Dataset S3); and (*iii*) CHH-context hypomethylation of the mg-stage SC relative to the EMB (*t* test, P < 0.001 and >1.5-fold change) (Fig. 6*D* and Dataset S3). Finally, genes encoding major classes of proteins important for seed development (e.g., storage proteins, oil biosynthesis, TFs) were: (*i*) localized within DMV regions with <5% bulk methylation levels across all of seed development (e.g., *AtCRA1*), (*ii*) methylated within their gene bodies (e.g., *AtFAB2*) or flanking regions (*AtLEC1*, *AtABI3*, *AtFUS3*), and (*iii*) either regulated in the absence of any detectable methylation events or changes that were not correlated with their expression programs (Fig. 6 *E* and *F*), similar to what was observed in soybean (Fig. 5 and Fig. S5). Together, these data suggest that the methylation landscape of soybean and *Arabidopsis* seeds is highly conserved despite a divergence of ∼90 My (14), and that the programmed changes in CHH-context methylation during development and between different seed regions may be a common feature of dicot seeds.

Seed Development Occurs Normally in a Mutant *Arabidopsis* Line Lacking CHG and CHH Methylation. To explore the possible role, if any, that the increase in CHH-context methylation plays in seed development, we took advantage of the *Arabidopsis ddcc* mutant (*drm1drm2cmt2cmt3*) (33) that is deficient in all methyltransferases [DOMAINS REARRANGED METHYLTRANSFERASE1 (DRM1), DOMAINS REARRANGED METHYLTRANSFERASE2 (DRM2), CHROMOMETHYLASE2 (CMT2), and CHROMOMETHYLASE3 (CMT3)] required for non-CG–context methylation (15). We reasoned that the *ddcc* mutant would provide an excellent test of the functional relevance of seed CHH-context methylation because CHG-context methylation levels do not change in soybean and *Arabidopsis* seeds from fertilization through dormancy and germination.

We did not detect any CHG- or CHH-context methylation in *ddcc* pmg seeds, dry seeds, or postgermination leaves in comparison with wild-type (Fig. 7*A*), as expected from knocking out genes required for non-CG–context methylation (15). *ddcc* seeds developed normally with no detectable morphological differences

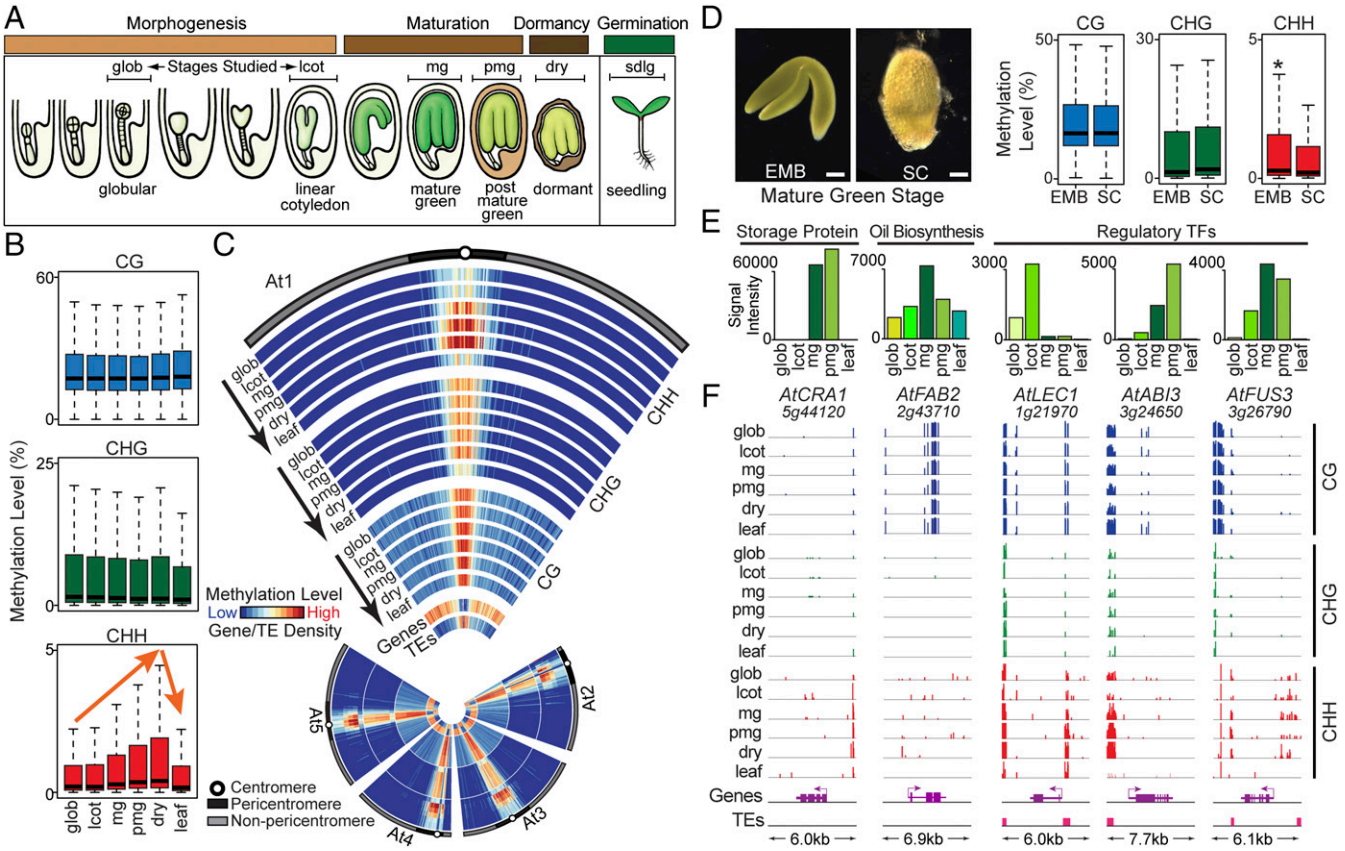

Fig. 6. Genome-wide methylation changes during *Arabidopsis* seed development. (*A*) *Arabidopsis* seed stages and major developmental events. Adapted from ref. 38). Seed and EMB images are not drawn to scale. Brackets indicate stages investigated. Box plots (*B*) and chromosome heat maps (*C*) of DNA methylation levels in 100-kb windows across the genome. The highest methylation levels were 96%, 80%, and 10% for CG, CHG, and CHH contexts, respectively. Gene and TE tracks represent densities of genes and TEs in 100-kb windows along the genome. (*D*) Box plots of DNA methylation levels in 100-kb windows across the genome in mg-stage EMB and SC. (Scale bars, 0.1 mm.) Asterisk indicates a significant comparison between EMB and SC (*t* test, $P < 0.001$ and fold change > 1.5). mRNA accumulation patterns (*E*) and genome browser views of DNA methylation levels (*F*) for major seed-specific gene classes. Transcript signal intensities were obtained from microarray analysis (38). Methylation levels of CG-, CHG-, and CHH-context sites are shown in genome browser view (vertical lines). Gene structures, transcription directions (arrows) and TEs are shown below each genome browser view. Adjacent genes are not shown. The size of each genomic region, including 2 kb of 5′ and 3′ flanking regions, is shown at the bottom. *AtCRA1*, Cruciferin 1. Names of other genes are defined in the legend to Fig. 5. At, *Arabidopsis thaliana*. See Table 1 for abbreviations of seed stages.

from wild-type at our level of resolution (Fig. 7*B*). In addition, *ddcc* COTL nuclei underwent normal shrinkage in pmg seeds, which is a marker for the desiccation events that occur at the end of seed development (Fig. 7 *C* and *D*), and then regained their size following germination (34). Finally, *ddcc* and wild-type seeds had the same levels of germination (Fig. 7 *E* and *F*). Taken together, these data suggest that non-CG methylation does not play a significant role in *Arabidopsis* seed development, and that the absence of programmed CHH-methylation changes does not affect seed morphogenesis, maturation, dormancy, or germination.

Gene Expression Is Not Affected Significantly in *Arabidopsis ddcc* Seeds. We compared the transcriptomes of *Arabidopsis ddcc* and wild-type pmg seeds to determine whether the loss of non-CG methylation, and CHH-context changes in particular, affected seed gene expression. Quantitative and qualitative mRNA levels in *ddcc* and wild-type seeds were not significantly different from each other (Fig. 8 *A* and *B*), including mRNAs encoding storage proteins (e.g., *AtCRA1*), fatty acids (e.g., *AtFAB2*), and major regulators of seed development (e.g., *AtABI3*, *AtFUS3*) (Fig. 8*C*). We obtained a 0.97 correlation coefficient between pmg *ddcc* and wild-type seed transcriptomes, which was the same as that obtained between either *ddcc* or wild-type biological replicates (Fig. S6*A*). Although the vast majority of the 19,638 mRNAs detected in *ddcc* pmg seeds were present in wild-type seeds at the same levels, we did detect a small number of *ddcc* mRNAs that were either up-regulated (90 genes) or down-regulated (17) [false-discovery rate (FDR) < 0.001 and >fivefold change] (Fig. 8*B* and Dataset S5). Gene ontology analysis showed that down-regulated genes were enriched for response to stress, while the up-regulated genes were associated with abiotic stress response and cellular component organization/anatomical structure formation (Dataset S6). We examined the 5′ flanking regions of the 107 differentially expressed *ddcc* genes, and found that ∼70% had no CHG- or CHH-context methylated sites in corresponding wild-type genes (Fig. 8*D* and Dataset S5). This suggests that mostly indirect effects were responsible for change in gene activity observed in *ddcc* pmg seeds. Together, these results suggest that non-CG methylation, including elevated CHH-context methylation levels, does not play a major role in regulating seed gene activity.

Many TE mRNAs Are Up-Regulated in *Arabidopsis ddcc* Seeds. We investigated whether any *Arabidopsis* TEs were de-repressed at the RNA level in pmg *ddcc* seeds as a consequence of non-CG methylation loss (Fig. 7*A*). Qualitative and quantitative differences were observed between pmg *ddcc* and wild-type TE RNAs (0.48 correlation coefficient) (Fig. 9*A*), by contrast with the biological replicate controls (Fig. S6*B*). We found that 96 TEs were de-repressed

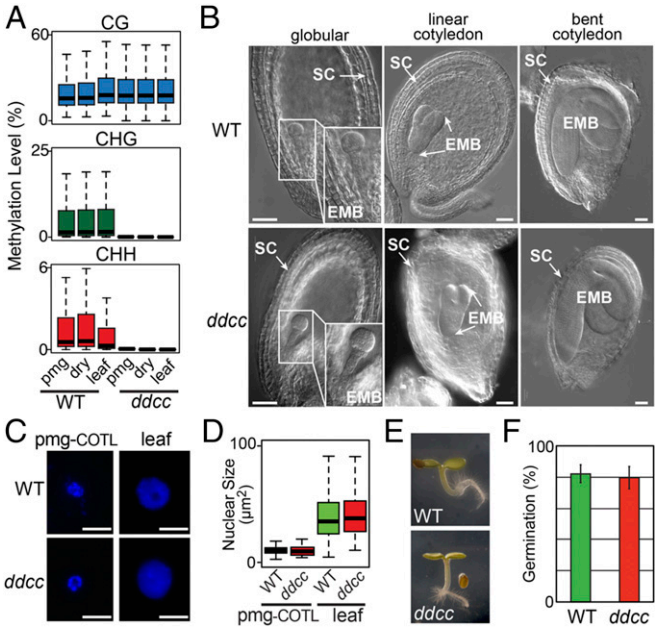

Fig. 7. Comparison between *Arabidopsis* wild-type and *ddcc* seed development and germination. (*A*) Box plots of methylation levels in 100-kb windows across the wild-type and *ddcc* genomes. (*B*) Nomarski photographs of wild-type and *ddcc* seeds at different developmental stages. (Scale bars, 50 μm.) (*C*) DAPI-stained nuclei of pmg-COTL and leaves. (Scale bars, 5 μm.) Comparison of nuclear sizes (110 nuclei) (*D*), 4-d sdlg morphologies (*E*), and germination percentages (*F*) for wild-type and *ddcc* seeds. Five replicates with 50 seeds each were used in the germination assays.

transcriptionally from a silenced state, while 10 TEs were up-regulated >sixfold (FDR < 0.01) (Fig. 9 *A*–*C* and Dataset S7). TE RNAs were transcribed from different retrotransposon and DNA TE families, although TE transcripts were most numerous from retrotransposon classes (Fig. S7 *A* and *B*). TE RNAs encoded proteins responsible for copy number increase and transposition, including transposase, integrase, and reverse transcriptase (Fig. 9*D*). We compared the copy number of de-repressed and up-regulated TEs in *ddcc* and wild-type genomes (35), and did not detect any significant differences, implying that these TEs did not undergo transposition events within the seed generation that we investigated (Fig. 9*E*).

We investigated the methylation landscape of the 106 *ddcc* de-repressed and up-regulated TEs in the wild-type pmg seed genome to determine how loss of non-CG methylation resulted in their transcriptional activation in *ddcc* seeds. We randomly selected 106 silenced TEs (i.e., no detectable RNA-Seq reads) with a similar distribution of TE classes and lengths as controls. The up-regulated and de-repressed TEs had significantly lower CG densities (*t* test, $P < 0.001$) compared with control TEs, by contrast with CHG and CHH densities, which did not differ from the control TE set (Fig. S7*C* and Dataset S8). CG-, CHG-, and CHH-context bulk methylation levels were significantly higher in the de-repressed and up-regulated TEs compared with the repressed controls, suggesting that there were insufficient methylated CG sites to prevent TE transcription in *ddcc* seeds (Fig. S7*D* and Dataset S8). Examination of the CG- and CHG-context methylation levels across the de-repressed and up-regulated TE bodies showed that they were similar to the TE controls (Fig. 9*F*). By contrast, the distribution of CHH-context methylation across the 106 de-repressed and up-regulated TEs differed significantly from the control set, and showed that there was a prominent increase in CHH methylation levels at the TE ends where the promoter sequences reside (Fig. 9*F*). The CHH-context methylation levels at the promoter sites and along the entire TE bodies increased significantly from fertilization through dormancy during seed development, by contrast to CG- and CHG-context methylation, which did not change (Fig. 9*G*). Together, these data imply that de-repression of TEs in *ddcc* pmg seeds might be caused primarily by the loss of highly methylated CHH-context sites within TE promoter regions, and suggest that the programmed increase in CHH-context methylation during soybean and *Arabidopsis* seed development might be a failsafe mechanism to ensure TE silencing.

Discussion

We determined that there is a programmed increase in CHH-context methylation during seed development from the glob stage through dormancy in both soybean and *Arabidopsis* seeds. The majority of this upswing occurs during the maturation phase in all major seed parts—SC, COTL, and AX—and is maintained during endoreduplication of the COTL genome. Although the SC layer is hypomethylated in CHH sites relative to other seed parts, the increase in CHH-context methylation occurs simultaneously throughout the seed. By contrast, CG- and CHG-context methylation does not change significantly during the same period of seed development, or in any specific seed part. Following germination, CHH-context methylation drops precipitously in germinating COTL and the emerging sdlg. Thus, the change that occurs in seed CHH-context methylation appears to be conserved in dicots, and is regulated with respect to space and time during the development of the seed. Very recently, an independent analysis of our *Arabidopsis* whole-seed data results in a similar conclusion (36).

The increase in seed CHH-context methylation occurs across the entire soybean and *Arabidopsis* genomes, targets primarily TEs, and is neutral with respect to TE class. Both DNA transposons and retrotransposons are targeted, as well as TEs that are either clustered in heterochromatic pericentromere regions or dispersed among genes in euchromatic chromosome arms. This implies that there is a coordinated targeting of CHH-context sites in TEs during seed development, most likely being directed by RdDM and non-RdDM pathway methylases DRM1, DRM2, CMT2, and CMT3, respectively (15, 20). Inspection of the soybean (GSE29163) (37) and *Arabidopsis* (GSE680) (38) seed transcriptome databases indicates that mRNAs encoding these methylases are present when the CHH-context methylation events occur both temporally and spatially during seed development, supporting this premise (Dataset S9). Following germination the decrease in CHH-context methylation is most likely caused by the methylation status of different dry seed AX regions that give rise to the sdlg. For example, methylated CHH sites within the AX-RT become diluted as the root cells divide and increase in number

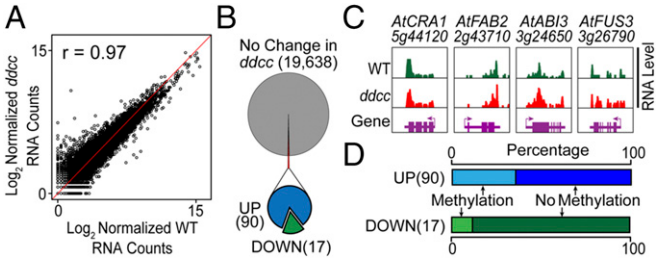

Fig. 8. Comparison between *Arabidopsis* wild-type and *ddcc* pmg seed transcriptomes. (*A*) Correlation between wild-type and *ddcc* seed mRNA accumulation levels. (*B*) Differentially expressed genes in *ddcc* seeds. DOWN, down-regulated; UP, up-regulated. (*C*) Genome browser view of major seed-specific mRNA accumulation patterns in *ddcc* and wild-type seeds. Arrows show the transcription directions. Gene names are defined in the legends to Figs. 5 and 6. (*D*) Methylation status of 5′ flanking 1-kb region of differentially expressed genes.

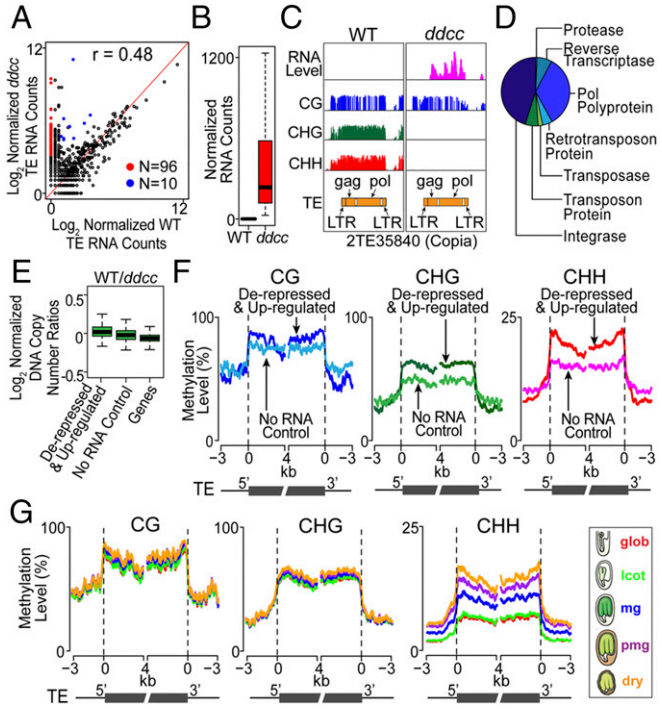

Fig. 9. TE transcriptional activity in *Arabidopsis* wild-type and *ddcc* pmg seeds. (*A*) Correlation between wild-type and *ddcc* seed TE RNA accumulations levels. Red and blue dots represent de-repressed and up-regulated TEs, respectively. (*B*) Box plots comparing de-repressed and up-regulated TE RNA accumulation levels in wild-type and *ddcc* seeds. (*C*) Genome browser view of the methylation pattern and RNA accumulation profile of a de-repressed Copia TE in pmg wild-type and *ddcc* seeds. LTR, long terminal repeat. (*D*) Major protein classes involved in TE transposition encoded by 106 de-repressed and up-regulated *ddcc* seed TE RNAs (*SI Materials and Methods*). (*E*) TE copy number changes between *ddcc* and wild-type seeds. Box plots show $\log_2$ ratios of normalized read depths from *ddcc* versus wild-type TEs. The gene control includes all genes in the *Arabidopsis* genome. (*F*) Methylation levels across 106 de-repressed and up-regulated TEs in wild-type seeds. The no RNA control used in *E* and *F* represent 106 randomly selected TEs which have (*i*) no detectable RNA wild-type reads and (*ii*) similar TE family distribution and lengths compared with the 106 de-repressed and up-regulated TEs. (*G*) *M*ethylation levels across 106 de-repressed and up-regulated TEs in Ws-0 wild-type seeds during *Arabidopsis* seed development.

within the germinating sdlg. By contrast, hypomethylated sites within the AX-PY and AX-PL retain their status as these regions give rise to the sdlg hypocotyl and leaf, respectively.

What role does the CHH-context methylation increase play in seed development? We investigated this issue by using an *Arabidopsis* mutant that is defective in both RdDM and non-RdDM pathways, and has no detectable non-CG methylation. *ddcc* seeds develop normally, have no detectable morphological defects, and germinate with frequencies indistinguishable from wild-type. In addition, the absence of non-CG context methylation does not appear to affect seed gene activity significantly because the mRNA profiles of *ddcc* and wild-type seeds are congruous with each other at both the qualitative and quantitative levels, including genes critical for seed development.

The simplest hypothesis for the role of increasing CHH-context methylation in seed development may be that it is a failsafe mechanism for reinforcing TE silencing in seeds. We favor this hypothesis because the programmed increase in CHH-context methylation during seed development targets TEs across the genome, and over 106 TEs rich in CHH sites at their ends (i.e., promoter regions) are de-repressed or up-regulated (>sixfold) at the RNA level in *ddcc* seeds. Although we found no evidence for these TEs moving, or increasing in copy number, in the generation we investigated, the up-regulated TE RNAs encode the requisite proteins (e.g., transposase, reverse transcriptase) that might unleash these TEs in subsequent generations (35). If this were the case, the consequences could be devastating for the seed and lead to lethality either before or after germination, or detrimental effects could accumulate more gradually over multiple generations. This hypothesis is consistent with the prevailing role for non-CG context methylation in higher plants (20).

One of the most intriguing aspects of our results is the observation that a large number of highly regulated soybean and *Arabidopsis* genes involved in major seed regulatory and metabolic events are localized in regions that are depleted of methylation (<5%) in any cytosine context during development, regardless of whether these genes are active or repressed. These include major regulatory genes, as well as genes encoding storage proteins and oil biosynthesis enzymes that are critical for maturation and germination. These seed genomic regions are similar to the DMVs observed in mammalian genomes during the differentiation of stem cells (28), and suggest strongly that genes present in these regions (e.g., storage protein genes) are not regulated directly by methylation changes, a conclusion that we made over three decades ago (26). Other highly regulated seed genes lie within regions containing heavily methylated TEs, or have methylation within their gene bodies, or both (Figs. 5*B* and 6*F*, Fig. S5, and Dataset S4). However, similar to genes in the seed DMVs, the methylation patterns of genes in these regions do not vary, whether the genes are active or not, indicating that methylation changes do not play a major role in regulating genes in these genomic regions as well (Figs. 5*B* and 6*F*, Fig. S5, and Dataset S4). This conclusion is enhanced by the fact that the loss CHG- and CHH-context methylation in pmg *ddcc* seeds does not significantly affect seed gene activity or development. We conclude that the major challenge for understanding the mechanisms that control seed development is to uncover *cis*-regulatory elements and corresponding TFs that control the spatial and temporal expression of essential seed genes, and determine how they are integrated into the genetic regulatory networks (39) that program seed development from generation to generation.

Materials and Methods

Specific details are contained within *SI Materials and Methods*.

Plant Material and LCM. The growth conditions and staging of soybean seeds [*Glycine max* (L.) cv. Williams 82] and *Arabidopsis* seeds [wild-type (Ws-0); *ddcc* (Col-0)] are detailed in *SI Materials and Methods*, following the procedures of Goldberg et al. (40) (soybean) and Le et al. (38) (*Arabidopsis*). AX, SC, and COTL were dissected manually from soybean seeds at the em and mm stages. LCM (41) was used to capture soybean: (*i*) cot-stage seed parts (AX, SC and COTL), (*ii*) em-stage seed parts (SC and COTL), (*iii*) em-stage COTL parenchyma tissues (ABPY and ADPY), (*iv*) em-stage SC layers (PA and PY), and (*v*) em-stage AX subregions and tissues (PL, PY, RT, and VS). EMB and SC were hand dissected from *Arabidopsis* seeds at the mg stage. Sample preparation procedures for the LCM experiments are detailed in *SI Materials and Methods*.

BS-Seq Library Construction, Methylome Sequencing, Data Processing, and Sequence Analysis. Genomic DNA was isolated from soybean and *Arabidopsis* whole seeds and hand-dissected seed parts using the DNEASY Plant Mini kit (Qiagen). DNA from seed tissue captured by LCM was isolated using the QIAMP FFPE DNA isolation kit (Qiagen). DNA was prepared for BS-Seq library preparation and methylome sequencing following the methods of Hsieh et al. (10) and Lister et al. (42), with modifications (*SI Materials and Methods*). DNA sequences were aligned to the soybean genome (version Wm82.a1; https://www.soybase.org) (43) or *Arabidopsis* genome (version TAIR10; https://www.arabidopsis.org/index.jsp) (44) using BS Seeker software (45), allowing up to two base mismatches. The sequencing depth for each cytosine in the reference genome was defined as the total number of detected cytosines (methylated C) or thymines (unmethylated C) across all reads. The specific procedures that we used to determine whether an individual cytosine site was methylated, as well as the bulk methylation level for

a given genomic feature (e.g., region, gene, TE) are described in detail in *SI Materials and Methods*.

RNA-Seq Library Construction, Sequencing, Data Processing, and Analysis. RNA was isolated from soybean whole seeds, seed parts, and sdlg using the Concert Plant RNA Reagent (Invitrogen) and treated with RNase-free DNase I (Ambion). Poly-A+ RNA was selected using oligo-dT_{25} magnetic beads (Dynabeads). Poly-A+ RNA was prepared for RNA-Seq library construction using the Illumina mRNA-Seq Sample Prep Kit (Illumina). For *Arabidopsis* pmg seeds, total RNA was used to generate double-stranded cDNA using the Ovation RNA-Seq System V2 (Nugen), and RNA-seq libraries were constructed using the Illumina TruSeq DNA Sample Prep Kit (Illumina). Bowtie (46) was used to map sequenced reads to: (*i*) the soybean genome (version Wm82.a1) and cDNA models (version Wm82.a1.v1.1) (https://www.soybase.org) (43) or (*ii*) the *Arabidopsis* genome (version TAIR10) and cDNA models (https://www.arabidopsis.org/index.jsp) (44). The EdgeR package (v3.18.1) (47) was used to identify differentially expressed RNAs.

***Arabidopsis ddcc* Seed Analysis.** *Arabidopsis ddcc* seeds were obtained from Steve Jacobsen, University of California, Los Angeles (15). Detailed information for characterization of (*i*) seed morphology, (*ii*) nuclear size, (*iii*) seed germination, and (*iv*) differentially expressed RNAs is presented in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Steve Jacobsen and his laboratory for *Arabidopsis ddcc* mutant seeds, excellent suggestions for how to analyze methylated transposable elements, and help with characterizing seed nuclei. This work was supported by a grant from the National Science Foundation Plant Genome Program (to R.B.G., M.P., and J.J.H.), and a National Institutes of Health Training Grant in Genomic Analysis and Interpretation T32HG002536 (to B.H.L.).

1. Goldberg RB, de Paiva G, Yadegari R (1994) Plant embryogenesis: Zygote to seed. *Science* 266:605–614.
2. Becker MG, Hsu S-W, Harada JJ, Belmonte MF (2014) Genomic dissection of the seed. *Front Plant Sci* 5:464.
3. Gehring M, Satyaki PR (2017) Endosperm and imprinting, inextricably linked. *Plant Physiol* 173:143–154.
4. Devic M, Roscoe T (2016) Seed maturation: Simplification of control networks in plants. *Plant Sci* 252:335–346.
5. Dhillon SS, Miksche JP (1983) DNA, RNA, protein and heterochromatin changes during embryo development and germination of soybean (*Glycine max* L.). *Histochem J* 15:21–37.
6. Larkins BA, et al. (2001) Investigating the hows and whys of DNA endoreduplication. *J Exp Bot* 52:183–192.
7. Sallon S, et al. (2008) Germination, genetics, and growth of an ancient date seed. *Science* 320:1464.
8. Bauer MJ, Fischer RL (2011) Genome demethylation and imprinting in the endosperm. *Curr Opin Plant Biol* 14:162–167.
9. Park K, et al. (2016) DNA demethylation is initiated in the central cells of *Arabidopsis* and rice. *Proc Natl Acad Sci USA* 113:15138–15143.
10. Hsieh T-F, et al. (2009) Genome-wide demethylation of *Arabidopsis* endosperm. *Science* 324:1451–1454.
11. Xing MQ, et al. (2015) Global analysis reveals the crucial roles of DNA methylation during rice seed development. *Plant Physiol* 168:1417–1432.
12. Raissig MT, Bemer M, Baroux C, Grossniklaus U (2013) Genomic imprinting in the *Arabidopsis* embryo is partly regulated by PRC2. *PLoS Genet* 9:e1003862.
13. Nodine MD, Bartel DP (2012) Maternal and paternal genomes contribute equally to the transcriptome of early plant embryos. *Nature* 482:94–97.
14. Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proc Natl Acad Sci USA* 97:4168–4173.
15. Stroud H, et al. (2014) Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nat Struct Mol Biol* 21:64–72.
16. Schmitz RJ, et al. (2013) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res* 23:1663–1674.
17. Song QX, et al. (2013) Genome-wide analysis of DNA methylation in soybean. *Mol Plant* 6:1961–1974.
18. Kim KD, et al. (2015) A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. *Plant Physiol* 168:1433–1447.
19. Goldberg RB, Hoschek G, Tam SH, Ditta GS, Breidenbach RW (1981) Abundance, diversity, and regulation of mRNA sequence sets in soybean embryogenesis. *Dev Biol* 83:201–217.
20. Kim MY, Zilberman D (2014) DNA methylation as a system of plant genomic immunity. *Trends Plant Sci* 19:320–326.
21. Du J, et al. (2012) Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* 151:167–180.
22. Kawashima T, Berger F (2014) Epigenetic reprogramming in plant sexual reproduction. *Nat Rev Genet* 15:613–624.
23. Martínez G, Panda K, Köhler C, Slotkin RK (2016) Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. *Nat Plants* 2:16030.
24. Ibarra CA, et al. (2012) Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science* 337:1360–1364.
25. Li S, Nielsen NC (2004) Endoreduplication during soybean seed development. PhD dissertation (Purdue University, West Lafayette, IN).
26. Walling L, Drews GN, Goldberg RB (1986) Transcriptional and post-transcriptional regulation of soybean seed protein mRNA levels. *Proc Natl Acad Sci USA* 83:2123–2127.
27. Comai L, Dietrich RA, Maslyar DJ, Baden CS, Harada JJ (1989) Coordinate expression of transcriptionally regulated isocitrate lyase and malate synthase genes in *Brassica napus* L. *Plant Cell* 1:293–300.
28. Xie W, et al. (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 153:1134–1148.
29. Lindstrom JT, Vodkin LO, Harding RW, Goeken RM (1990) Expression of soybean lectin gene deletions in tobacco. *Dev Genet* 11:160–167.
30. de Paiva G (1994) Transcriptional regulation of seed protein genes. PhD dissertation (University of California, Los Angeles).
31. Yadegari R (1996) Regional specification and cellular differentiaiton during early plant embryogenesis. PhD dissertation (University of California, Los Angeles).
32. Cokus SJ, et al. (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452:215–219.
33. Du J, Johnson LM, Jacobsen SE, Patel DJ (2015) DNA methylation pathways and their crosstalk with histone methylation. *Nat Rev Mol Cell Biol* 16:519–532.
34. van Zanten M, et al. (2011) Seed maturation in *Arabidopsis thaliana* is characterized by nuclear size reduction and increased chromatin condensation. *Proc Natl Acad Sci USA* 108:20219–20224.
35. Marí-Ordóñez A, et al. (2013) Reconstructing de novo silencing of an active plant retrotransposon. *Nat Genet* 45:1029–1039.
36. Kawakatsu T, Nery JR, Castanon R, Ecker JR (2017) Dynamic DNA methylation reconfiguration during seed development and germination. *Genome Biol* 18:171.
37. Danzer J, et al. (2015) Down-regulating the expression of 53 soybean transcription factor genes uncovers a role for SPEECHLESS in initiating stomatal cell lineages during embryo development. *Plant Physiol* 168:1025–1035.
38. Le BH, et al. (2010) Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. *Proc Natl Acad Sci USA* 107:8063–8070.
39. Peter IS, Davidson EH (2011) Evolution of gene regulatory networks controlling body plan development. *Cell* 144:970–985.
40. Goldberg RB, Hoschek G, Ditta GS, Breidenbach RW (1981) Developmental regulation of cloned superabundant embryo mRNAs in soybean. *Dev Biol* 83:218–231.
41. Kerk NM, Ceserani T, Tausta SL, Sussex IM, Nelson TM (2003) Laser capture microdissection of cells from plant tissues. *Plant Physiol* 132:27–35.
42. Lister R, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462:315–322.
43. Schmutz J, et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183.
44. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.
45. Chen P-Y, Cokus SJ, Pellegrini M (2010) BS seeker: Precise mapping for bisulfite sequencing. *BMC Bioinformatics* 11:203–208.
46. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25–R34.
47. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.
48. Le BH, et al. (2007) Using genomics to study legume seed development. *Plant Physiol* 144:562–574.
49. Saski C, et al. (2005) Complete chloroplast genome sequence of *Gycine max* and comparative analyses with other legume genomes. *Plant Mol Biol* 59:309–322.
50. Gan X, et al. (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477:419–423.
51. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB (1982) Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* 162:729–773.
52. Sanger F, et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265:687–695.
53. Lister R, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133:523–536.
54. Robinson JT, et al. (2011) Integrative genomics viewer. *Nat Biotechnol* 29:24–26.
55. Krzywinski M, et al. (2009) Circos: An information aesthetic for comparative genomics. *Genome Res* 19:1639–1645.
56. Du J, et al. (2010) SoyTEdb: A comprehensive database of transposable elements in the soybean genome. *BMC Genomics* 11:113–119.
57. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
58. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
59. Katari MS, et al. (2010) VirtualPlant: A software platform to support systems biology research. *Plant Physiol* 152:500–515.
60. Lee T-F, et al. (2012) RNA polymerase V-dependent small RNAs in *Arabidopsis* originate from small, intergenic loci including most SINE repeats. *Epigenetics* 7:781–795.
61. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* 19:1586–1592.
62. Moissiard G, et al. (2012) MORC family ATPases required for heterochromatin condensation and gene silencing. *Science* 336:1448–1451.
63. Yadegari R, et al. (1994) Cell differentiation and morphogenesis are uncoupled in *Arabidopsis* raspberry embryos. *Plant Cell* 6:1713–1729.

Supporting Information

Lin et al. 10.1073/pnas.1716758114

SI Materials and Methods

Plant Growth and Tissue Collection. Soybean plants [*Glycine max* (L.) cv. Williams 82] were grown at 22 °C under long-day conditions with a 16-h light to 8-h dark cycle in the University of California, Los Angeles Plant Growth Center. Seeds were staged based on length, weight, and embryonic characteristics, such as shape and color (19). Hand-made drawings of soybean seeds and their corresponding EMBs at all developmental stages used in our experiments are shown in Fig. 1. glob, cot, em, mm, lm, pd1, and pd2 seeds had lengths of 1.0–1.5 mm, 3.0–3.5 mm, 6.0–7.0 mm, 11.0–12.0 mm, 12.0–14.0 mm, 8.0–10.0 mm, and 8–9 mm, respectively. In addition, mm, lm, pd1, pd2, and dry seeds weighed 150–250 mg, 230–350 mg, 150–260 mg, 120–150 mg, and 155 mg, respectively. mm EMBs had green COTL and a yellow AX tip, while lm EMBs were completely yellow (Fig. 1). sdlg were collected 6 d after imbibition, having primary roots, hypocotyls [average length 8 cm, flat green COTL (average length 19 mm)], and unifoliate leaves (Fig. 1). Postgermination COTL were dissected from sdlg 6 d after imbibition and weighed 250–400 mg. AX, COTL, and SC were manually separated from em and mm seeds (Fig. 3*B*). AX and COTL were harvested without the PL. Whole seeds and seed parts were harvested, frozen in liquid nitrogen, ground to a fine powder, and stored at −80 °C before DNA or RNA isolation.

Arabidopsis thaliana of ecotype Wassilewskija (Ws-0) were grown, and seeds were collected at glob, lcot, mg, pmg, and dry stages from 3-mo-old plants, as described in detail previously (38). Artistic renditions of *Arabidopsis* seeds at all stages of development are shown in Fig. 6*A*. EMB and SC were hand-dissected from mg-stage seeds (Fig. 6*D*). Leaves were collected from 4-wk-old plants. *A. thaliana* ecotype Columbia (Col-0) pmg seeds were collected from 3-mo-old *ddcc* mutant (15) and wild-type plants, and leaves were collected from 3-wk-old postgermination *ddcc* mutant and wild-type plants. Whole seeds and seed parts were harvested, frozen in liquid nitrogen, ground to a fine powder, and stored at −80 °C before DNA or RNA isolation.

LCM of Soybean Seed Regions and Tissues. em seeds were harvested, cut in half transversely, fixed in ethanol:acetic acid [3:1 (vol/vol)], dehydrated, infiltrated, and embedded in paraffin solution containing Paraplast-X-Tra tissue embedding medium (Fisher Scientific) according to the methods of Kerk et al. (41). Ten-micrometer paraffin cross-sections were prepared for each seed half using a Reicher-Jung 4 Rotary Histocut 820 Microtome, and floated in diethylpyrocarbonate-treated water to stretch ribbons containing seed serial sections. Seed sections were placed on PEN-foil slides (Leica Microsystems) and deparaffinized with two consecutive 2-min xylene treatments before LCM. Tissue sections were captured using a Leica LMD6000 microdissection scope into a PCR tube cap containing DNA isolation solution included in the FFPE DNA isolation kit (Qiagen).

Isolation of Endoreduplicated and Nonendoreduplicated COTL Tissues. The endoreduplication studies of Li and Nielsen (25) demonstrated that cells of the em-stage COTL ABPY tissue have undergone endoreduplication, whereas the COTL ADPY cells have not. To isolate parenchyma ABPY and ADPY tissues, ∼350 μm from the em-stage cotyledon ends was excluded from each cross-section, and only the middle sections of em seeds with COTL sizes 2,900–3,600 μm were used for LCM. The epidermis layer of the COTL was excluded from the ABPY collection, but was captured with the ADPY layer. Four ABPY cell layers and three ADPY cell layers were captured, respectively (Fig. 4*A*).

Isolation of SC Tissue Layers. To compare the methylation levels of different SC layers, SC-PY and SC-PA tissues were captured using LCM from the same em-stage paraffin sections used to capture the ABPY and ADPY tissue layers (Fig. 3*E*).

Isolation of cot-Stage Seed Parts. cot-stage seeds were harvested, fixed as described for em-stage seeds, and sectioned longitudinally. SC, AX, and COTL were captured from all sections using LCM (Fig. 3*A*). To avoid endosperm contamination, we used the laser to destroy the aleurone layer before capturing the SC.

Isolation of em-Stage Seed Parts, and AX Subregions and Tissues. em-stage seeds were harvested, fixed, and sectioned longitudinally. The entire SC and COTL were captured from all sections using LCM (Fig. 3*D*). PL, PA, PY, and RT were isolated from the same AX sections using LCM (Fig. S4*A*).

BS-Seq Library Construction. Approximately 100–1,000 ng of genomic DNA was used for BS sequencing library preparation following the methods of Hsieh et al. (10), with modifications. Specifically, 3 ng of unmethylated *cI857 Sam7* λ DNA (GenBank Accession no. NC_001416; Promega) was spiked-in with genomic DNA before DNA sonication to serve as an internal control for estimating BS conversion efficiency. Adapter-ligated genomic DNA was subjected to two rounds of BS treatment using the EpiTect Bisulfite Conversion kit (Qiagen). BS-treated DNA was purified and amplified for 10 cycles using ExTaq (Takara) DNA polymerase. PCR-amplified DNA fragments were size-selected using the AMpure XP beads (Beckman).

RNA-Seq Library Construction. Approximately 100 ng of soybean poly-A+ RNA was used for RNA-Seq library preparation according to the Illumina RNA-Seq Sample Prep Kit (Illumina). For *Arabidopsis* pmg wild-type and *ddcc* seeds, 25 ng total RNA was used to generate double-stranded cDNA using Ovation RNA-Seq System v2 (Nugen), and then 1 μg of double-stranded cDNA was used for RNA-Seq library preparation with the Illumina TruSeq DNA Sample Prep Kit (Illumina).

Small RNA-Seq Library Construction. Total RNA was isolated from em AX, COTL, and SC using the Concert Plant RNA Reagent (Invitrogen), according to the manufacturer's instructions. Appropriately 250 ng total RNA was used for the TruSeq Small RNA Sample Preparation kit (Illumina), and 15 PCR cycles were used for the final PCR enrichment step.

Illumina Next-Generation Sequencing. Single-end 50-bp reads were generated for the RNA-Seq and small RNA-Seq libraries, whereas 100-bp reads were generated for BS-Seq library using the Illumina Genome Analyzer IIx or HiSeq 2000 sequencing machines in the University of California, Los Angeles Genome Sequencing Center or Broad Stem Cell Research Center High Throughput Sequencing Core. A ϕX174 DNA control was spiked into each library by the sequencing facility before cluster formation and sequencing.

BS-Seq Data Processing and Analysis. Sequences were aligned to a reference genome using the BS Seeker program (45) allowing up to two mismatches. We used soybean genome build version Wm82.a1 (https://www.soybase.org) as a reference, which consists of scaffold sequences, including the 20 nuclear chromosomes, mitochondrial DNA, and unanchored sequences (43). Scaffolds containing chloroplast sequences were replaced with the 152,218-bp fully sequenced

soybean chloroplast genome sequence (DQ317523) (49). For *Arabidopsis*, we used both TAIR10 Columbia (Col-0) (https://arabidopsis.org/index.jsp) and Ws-0 (mtweb.cs.ucl.ac.uk/mus/www/19genomes/) as reference genomes (44, 50). We compared the Col-0 and Ws-0 genomes and found that: (*i*) only 0.4% of cytosines were affected by single nucleotide polymorphisms that could affect their methylation status, and (*ii*) similar CG-, CHG-, and CHH-context bulk methylation results across seed development were obtained with both *Arabidopsis* ecotypes. That is, the results shown in Fig. 6 are valid for both Col-0 and Ws-0 ecotypes. Finally, we used the 48,502-bp *cI857 Sam7* λ genome (NC_001416) (51) and the 5,386-bp ϕX174 genome (NC_001422) (52) to map reads from our λ and ϕX174 DNA controls.

Only reads that mapped uniquely to the reference genomes were retained for detailed analysis. The BS-Seeker output was subjected to postprocessing that consisted of two steps: (*i*) clonal reads, or reads containing identical 5′ mapped positions and exact nucleotide sequences, were collapsed and all but one read was retained to reduce PCR amplification bias for each library; and (*ii*) reads containing three or more consecutive cytosines in the CHH context were removed as they are likely not bisulfite converted (32).

Analysis of BS Conversion Efficiency. During BS conversion experiments, unmethylated cytosines might not be converted to uracil (thymine) and will appear erroneously as methylated cytosines in the final sequencing reads. To assess the extent of nonconversion during BS treatment, we spiked in unmethylated λ DNA into each genomic DNA sample during the library preparation step. We obtained at least 300× coverage of the λ genome, and detected up to 99.98% of the genomic cytosines, representing comprehensive coverage of the λ genome. Overall, we obtained excellent BS conversion efficiency of unmethylated cytosine to uracil (thymine), averaging 99.55% over 49 BS-Seq libraries, which were similar to conversion rates obtained by others using the same BS kits (42) (Dataset S1).

Estimate of Genome Sequencing Coverage. Because the soybean genome is an ancient polyploid with >60% of the genome represented by repetitive sequences (43), we asked what fraction of the nuclear genome sequences we could detect and map uniquely with 100-bp sequencing reads. Using a sliding-window approach, we generated >950 million 100-bp reads with a 99-bp overlap covering the entire genome. We next collapsed and removed identical redundant reads leaving 805 million (~85%) reads that are unique. The 805 million reads were aligned to the genome using BS-Seeker to mimic our data processing pipeline allowing for no mismatch. Approximately 782 million reads aligned to the genome, uniquely covering 857 million bases, or 90% of the nuclear genome, including ~325 million cytosines or 90% of the genomic cytosines. These results suggest that although soybean is an ancient polyploid with a large repeat-rich genome, most of the sequences have diverged significantly and can be distinguished clearly from 100-bp reads using BS-Seeker, indicating that our whole-genome BS sequencing can interrogate most, if not all, of the genome sequences. We used 90% of cytosines in the soybean genome as the basis for the calculation of cytosines detected in Dataset S1 by our BS sequencing. We didn't carry out an analogous simulation for *Arabidopsis*, as it contains much fewer repetitive sequences in its genome, and used the total number of genomic cytosines as the basis for the numbers in Dataset S1 (44).

Determination of Whether a Genomic Site Is Methylated or Not. Whole seeds were used for most of our BS-Seq experiments and, therefore, the BS-Seq reads represent the genomes of different cells within the seeds. If the methylation status of a given cytosine site in the different cell types differs, then BS-Seq will detect both the methylated and unmethylated cytosines. That is, not all equivalent genomic sites within different seed cells might be methylated. For example, it is possible that within a seed only a fraction of the cells might be methylated at a specific cytosine location resulting in a mixture of methylated and unmethylated cytosines in the BS-Seq data for that genomic site. This also applies to tissues captured by LCM, as not all cells within a tissue might have the same developmental equivalency and methylation states.

We used the following approach to assign the methylation status of a given seed genomic site. For each cytosine site in the reference genome, the total number of cytosine reads (representing methylated cytosines) and thymine reads (representing unmethylated cytosines) were summarized using custom scripts. We then eliminated cytosine sites that had only one mapped read (cytosine or thymine). Next, we used a statistical test (binominal test, $P < 0.05$) and only cytosines that passed this filter were used for downstream analysis to ensure that they were not false positives because of: (*i*) nonconversion of unmethylated cytosines to thymines (i.e., they remain cytosines after BS treatment and scored falsely as methylated in the BS-Seq analysis) and (*ii*) the amount of sequence coverage (53). We used 0.5% as the value for nonconversion of unmethylated cytosines in the binomial test, which was similar that obtained in our λ genome spiked-in controls (Dataset S1). In general, genomic sites with two or more cytosine reads passed our statistical filter and were considered methylated. By contrast, genomic sites were considered unmethylated if there was at least one thymine BS-Seq read, indicating that BS converted an unmethylated cytosine to uracil (thymine) at that site. By using these criteria, we were able to assign the methylation status of each cytosine site in the genome regardless of which seed cell from which it was derived.

Calculation of Average or Bulk Methylation Levels Across the Entire Seed. Bulk methylation is the average cytosine methylation percentage for all filtered reads, irrespective of their genomic sites, and represents the average methylation levels across all genomes within the entire seed or tissue. Bulk methylation levels for all cytosines (C), and those in CG, CHG, and CHH contexts, were calculated by using the formula $[C/(C + T) \times 100]$, where C and T represent total cytosine (methylated) or thymine (unmethylated) reads (10). For example, 10 methylated cytosines of 100 detected cytosines in all reads mapped within a specific genomic element [e.g., 500-kb genomic windows (soybean), 100-kb windows (*Arabidopsis*), genes, or TEs] is a 10% bulk cytosine methylation level. Bulk methylation levels were used for construction of methylation box plots and pair-wise seed-stage comparisons presented in this paper (e.g., see Fig. 2*A*). Each box plot represents the middle 50% of methylation levels. The black bar in Fig. 2*A* indicates the median methylation level, while the whiskers indicate 1.5 times the box length.

Visualization of Methylation Levels. The methylation levels of cytosine sites were converted to bigwig format and viewed using the Integrative Genomics Viewer (54) (e.g., Fig. 5*B*). The heat maps in Fig. 2*B* and Fig. 6C were drawn using Circos (55). In the heat maps, chromosome, centromere, and pericentromere coordinates were obtained from Phytozome (phytozome.net) for soybean or from TAIR10 (www.arabidopsis.org) for *Arabidopsis*. Tracks representing the densities of genes and TEs in 500- or 100-kb windows along the genome were based on soybean version W82.a1.v1.1 annotations from SoyBase (https://www.soybase.org), the SoyTEdb database (56), or the TAIR10 genome release (www.arabidopsis.org).

RNA-Seq Data Processing and Analysis. Illumina raw reads were first filtered using the Illumina purity filter, and then trimmed at their 5′ and 3′ ends based on positions with error rates >0.1%. rRNA reads were identified by mapping trimmed reads against a rRNA database using Bowtie (v0.12.7) and then removed from further analysis (46). The remaining high-quality reads were mapped to either the soybean or *Arabidopsis* reference genomes and cDNA models using Bowtie, allowing up to two mismatches. Only uniquely mapped reads (i.e., reads that map to one unique genomic locus) were used

for subsequent analysis. Read counts for each gene model were computed using a customized script. RNA-Seq expression values for each gene within a dataset were normalized as RPKM (57). Genes and TEs were identified as differentially expressed in the *ddcc* mutant relative to wild-type with the EdgeR package (v3.18.1) (47) using the following filtering parameters: (*i*) genes: a corrected *P* value < 0.001 and >fivefold difference; and (*ii*) TEs: a corrected *P* value < 0.01 and detection in both replicates. We used BEDTools (58) to obtain normalized coverage per base from the alignments of RNA-Seq data, converted to BAM format, and viewed using the Integrative Genomics Viewer (54). VirtualPlant 1.3 (59) was used for *Arabidopsis* gene Gene Ontology enrichment analysis with a cut-off FDR value < 0.01. Proteins encoded by the 106 up-regulated and de-repressed TE RNAs were determined using the corresponding reads as query sequences in translated BLAST (blastx) against *Arabidopsis* TE proteins in the National Center for Biotechnology Information protein database (https://www.ncbi.nlm.nih.gov/protein/) with 1E-4 as a cut-off criterion.

Small RNA Sequence Processing and Analysis. Quality filtered small RNA sequences were trimmed to remove the adapter sequences. Trimmed reads that mapped to soybean ribosomal RNA (rRNA) and transfer RNA sequences were removed, and then 18- to 24-nt reads were kept for further analysis. These reads were aligned to the soybean genome (version Wm82.a1) using Bowtie (v0.12.7) with no mismatches, but allowing for matches to multiple positions within the genome (60). The filtered reads were then sorted into 21-, 22-, and 24-nt siRNAs (Fig. 3*G*). We used BEDTools (58) to obtain normalized counts within TEs from all mapped sequences for 21-, 22-, and 24-nt small RNAs in em-stage AX, COTL, and SC seed parts (Fig. 3*G*).

Copy Number Analysis. The sequencing read depth method of Yoon et al. (61) was used to detect copy number variation between: (*i*) soybean seed endoreduplicated and nonendoreduplicated ABPY and ADPY COTL regions (Fig. 4*C*) and (*ii*) *Arabidopsis* TEs in *ddcc* and wild-type pmg seeds (Fig. 9*E*). We used BEDTools (58) to obtain normalized read depth per base from the alignments of BS-Seq data. The average normalized read depth in a soybean genomic window (500 kb), or an *Arabidopsis* TE, was calculated from single base normalized read depth.

***Arabidopsis* Nuclear Size Assay.** Nuclear size assays were performed according to Moissiard et al. (62) with the following modifications. Two hundred COTL from hand-dissected *Arabidopsis* pmg seeds, or 0.5 g of leaves from 2-mo-old plants were fixed in Tris buffer (10 mM Tris pH 7.5, 10 mM EDTA, 100 mM NaCl) containing 4% paraformaldehyde for 20 min and washed twice in Tris buffer without paraformaldehyde. Samples were ground in 45 μL lysis buffer (15 mM Tris pH 7.5, 2 mM EDTA, 0.5 mM spermine, 80 mM KCl, 20 mM NaCl, 0.1% Triton X-100) using a glass grinder and filtered through a 35-μm cell strainer. The suspension containing nuclei was added to sorting buffer (100 mM Tris pH 7.5, 50 mM KCl, 2 mM $MgCl_2$, 0.05% Tween-20, 20.5% sucrose) and transferred to slides to air dry for 1 h. Slides were postfixed in a PBS (10 mM sodium phosphate, pH 7, 143 mM NaCl) containing 4% paraformaldehyde for 20 min followed by three washes with PBS alone. The mounting medium Vectashield containing DAPI (Vector Laboratories H-1200) was added to the slides to hold nuclei in place between the coverslip and the slide and to stain the nuclei. Nuclei were observed using a Zeiss Imager D2 microscope and more than 110 nuclei were analyzed for each genotype.

***Arabidopsis* Seed Morphology and Germination Assays.**

Seed development. The development of *Arabidopsis* wild-type and *ddcc* seeds was characterized using the procedures of Yadegari et al. (63) with the following modifications. Siliques were fixed in ethanol:acetic acid [9:1 (vol/vol)] overnight followed by two washes in 90% and 70% ethanol for 1 h, respectively. Siliques were cleared with a chloral hydrate/glycerol/water solution [8:1:2, (w/vol/vol)] for 1 h, and seeds were visualized using a Zeiss Imager D2 microscope equipped with Nomarski optics.

Seed germination analysis. *Arabidopsis* dry seeds were washed with 70% ethanol twice, 50% bleach twice, and sterilized water four times, then resuspended in sterilized water and put in refrigerator for 3 d. Fifty seeds were paced in a grid alignment on five replicate agar plates (Phytoblend; Caisson Labs) containing Murashige and Skoog medium, pH 5.7. Plates were put in a growth chamber at 22 °C with a 16-h light to 8-h dark cycle, and sdlg numbers were counted after 4 d.

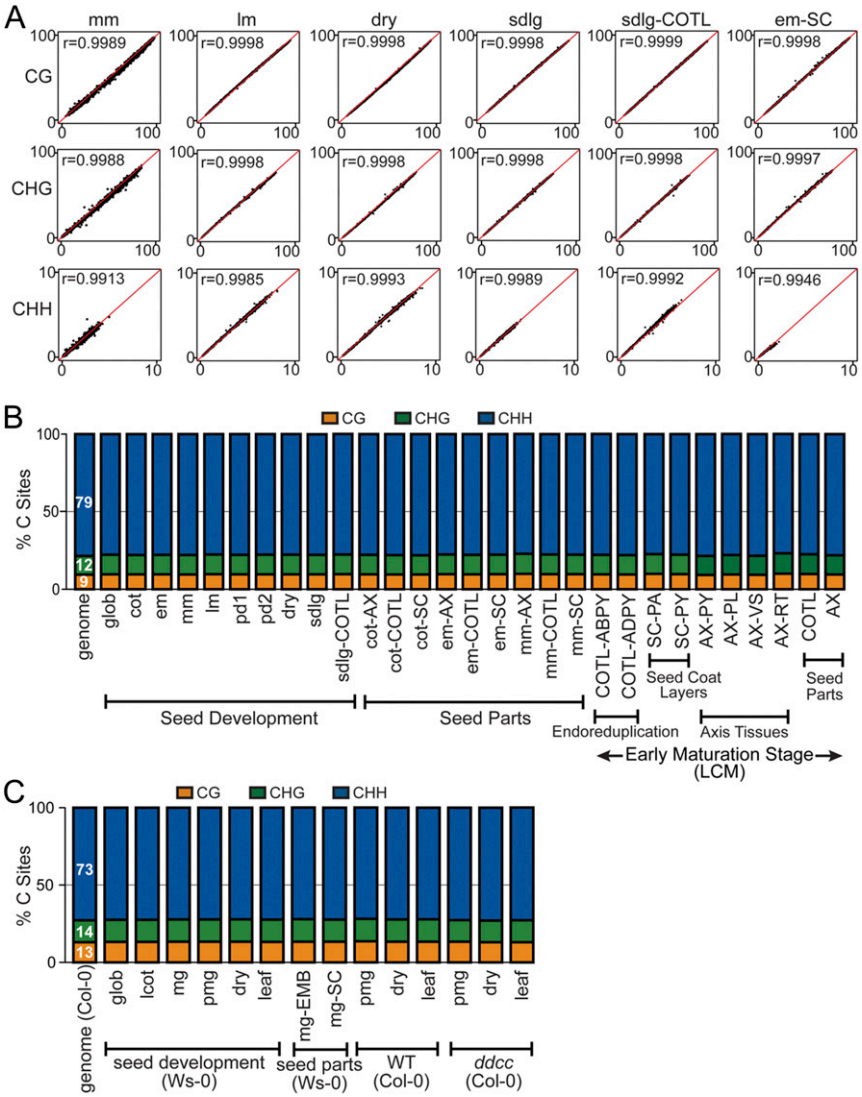

Fig. S1. Quality of BS-Seq methylome libraries. (*A*) Correlation coefficients between biological replicates of soybean seed BS-Seq libraries. The average methylation levels in 500-kb windows across the genome from biological replicates with similar sequencing depths were used to determine the correlation coefficients. Proportions of cytosine bases in CG, CHG, and CHH contexts for the soybean genome (*B*) and *Arabidopsis* genome (*C*) were determined from the BS-Seq results reported here, the soybean genome sequence (version Wm82.a1) (https://www.soybase.org) (43), and the *Arabidopsis* genome sequence (version TAIR10) (https://www.arabidopsis.org/index.jsp). See Table 1 for definition of abbreviations.

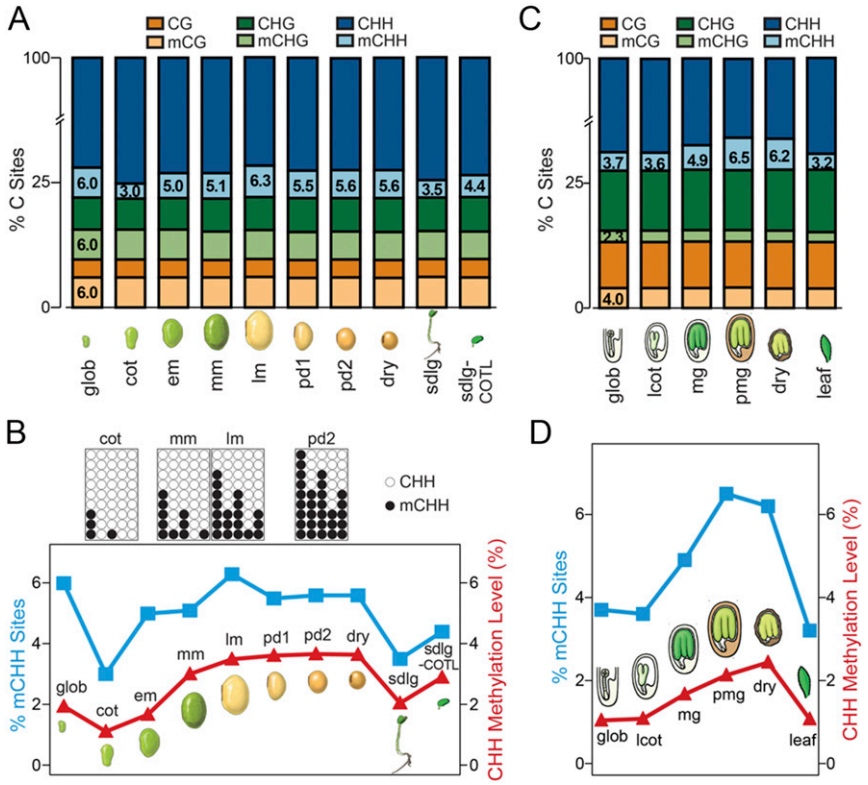

Fig. S2. DNA methylation changes during seed development and germination. (*A*) Changes in the proportion of methylated cytosine sites in CG, CHG, and CHH contexts during soybean seed development. CG, CHG, and CHH represent unmethylated cytosines, whereas mCG, mCHG, and mCHH indicate methylated cytosines. Numbers represent the proportions of methylated cytosine sites in each sequence context during seed development. (*B*) Comparison between the changes in methylated cytosine sites (blue) and bulk methylation percentages (red) in CHH context during soybean seed development. Methylated cytosine site data were taken from values in *A*, and those for bulk methylation percentages were taken from Dataset S2. The latter represents mean DNA methylation levels in 500-kb windows across the soybean genome. Boxes represent a heuristic model for the changes in methylated cytosine sites during seed development. Each circle along the bottom row represents one cytosine site in the genome, whereas the vertical circles represent cytosines in the same genomic positions in different seed cells. During seed development individual CHH-context sites across the genome become methylated (dark circles), resulting in an increase in the absolute number of methylated CHH sites. These same sites become methylated stochastically in different seed cells during development, resulting in an increase in bulk methylation percentage. (*C*) Changes in the proportion of methylated cytosine sites in CG, CHG, and CHH contexts during *Arabidopsis* seed development similar to those presented for soybean seeds in *A*. (*D*) Comparison between the changes in methylated cytosine sites (blue) and bulk methylation percentages (red) in CHH context during *Arabidopsis* seed development similar to that presented for soybean in *B*. See Table 1 for abbreviations of soybean and *Arabidopsis* seed stages.

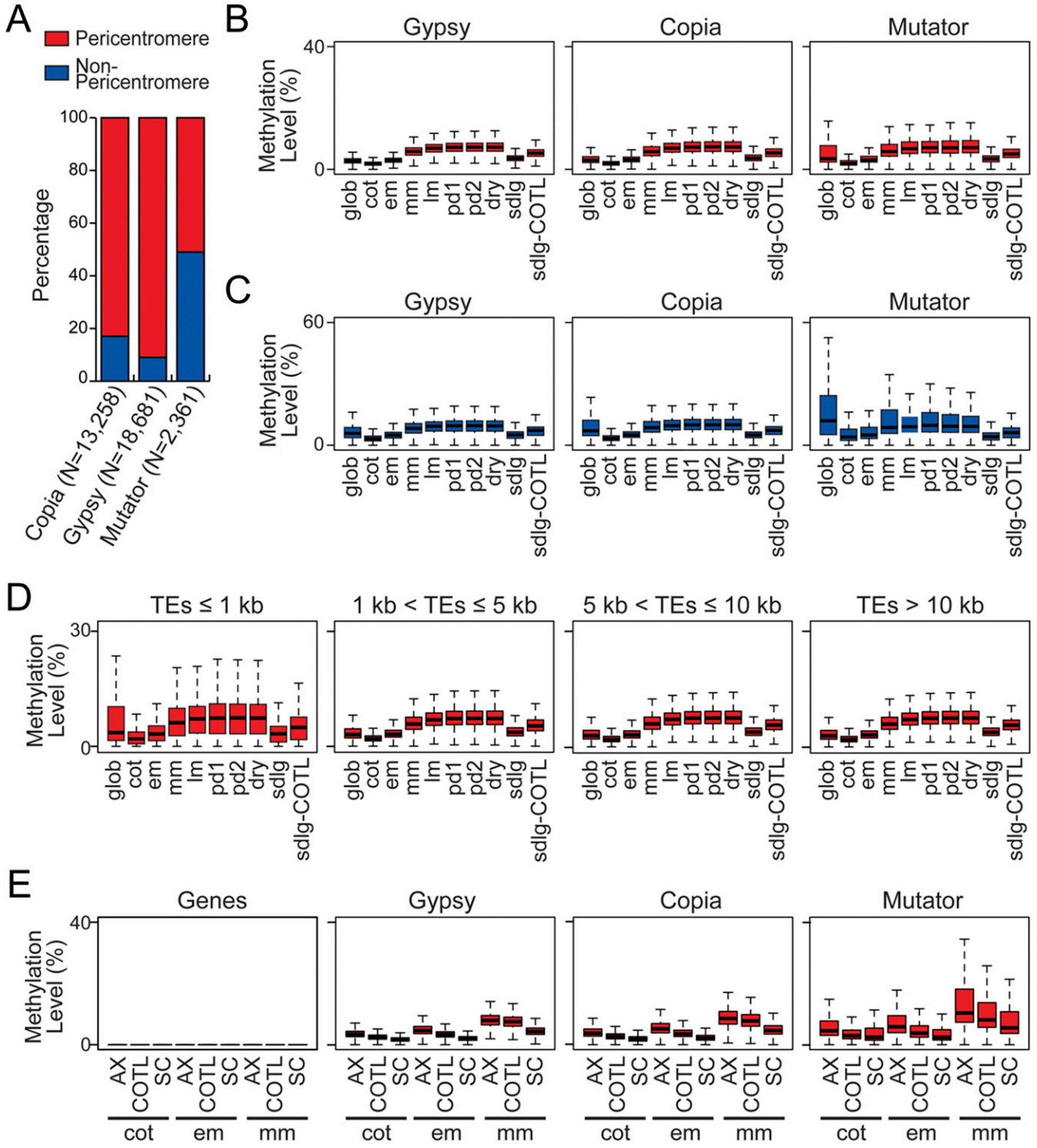

Fig. S3. Changes in CHH-context DNA bulk methylation levels in different TE classes during soybean seed development and germination. (*A*) Proportion of TE classes in chromosomal pericentromeric and arm regions. Box plots of methylation levels for different TE classes in pericentromere regions (*B*) and arm regions (*C*) during seed development. (*D*) Box plots of methylation levels for different-sized TEs across seed development. (*E*) Box plots of methylation levels for genes and TEs in different seed parts. See Table 1 for abbreviations of seed stages and parts.

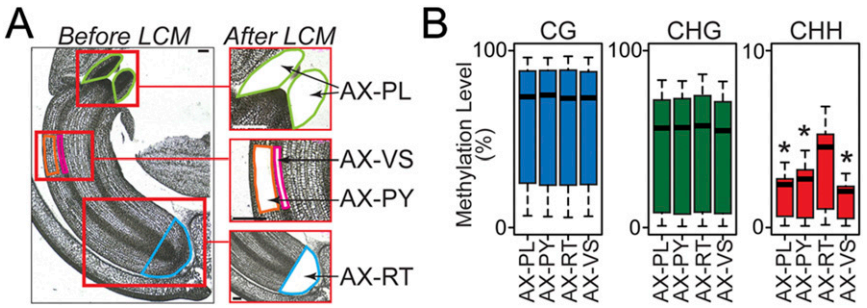

Fig. S4. Comparison of methylation levels within different soybean AX regions. (*A*) Paraffin sections of early maturation stage AX-PL, AX-PA, AX-VS, and AX-RT regions before and after capture by LCM. (Scale bar, 200 μm.) (*B*) Box plots of DNA methylation levels in 500-kb windows across the genome in different AX tissues. Asterisks indicate significant comparisons between RT and other AX tissues (*t* test, $P < 0.001$ and fold change > 1.5) (Dataset S3).

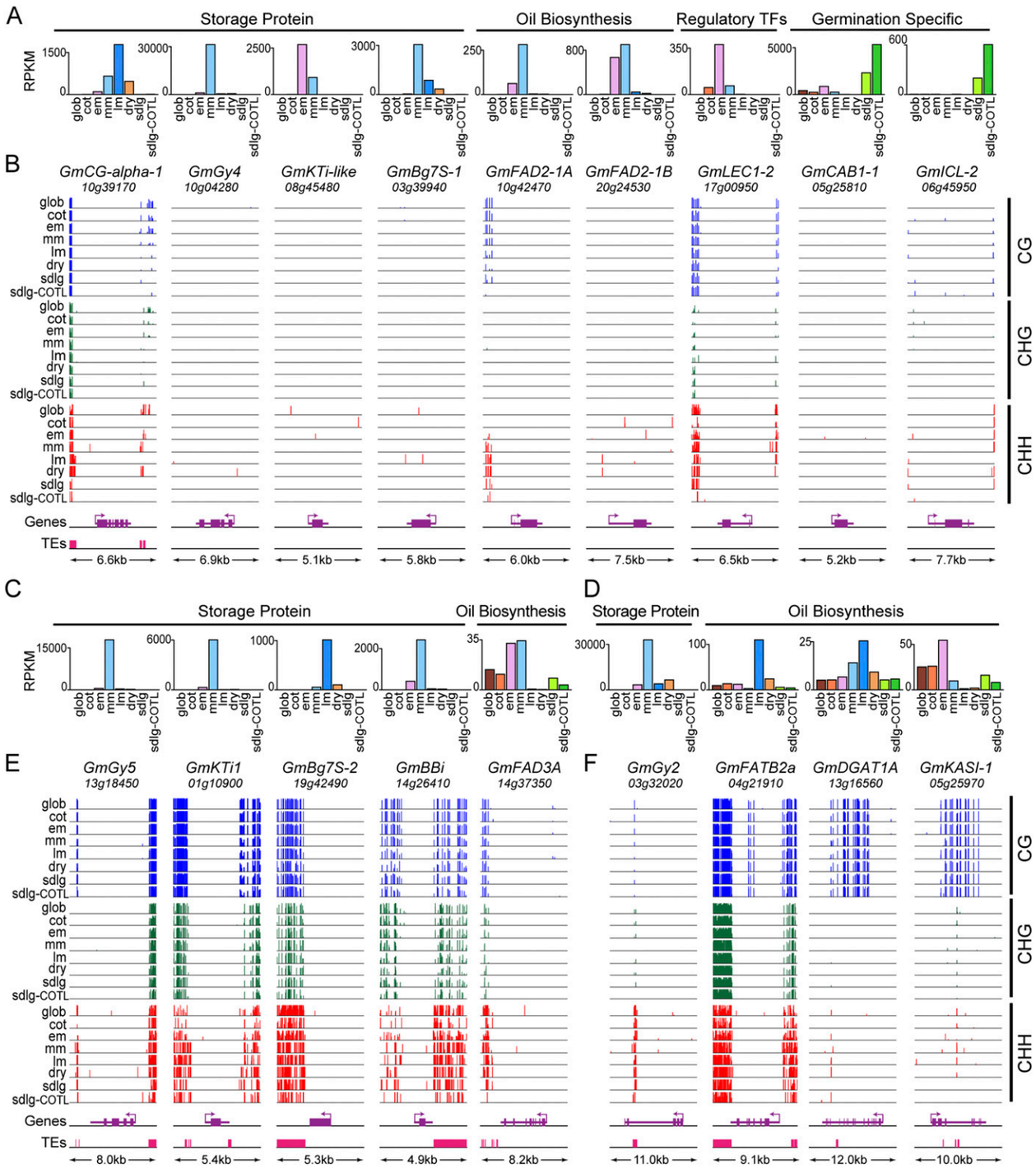

Fig. S5. Methylation levels and mRNA accumulation patterns of major soybean seed-specific gene classes during seed development and germination. (*A*, *C*, and *D*) mRNA accumulation levels for major seed and germination mRNAs. RPKM were taken from the Goldberg-Harada soybean (*i*) whole-seed RNA-Seq dataset GEO number GSE29163 (37), and (*ii*) cotyledon-specific RNA-Seq dataset GSE29134 (sdlg-COTL). (*B*, *E*, and *F*) Methylation levels of CG-, CHG-, and CHH-context cytosine sites are shown in genome browser view (vertical lines). Genes in *A* and *B* have either zero or <5% cytosine methylation in their gene bodies and at least 1 kb of upstream and downstream regions. Genes in *C* and *E* have no detectable cytosine methylation in their gene bodies, but have methylated TEs within 1 kb of flanking gene regions. Genes in *D* and *F* have methylated cytosines within their gene bodies. Additional soybean seed and germination genes with similar methylation patterns [described as classes (*i*), (*ii*), and (*iii*)] are listed in Dataset S4. Gene structures, transcription directions (arrows), and TEs are shown below each genome browser view. Adjacent genes are not shown. The size of each genomic region, including 2 kb of gene flanking region, is shown at the bottom of the browser views. *GmBBi*, Bowman Birk inhibitor; *GmBg7S-1*, basic 7S globulin-1; *GmBg7S-2*, basic 7S globulin-2; *GmCAB1-1*, Chlorophyll A/B Protein 1-1; *GmCG-α-1*, *Gm*β-conglycinin-α-1; *GmDGAT1A*, diacylglycerol acyl-transferase 1A; *GmFAD2-1A*, oleoyl desaturase 2-1A; *GmFAD2-1B*, oleoyl desaturase 2-1B*; Gm FAD3A*, linoleoyl desaturase 3A; *GmFATB2a*, fatty acyl-ACP thioesterase B2a; *GmGy2*, glycinin2; *GmGy4*, glycinin4; *GmGy5*, glycinin5; *GmICL-2*, isocitrate lyase-2; *GmKASI-1*, 3-ketoacyl-ACP synthase I-1; *GmKTi1*, Kunitz trypsin inhibitor 1; *GmKTi-like*, Kunitz trypsin inhibitor-like; *GmLEC1-2*, Leafy cotyledon 1-2. See Table 1 for developmental stage abbreviations.

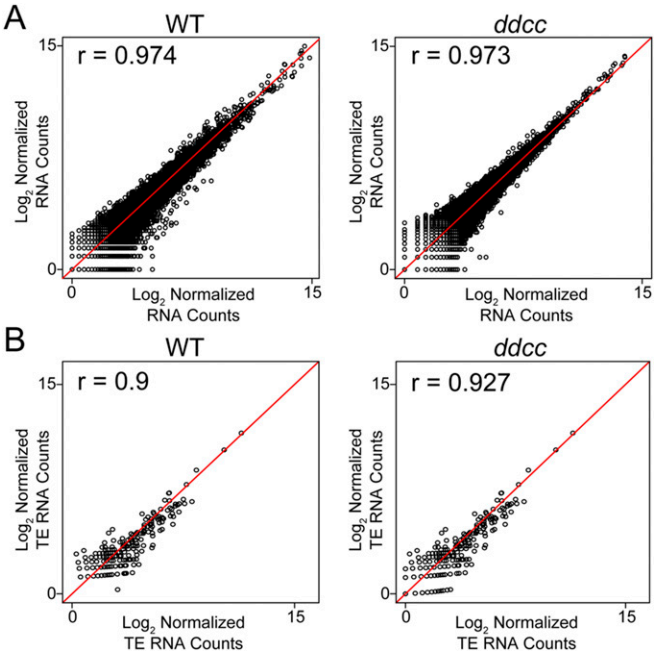

Fig. S6. Comparison between biological replicates of *Arabidopsis* wild-type (WT) and *ddcc* pmg seed gene (*A*) and TE (*B*) transcriptomes.

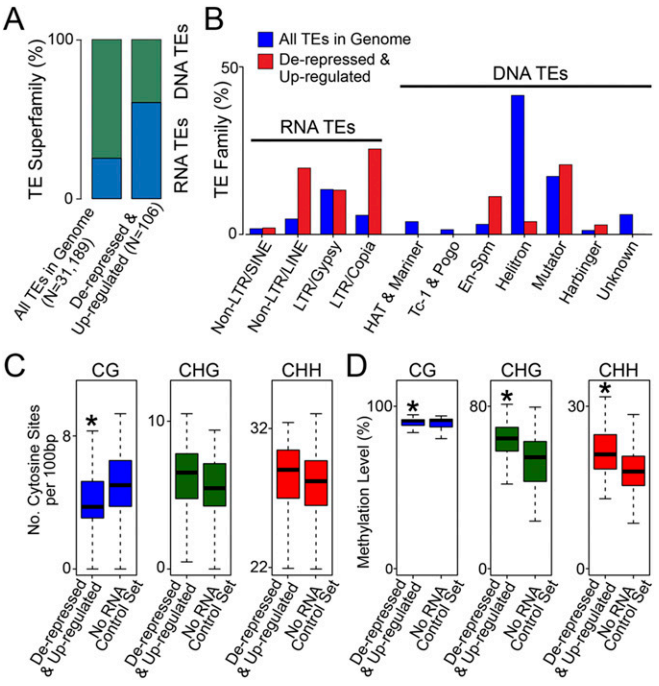

Fig. S7. De-repressed and up-regulated TEs in *Arabidopsis ddcc* pmg seeds. Representation of RNA and DNA TEs (*A*) and individual TE families (*B*) in *ddcc* seed de-repressed and up-regulated TEs. (*C*) Box plots of the number of cytosine sites per 100 nt for each sequence context in *ddcc* seed de-repressed and up-regulated TEs. (*D*) Bulk DNA methylation level box plots for each sequence context in *ddcc* seed de-repressed and up-regulated TEs. The no RNA control used in *C* and *D* represent 106 randomly selected TEs, which have (*i*) no detectable RNA wild-type reads and (*ii*) similar TE family distribution and lengths compared with the 106 de-repressed and up-regulated TEs (Fig. 9). The asterisks indicate statistically significant comparisons between the no RNA control TEs and de-repressed and up-regulated TEs (*t* test, $P < 0.01$).

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)
Dataset S4 (XLSX)
Dataset S5 (XLSX)
Dataset S6 (XLSX)
Dataset S7 (XLSX)
Dataset S8 (XLSX)
Dataset S9 (XLSX)