Seed genome hypomethylated regions are enriched in transcription factor genes

Min Chen[a], Jer-Young Lin[a], Jungim Hur[a], Julie M. Pelletier[b], Russell Baden[b,1], Matteo Pellegrini[a], John J. Harada[b], and Robert B. Goldberg[a,2]

[a]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA 90095; and [b]Department of Plant Biology, College of Biological Sciences, University of California, Davis, CA 95616

Contributed by Robert B. Goldberg, July 18, 2018 (sent for review June 26, 2018; reviewed by James J. Giovannoni and Brian Larkins)

The precise mechanisms that control gene activity during seed development remain largely unknown. Previously, we showed that several genes essential for seed development, including those encoding storage proteins, fatty acid biosynthesis enzymes, and transcriptional regulators (e.g., *ABI3*, *FUS3*) are located within hypomethylated regions of the soybean genome. These hypomethylated regions are similar to the DNA methylation valleys (DMVs), or canyons, found in mammalian cells. Here, we address the question of the extent to which DMVs are present within seed genomes and what role they might play in seed development. We scanned soybean and *Arabidopsis* seed genomes from postfertilization through dormancy and germination for regions that contain <5% or <0.4% bulk methylation in CG, CHG, and CHH contexts over all developmental stages. We found that DMVs represent extensive portions of seed genomes, range in size from 5–76 kb, are scattered throughout all chromosomes, and are hypomethylated throughout the plant life cycle. Significantly, DMVs are enriched greatly in transcription factor (TF) genes and other developmental genes that play critical roles in seed formation. Many DMV genes are regulated with respect to seed stage, region, and tissue, and contain H3K4me3, H3K27me3, or bivalent marks that fluctuate during development. Our results indicate that DMVs are a unique regulatory feature of both plant and animal genomes, and that a large number of seed genes are regulated in the absence of methylation changes during development, probably by the action of specific TFs and epigenetic events at the chromatin level.

seed development | DNA methylation valleys | transcription factor genes | soybean | *Arabidopsis*

Seeds are the agents for higher plant sexual reproduction and an essential source of food for human and animal consumption. They consist of three major regions—the seed coat, endosperm, and embryo—that have different genetic origins, unique functions, and distinct developmental pathways (1). The seed coat transfers nutrients from the maternal plant to the embryo during seed development and protects the seed during dormancy, a period of quiescence where growth and development have ceased temporarily. The endosperm also provides nourishment to the embryo, particularly during early embryogenesis, and in dicots, it degenerates and remains as a vestigial cell layer in the mature seed (2). The embryo, on the other hand, differentiates into axis and cotyledon regions, with the former giving rise to the mature plant after germination, while the latter is terminally differentiated and accumulates storage reserves that are used as an energy source for the germinating seedling (1, 3). Gene activity is highly regulated with respect to space and time within each seed region (4). However, the detailed mechanisms required for the differentiation of each seed region remain to be identified.

Selective DNA methylation of maternal and paternal alleles, or imprinting, plays a critical role in endosperm development (5–7). Imprinting, however, does not appear to play a major role in embryo formation (6, 7). The extent to which DNA methylation events regulate the activity of specific genes during embryogenesis is largely unknown. Recently, we (8) and others (9–11) showed that, on a global basis, CG- and CHG-context methylation does not change significantly during seed development. By contrast, CHH-context methylation increases primarily within transposons during the period leading up to dormancy (8–11), and it appears to be a fail-safe mechanism to ensure that transposons remain silent and do not inactivate genes essential for seed development and germination (8).

In the course of our seed methylome studies with both soybean and *Arabidopsis*, we identified several genes important for seed formation, including storage protein and fatty acid metabolism genes, which are located within hypomethylated genomic regions that remain static with respect to DNA methylation throughout development (8). These regions resemble the DNA methylation valleys (DMVs) (12), hypomethylated canyons [undermethylated regions (UMRs)] (13), and nonmethylated islands (NMIs) (14) found in many animal cell types that are enriched with transcription factor (TF) genes and coated with specific histone marks. Genes within these DMVs are not regulated by DNA methylation events, but by both transcriptional and epigenetic processes at the chromatin level (12–15).

Significance

We scanned soybean and *Arabidopsis* seed genomes for hypomethylated regions, or DNA methylation valleys (DMVs), present in mammalian cells. Seeds contain DMV regions that have <5% bulk DNA methylation or, in many cases, no detectable DNA methylation. Methylation levels of seed DMVs do not vary detectably during seed development and are present prior to fertilization. Seed DMVs are enriched in transcription factor (TF) genes and are decorated with histone marks that fluctuate developmentally, resembling their animal counterparts in significant ways. We conclude that many genes playing important roles in seed formation are regulated without detectable DNA methylation events and suggest that selective action of TFs, as well as chromatin epigenetic events, play important roles in making a seed.

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

Reviewers: J.J.G., US Department of Agriculture and Boyce Thompson Institute; and B.L., University of Arizona.

The authors declare no conflict of interest.

Data deposition: All ChIP-sequencing data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo (accession no. GSE114879).

[1]Present address: California Animal Health and Food Safety Laboratory, University of California, Davis, CA 95616.

[2]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.1811017115/-/DCSupplemental.

Published online August 13, 2018.

PLANT BIOLOGY

In this study, we scanned soybean and *Arabidopsis* seed genomes for DMVs. We found that a significant fraction of these seed genomes contains DMV regions that have <5% bulk DNA methylation or, in many cases, no detectable DNA methylation. Methylation levels of seed DMVs do not vary detectably during seed development with respect to time, region, and tissue, and appear to be present before fertilization. Seed DMVs are enriched in TF genes and other genes critical for seed development, and are also decorated with histone marks that fluctuate with developmental stage, resembling their animal counterparts in significant ways. We conclude that many genes playing important roles in seed formation are regulated in the absence of detectable DNA methylation events, and suggest that selective action of transcriptional activators and repressors, as well as chromatin epigenetic events, play important roles in making a seed, particularly embryo formation.

Results

DMVs Represent a Significant Portion of Soybean Seed Genomes. We scanned soybean seed methylomes from the globular stage through dormancy and germination for regions with <5% bulk methylation in all cytosine contexts (CG, CHG, and CHH) using a sliding 5-kb window with 1-kb smaller steps to search for hypomethylated regions, or DMVs (Figs. 1 and 2*A*). Seed DMVs were defined as genomic regions with <5% bulk methylation over all developmental stages investigated (*Materials and Methods*). This strategy identified 21,669 seed DNA regions, or 21% of the soybean genome (210 Mb), that were hypomethylated and did not vary significantly throughout seed development, or during early seed germination, with respect to methylation status (Fig. 2*B*, Table 1, and Dataset S1). Ninety-nine percent of seed DMVs identified during seed development were also shared with seedling and seedling cotyledon DMV regions, and, as such, we refer to all DMVs uncovered by our genome scans as seed DMVs (*Materials and Methods*). Detailed analysis showed that (*i*) the majority of seed DMVs had bulk methylation levels between 0% and 1% in all cytosine contexts (*SI Appendix*, Fig. S1*A*) and (*ii*) the average bulk methylation level per DMV was 0.14% (*SI Appendix*, Fig. S1*B*) or, on average, one methylated cytosine per 1 kb of DMV (*SI Appendix*, Fig. S1*B*), indicating that our strategy was robust enough to identify genomic regions that were significantly hypomethylated over all developmental stages. By contrast, the average bulk methylation level of the soybean genome was 11.5% or, on average, 43 methylated cytosines per 1 kb (*SI Appendix*, Fig. S1*B*). We validated this approach by scanning soybean seed methylomes at a <0.4% bulk methylation criterion, which was the lowest level of bulk methylation identifiable using bisulfite sequencing (BS-Seq) (8) (*Materials and Methods*), and uncovered 14,558 seed DMVs, or 112 Mb of the seed genomic regions, that effectively had no detectable cytosine methylation over all of the developmental stages examined (Table 1, *SI Appendix*, Fig. S1*B*, and Dataset S1).

Soybean seed DMVs averaged 10 kb in length and extended up to 76 kb, and 50% of the DMVs ranged in size from 6–11 kb at the <5% scanning criterion (Fig. 2*C* and Table 1). These values were somewhat smaller using the more stringent <0.4%

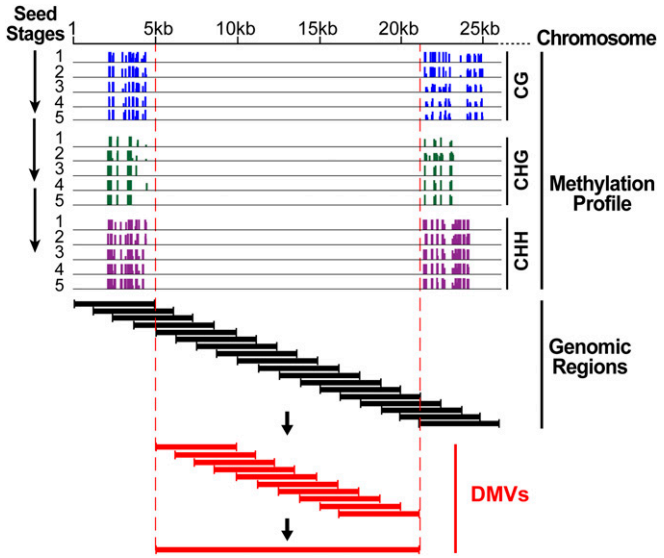

Fig. 1. Strategy to identify seed DMVs. Seed methylomes (8) were scanned across the genome at each stage of development (arrows) using a 5-kb sliding window with smaller 1-kb incremental steps (*Materials and Methods*) (dark bracketed lines). The bulk methylation levels in CG, CHH, and CHG contexts were calculated for each window at every developmental stage (8). Genomic regions with bulk methylation levels of <5% or <0.4% across all developmental stages were designated as DMVs, and overlapping DMVs were merged to define DMV regions (red line).

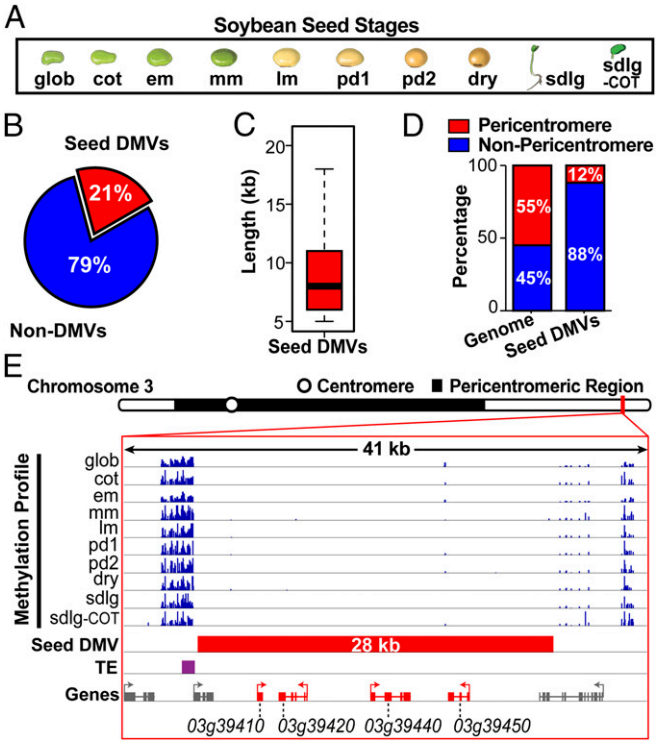

Fig. 2. Identification of soybean seed DMVs with <5% bulk methylation level. (*A*) Seed and postgermination methylomes used to identify DMVs (8). cot, cotyledon seed developmental stage; em, early-maturation seed developmental stage; glob, globular seed developmental stage; lm, late-maturation seed developmental stage; mm, midmaturation seed developmental stage; pd1, early predormancy seed developmental stage; pd2, late-predormancy seed developmental stage; sdlg, 6-d postgermination seedling; sdlg-COT, 6-d postgermination seedling cotyledon. Seed images are not drawn to scale. (*B*) Percentage of seed DMVs with <5% bulk methylation in the soybean genome (*Materials and Methods*). (*C*) Box plot of seed DMV lengths. The horizontal bar represents a median length of 8 kb. (*D*) Percentages of seed DMVs in chromosomal pericentromeric and nonpericentromeric regions. (*E*) Genome browser view of a 28-kb DMV located on chromosome 3. Genes in red color (Glyma03g39410, *EPOXIDE HYDROLASE*; Glyma03g39420, *60S RIBOSOMAL PROTEIN L18-3*; Glyma03g39440, *SERINE/THREONINE PROTEIN PHOSPHATASE 2A*; and Glyma03g39450, *DORMANCY/AUXIN ASSOCIATED PROTEIN*) are located within this DMV, including 1 kb of 5′ and 3′ flanking regions. Genes in gray color are located either partially or outside this DMV region. TE, transposable element.

Table 1. Summary of soybean and *Arabidopsis* DMV characteristics

	Soybean		*Arabidopsis*	
DMV characteristics	<5%*	<0.4%*	<5%*	<0.4%*
No. of DMVs	21,669	14,558	4,829	3,386
DMV genome size	210 Mb	112 Mb	49 Mb	27 Mb
Average DMV length†	10 ± 6 kb	8 ± 3 kb	10 ± 6 kb	8 ± 4 kb
Maximum DMV length	76 kb	41 kb	137 kb	39 kb
No. of DMV genes	14,328	6,377	8,710	3,736
No. of DMV TF genes	1,721	924	835	457

*Maximum DMV bulk methylation level in CG, CHG, and CHH contexts across all developmental stages.
†Mean length ± SD.

scanning criterion (Table 1). Soybean seed DMVs were located primarily on the arms (Fig. 2*D*) of all chromosomes (*SI Appendix*, Fig. S2), although over 2,000 DMVs were embedded within highly methylated pericentromeric regions (Fig. 2*D*). One example of a 28-kb seed DMV that was located on the distal arm of chromosome 3 is shown in Fig. 2*E*. This DMV region had four genes and an average bulk cytosine methylation level of 0.04%, and it was static with respect to methylation level during seed development, from the globular stage through dormancy and early germination (Fig. 2 *A* and *E*). Taken together, these results indicate that DMVs represent a significant portion of the soybean seed genome, and do not vary significantly with respect to methylation status during seed development and early germination.

Soybean DMVs Are Present in Different Seed Regions and Tissues. Soybean DMVs were identified using methylome data from whole seeds (8) (*Materials and Methods* and Fig. 2*A*). To determine whether the seed DMVs were present within different seed regions and tissues, we scanned soybean seed methylomes from (*i*) embryonic axis and cotyledon regions at three developmental stages and (*ii*) specific seed coat and cotyledon tissues at early maturation for DMVs at the <5% methylation criterion (Dataset S1), and compared these DMVs with those identified in seeds as a whole (Fig. 3 *A*–*C*). We generated these methylomes previously from hand-dissected seed regions (seed coat, axis, and cotyledons) and from specific seed tissues that were captured using laser capture microdissection (LCM) (8).

The vast majority of whole-seed DMVs overlapped with those identified within different seed regions and tissues (Fig. 3*D*). For example, 99% of the whole-seed DMVs were found within the seed coat and axis, while 98% were present within the cotyledons. Similarly, 99% of the whole-seed DMVs were represented in specific seed coat tissue layers (parenchyma and palisade) and within adaxial and abaxial cotyledon parenchyma tissues (Fig. 3*D*). These results indicate that soybean seed DMVs are not confined to one seed compartment but are present throughout the seed within diverse regions and tissues that have unique functions and developmental fates, and are a unique feature of seed and seedling genomes.

Soybean Seed DMVs Are Enriched in TF Genes. We searched seed DMVs for genes and scored identified genes as DMV genes if they had their gene bodies and 1 kb of 5′ and 3′ flanking regions within DMV regions (*Materials and Methods* and Fig. 2*E*). We uncovered 14,328 and 6,377 diverse genes within DMVs at the <5% and <0.4% scanning criteria, respectively (Table 1 and Dataset S2). The former represented 26% of all genes within the soybean genome (Fig. 4*A*).

Remarkably, 46% of soybean TF genes (n = 1,721) were located within seed DMV regions (Fig. 4*A* and Table 1), which represents a significant enrichment of TF genes within the soybean genome (χ^2 test, $P < 0.0001$). Gene ontology (GO) analysis indicated that the most enriched seed DMV gene GO functional group was regulation of transcription [false discovery rate (FDR) of 4.9×10^{-80}], containing 1,413 TF genes, as predicted from the large number of TF genes represented within seed DMVs (Fig. 4*A*, Table 1, and Dataset S3). A control group of 14,328 randomly selected soybean genes had no GO enrichment terms. DMV genes were distributed into many developmentally and physiologically relevant GO functional classes, such as developmental processes, response to hormones, response to stimulus, and transport, among others (Fig. 4*B*). Seed relevant developmental categories included pattern specification, cotyledon vascular tissue formation, radial pattern formation, organ boundary formation, and multicellular development (Fig. 4*B*). Each of these functional categories contained large numbers of specific TF genes that most likely guide and control these processes (Fig. 4*C*). Together, these data show that TF genes that play major roles in soybean seed formation are preferentially located within DMV regions.

Many Soybean DMV Genes Are Regulated During Seed Development. We analyzed soybean whole-seed RNA-Seq data, as well as RNA-Seq data from specific seed regions, subregions, and tissues captured using LCM, for DMV genes that were regulated during seed development (*Materials and Methods*). We found a large number of DMV genes that were up-regulated greater than fivefold in specific seed developmental stages, including many TF genes (Fig. 5 *A* and *B* and Dataset S2). For example, *GmWOX9a*, *GmSPCH*, and *GmTOM3* TF genes were expressed specifically at the globular, early-maturation, and late-maturation stages of seed development, respectively, and were present in DMV regions that did not vary significantly with respect to methylation during seed formation (Fig. 5*C*). All of the DMV stage-specific TF genes (Fig. 5*B*) were represented within the regulation of transcription and developmentally relevant functional GO groups (Fig. 4 *B* and *C*).

A large number of DMV genes, including those encoding TFs, were regulated with respect to specific seed regions, subregions, and tissues at different developmental stages as well (*SI Appendix*, Fig. S3 *A*–*C* and Dataset S2). Using a greater than fivefold up-regulation criterion (*Materials and Methods*), ~22% of all DMV

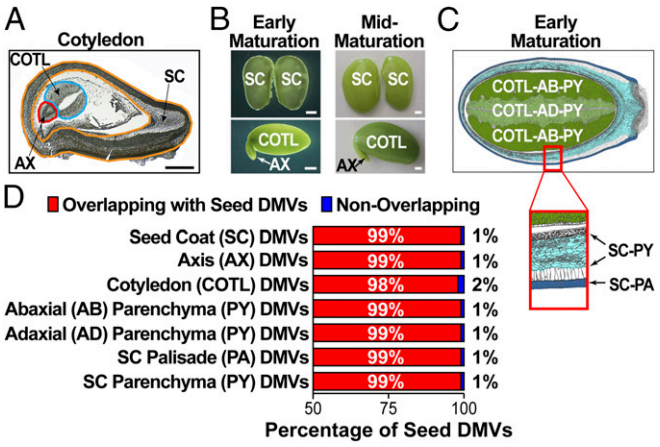

Fig. 3. Soybean seed DMVs are present in specific seed regions and tissues. (*A*) Paraffin section of a cotyledon stage seed. AX, axis; COTL, cotyledon; SC, seed coat. (*B*) Whole-mount photographs of embryos and SCs from early-maturation and midmaturation stage seeds. (*C*) Hand-drawn, colorized cross-section of an early-maturation stage seed. The expanded red box represents an enlarged SC region. COTL-AB-PY, cotyledon abaxial parenchyma tissue (dark green); COTL-AD-PY, cotyledon adaxial parenchyma tissue (light green); SC-PA, seed coat parenchyma tissue (light blue); SC-PY, seed coat palisade tissue (dark blue). (*D*) Percentage of seed DMVs that overlap with specific seed region and tissue DMVs. (Scale bars: *A*, 100 μm; *B*, 1 mm.)

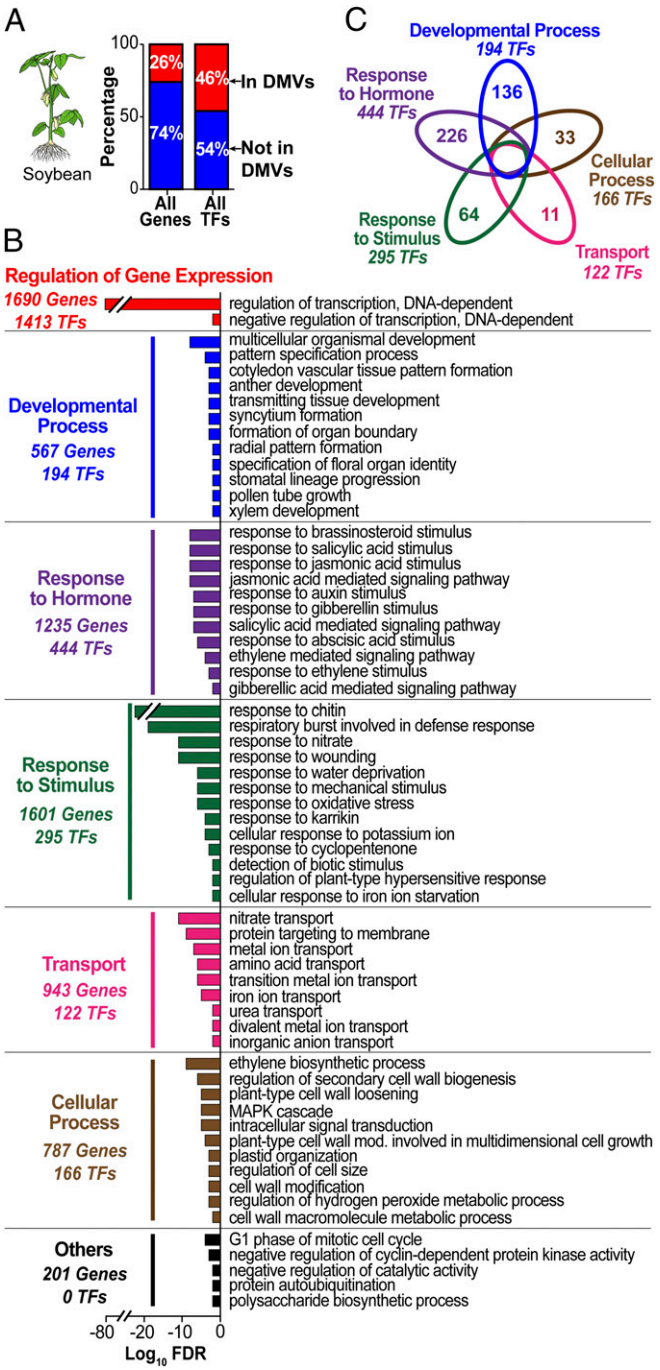

Fig. 4. Biological process GO terms that are enriched in soybean seed DMV genes. (*A*) Proportion of seed DMV genes and TF genes with <5% bulk methylation in the soybean genome. (*B*) Enriched GO terms with a FDR < 0.05, which are also listed in Dataset S3 (*Materials and Methods*). (*C*) Venn diagram showing the number of DMV TF genes that are unique to each of the five GO term biological function groups.

genes (n = 3,189) (*SI Appendix*, Fig. S3*A*) were expressed preferentially within specific seed parts, representing almost half of all soybean seed region-, subregion-, and tissue-specific genes (*SI Appendix*, Fig. S3*B*). For example, a *TRIHELIX DNA BINDING PROTEIN* gene located within a 10-kb seed DMV region that was shared by all seed developmental stages, regions, and tissues investigated (8) was expressed specifically within the early-maturation stage seed coat palisade tissue layer (*SI Appendix*, Fig. S3 *D* and *E*). This DMV had a maximum of three methylated cytosines over its entire 10-kb length, did not change its methylation status during development, and was flanked on either side by cliffs of highly methylated transposable elements (*SI Appendix*, Fig. S3*E*). Taken together, these data show that a large number of soybean genes that reside within seed DMV regions are regulated with respect to space and time and participate in important seed developmental processes (Fig. 4).

DMV Genes Are Enriched in H27Kme3 and Bivalent Histone Marks. We investigated soybean embryos at different developmental stages

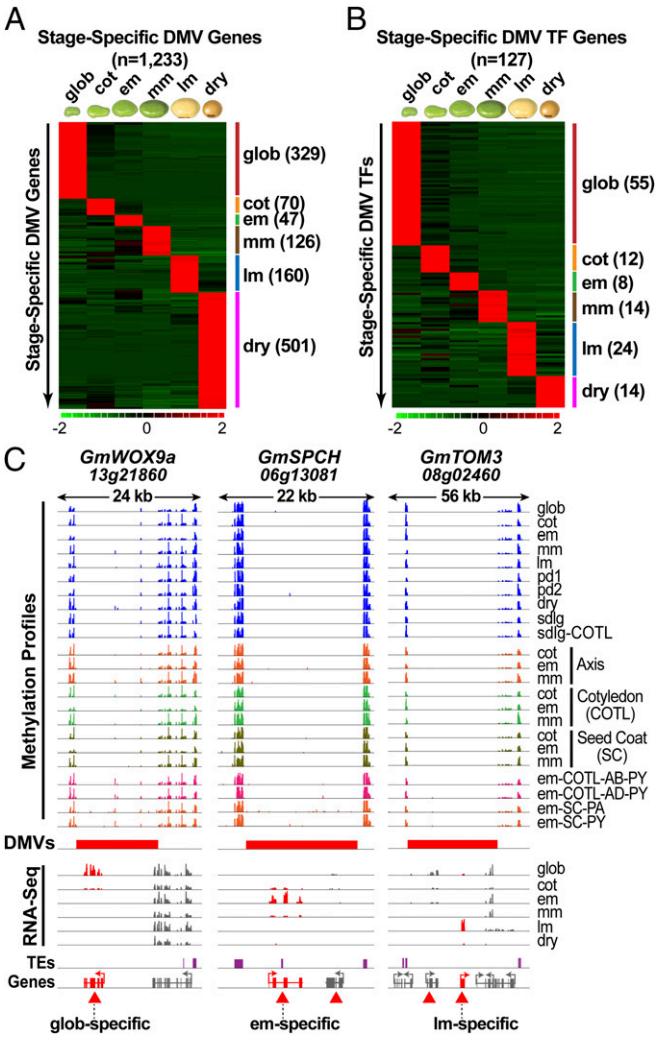

Fig. 5. Expression profiles of soybean DMV genes that are up-regulated greater than fivefold during seed development. Heat maps of up-regulated seed DMV genes (*A*) and TF genes (*B*) generated by edgeR analysis (33) of soybean RNA-Seq whole-seed datasets (GEO accession no. GSE29163) (8, 23) (*Materials and Methods*). The number of up-regulated DMV genes and TF genes in each stage is listed on the right side of the heat maps and in Dataset S2. Seed images are not drawn to scale. (*C*) Methylome and RNA-Seq genome browser views of three greater than fivefold up-regulated stage-specific TF genes [*GmWOX9a*, *WUSCHEL RELATED HOMEOBOX 9A*; *GmSPCH*, *SPEECHLESS*; and *GmTOM3*, *TARGET OF MONOPTEROS 3* (red gene models)]. Red triangles mark all genes within the DMVs, including those that are not specific to seed stage. Seed stage, region, and tissue abbreviations are as defined in the legends for Figs. 2 and 3. TEs, transposable elements. The scale ranging from −2 (green) to +2 (red) represents the relative number of SDs from the mean normalized RNA count for each DMV gene across all developmental stages.

to determine whether DMV genes were coated with specific histone marks (*Materials and Methods*). We used embryos instead of whole seeds to minimize the representation of different cell types in our chromatin preparations. Approximately 90% of all cells were derived from cotyledon parenchyma tissue at all stages of embryo development investigated for histone marks (16).

DMV genes were enriched significantly (hypergeometric test, $P < 0.001$) in H3K27me3 and bivalent (H3K27me3 and H3K4me3) marks at the cotyledon, early-maturation, mid-maturation, and late-maturation stages, as well as in seedlings, compared with all soybean genes (Fig. 6*A*), which is a distinctive feature of animal cell DMVs (12, 13). By contrast, there was no significant enrichment of H3K4me3 at any developmental stage, although a large fraction of expressed DMV genes was coated with this histone mark (Fig. 6*A*). DMV genes marked with H3K4me3 had the highest expression levels, followed by bivalent- and H3K27me3-marked genes, respectively (*SI Appendix*, Fig. S4). H3K27me3-marked DMV genes were repressed, or expressed at very low levels, consistent with the repressive nature of the H3K27me3 mark (13). Surprisingly, DMV genes coated with H3K27me3 and bivalent marks were enriched significantly for the transcriptional regulation functional GO term group, a property also similar to their animal DMV gene counterparts (12, 13). For example, at midmaturation, H3K27me3- and bivalent-marked DMV genes had FDRs of 2.93×10^{-24} and 3.92×10^{-108}, respectively, for this GO term group. Supporting these results, 62–74% of the 1,721 DMV TF genes (Table 1) contained bivalent or H3K27me3 marks depending upon the developmental stage, representing a significant enrichment compared with all TF genes (hypergeometric test, $P < 0.001$) (Fig. 6*B*). By contrast, DMV genes marked with H3K4me3 did not generate a transcriptional regulation enrichment group by GO analysis.

The chromatin marks for many seed DMV genes changed during development in parallel with their expression levels at specific stages (Fig. 6*A* and *SI Appendix*, Fig. S4). For example, the *P34 ALLERGEN PROTEIN* gene had an H3K27me3-repressive mark at the cotyledon stage, where it was repressed; an H3K4me3-active mark during early and midmaturation, where it was active at a high level; and an H3K27me3 mark in late maturation and postgermination seedlings, where it was silent (Fig. 6*C*). The *ABSCISIC INHIBITOR3-1* (*ABI3-1*) TF gene had an active H3K4me3 mark during seed development when it

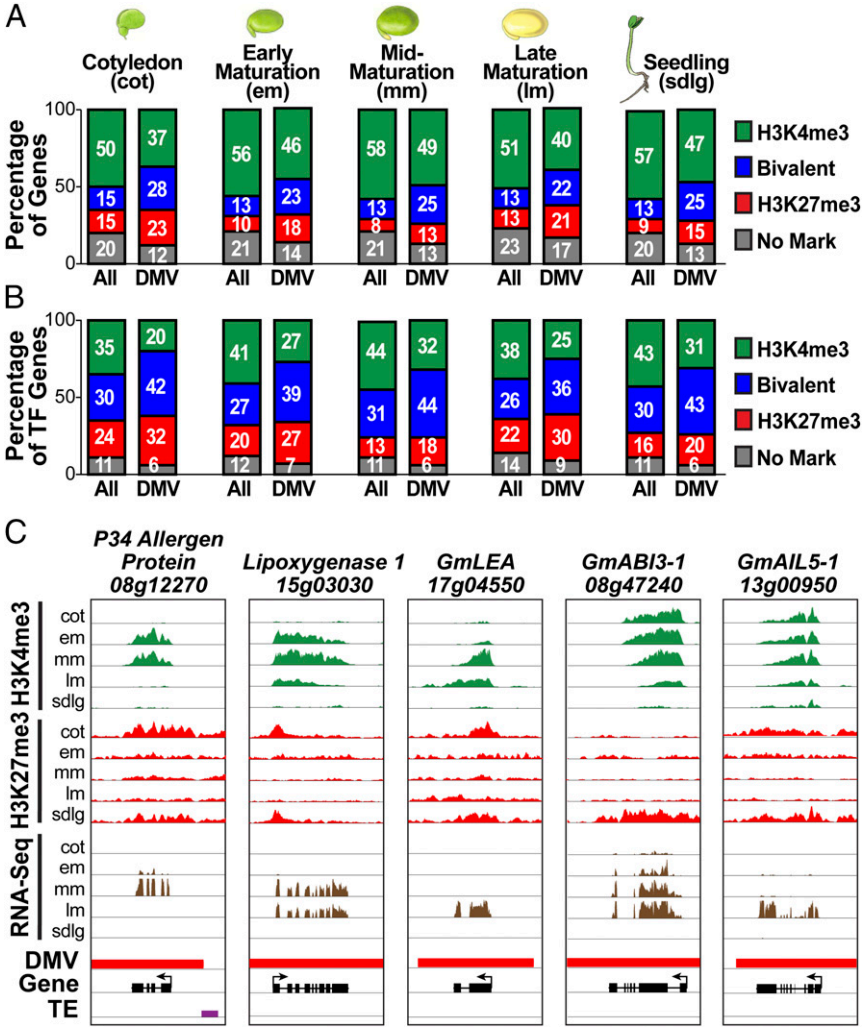

Fig. 6. DMV genes with histone marks across soybean seed development. (*A*) Percentage of DMV genes and all genes in the genome marked with H3K4me3, H3K27me3, H3K4me3, and H3K27me3 (bivalent mark) or with no mark at each developmental stage (GEO accession no. GSE114879). (*B*) Percentage of DMV TF genes and all TF genes in the genome marked with H3K4me3 and H3K27me3 (bivalent mark) or no mark at each developmental stage. (*C*) Genome browser views showing histone marks and RNA-Seq levels for specific DMV genes during seed development and germination. *GmABI3*, *ABSCISIC ACID INHIBITOR3*; *GmAIL5-1*, *AINTEGUMENTA LIKE5-1*; *GmLEA*, *LATE EMBRYO ABUNDANT*.

was expressed, but an H3K27me3 mark following germination when it was repressed (Fig. 6*C*). Finally, the *AINTEGUMENTA-LIKE5-1* (*AIL5-1*) TF gene had bivalent marks at cotyledon and early-maturation stages, where it was silent or expressed at low levels; an H3K4me3 mark during midmaturation and late-maturation stages, where it became active; and a bivalent mark within post-germination seedlings, where it was repressed (Fig. 6*C*). Together, these results indicate that (*i*) seed DMVs are enriched in TF genes that are preferentially coated with H3K27me3 and bivalent marks, (*ii*) genes within DMV regions undergo modifications in chromatin epigenetic state in the absence of DNA methylation changes during development, and (*iii*) seed DMVs have features strikingly similar to their animal cell counterparts.

***Arabidopsis* Seeds also Contain DMV Regions Enriched with TF Genes.** We scanned *Arabidopsis* methylomes from seeds at different developmental stages (globular stage through dry seed) and leaves from postgermination plants to determine whether DMVs were a conserved feature of plant genomes (Fig. 7*A*). We used the same strategy to identify *Arabidopsis* seed genomic regions with <5% and <0.4% average bulk methylation levels as we did to uncover soybean seed DMVs (*Materials and Methods* and Fig. 1). Ninety-nine percent of the DMVs identified during seed development were also shared with leaf DMVs, and, as such, we refer to all DMVs as seed DMVs (*Materials and Methods*).

Approximately 41% of the *Arabidopsis* genome, or 4,829 regions, consisted of seed DMVs at the <5% scanning criterion (Fig. 7*B*, Table 1, and Dataset S1). These regions had a bulk cytosine methylation level of 0.24%, on average, compared with 5.8% for the *Arabidopsis* genome as a whole (*SI Appendix*, Fig. S1 *C* and *D*). *Arabidopsis* DMV regions averaged 10 kb in length and were localized primarily on the arms of all chromosomes, similar to what was observed in soybeans (Fig. 7 *C* and *D* and Table 1). A smaller number of seed DMVs (n = 3,386) were uncovered using the <0.4% scanning criterion that identified genomic regions with essentially no methylation (Table 1).

Approximately 32% of all *Arabidopsis* genes (n = 8,710), including 48% of those encoding TFs (n = 835), were localized within seed DMV regions at the <5% scanning criterion (Fig. 7*E*, Table 1, and Dataset S2). *Arabidopsis* TF genes were enriched significantly within DMVs (χ^2 test, $P < 0.0001$), and about one-half overlapped with those present within soybean DMV regions. GO enrichment analysis of all *Arabidopsis* seed DMV genes showed that the most significant GO functional groups were sequence-specific DNA-binding TF activity, nucleic acid TF-binding activity, and regulation of transcription (Fig. 7*F* and Dataset S4), similar to what was observed for soybean DMVs (Fig. 4). Other developmentally relevant GO functional groups, such as response to hormones, also showed significant enrichment (*SI Appendix*, Fig. S5 and Dataset S4). DMV genes shared between *Arabidopsis* and soybeans showed GO enrichment for many biological processes related to development, as predicted from the DMV GO analysis of each plant species separately (*SI Appendix*, Fig. S6). Finally, large numbers of *Arabidopsis* genes that reside within DMVs, including those encoding TFs, were regulated with respect to time and space during seed development (*SI Appendix*, Figs. S7 and S8 and Dataset S2). Together, these results show that *Arabidopsis* seed DMV regions have features identical to those in soybean seeds, and that DMVs enriched for TF genes and other genes that carry out important seed developmental functions are a conserved feature of seed genomes.

Most Seed DMV Regions Are Present Before Fertilization. In both soybeans and *Arabidopsis*, 99% of DMVs uncovered during seed development were conserved within the genomes of seedlings (soybeans) and leaves (*Arabidopsis*), indicating that the seed DMV methylation status did not change significantly following germination (*Materials and Methods*). We scanned *Arabidopsis* sperm, vegetative, and central cell methylomes generated by others for DMVs (17, 18) (Fig. 1), and compared DMVs present in these gametophytic cells with those uncovered from our *Arabidopsis* postmature green and dry seed methylomes (8) to determine whether seed DMVs were present before fertilization (Fig. 7*G*). We used seed methylomes from the Col-0 ecotype (Fig. 7*G*) instead of Ws-0 (Fig. 7 *A*–*F*), because the gametophytic cell methylomes were from Col-0, and we found that 10% of Ws-0 seed DMVs differed from those in Col-0 at the same developmental stages due to cytosine polymorphisms (8, 19, 20). Thus, it was important to carry out within-ecotype DMV comparisons.

The vast majority (86–95%) of *Arabidopsis* seed DMV regions were also present within female gametophyte (central cell) and male gametophyte (sperm and vegetative cells) genomes before fertilization (Fig. 7*G*). The small number of seed DMV regions that were not scored as DMVs in sperm and central cells had average bulk methylation levels >5% and, therefore, higher methylation levels before fertilization and seed development. These results indicate that most seed DMV regions are highly conserved and do not change significantly with respect to methylation status during major periods of the plant life cycle. They are present within gametophytic cells before fertilization, developing seeds after fertilization, growing sporophytic seedlings following seed germination, and leaves of the young sporophytic plant.

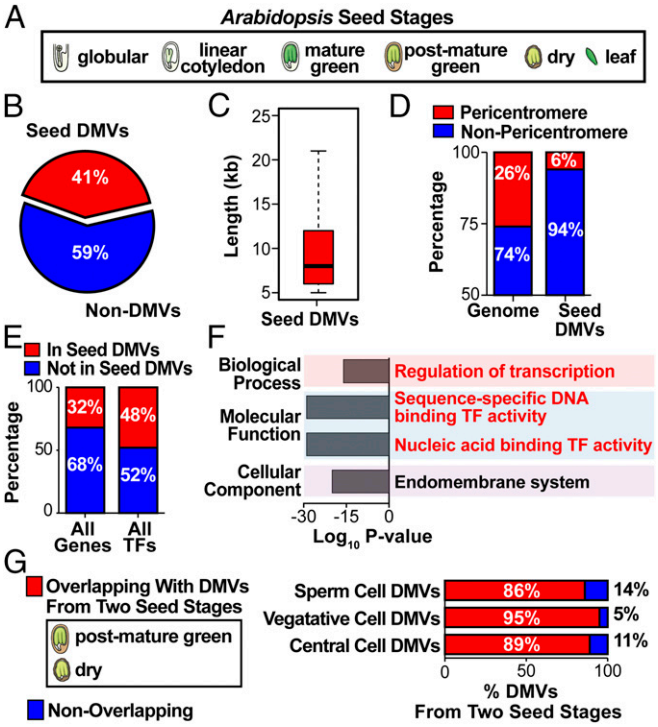

Fig. 7. Identification of *Arabidopsis* seed DMVs with <5% bulk methylation level. (*A*) *Arabidopsis* methylomes used to identify DMVs (8). Images are not drawn to scale. (*B*) Proportion of seed DMVs with <5% bulk methylation level in the *Arabidopsis* genome. (*C*) Box plot of seed DMV lengths. The horizontal bar represents a median length of 8 kb. (*D*) Proportion of seed DMVs in chromosomal pericentromeric and arm regions. (*E*) Proportion of seed DMV genes and DMV TF genes with <5% bulk methylation in the *Arabidopsis* genome. (*F*) Bar plots of the most significantly enriched GO terms of *Arabidopsis* seed DMV genes, which are also listed in Dataset S4. GO terms related to TF activity are highlighted in red. (*G*) Percentages of seed DMVs that overlap with sperm cell, vegetative cell, and central cell DMVs.

Many Genes Known to Play Important Roles in Seed Development Are Present Within Soybean and *Arabidopsis* DMVs. GO analysis indicated that genes contained within both soybean and *Arabidopsis* DMVs were significantly enriched for those that play major regulatory roles during seed formation (Figs. 4 and 7*F* and *SI Appendix*, Figs. S5 and S6). We searched soybean and *Arabidopsis* DMVs for genes that were known to play critical roles in seed differentiation and development (Dataset S5). Many genes essential for embryo formation were localized within both soybean and *Arabidopsis* DMVs, including several *WUS HOMEOBOX-CONTAINING* (*WOX*) genes (e.g., *WOX1*, *WOX3*, *WOX8/9*), *BABY BOOM*, *SCARECROW*, *SHATTERPROOF1*, *PLETHORA*, *TARGET OF MONOPTEROS5*, and *CUP-SHAPED COTYLEDON* family genes (e.g., *CUC2*, *CUC3*). Other genes playing major roles in seed formation, such as *CLAVATA3*, *PIN1*, *YUCCA4*, *BODENLOS*, and *HANABA TARANU*, were also present within both soybean and *Arabidopsis* DMVs. In addition, several gene classes that play critical roles in seed physiological processes, such as those encoding storage proteins (e.g., *GmGlycinin1*, *AtCruciferin1*) and hormone biosynthesis enzymes [e.g., gibberellic acid oxidase (*GmGA20Ox2*, *GmGA3Ox1*, *AtGm20Ox2*, and *AtGA3Ox1*)], were also localized with both soybean and *Arabidopsis* DMVs. Together, these results suggest that many genes playing essential roles in seed formation are conserved within the DMVs of divergent plant species.

Discussion

We have shown that there are large portions of soybean and *Arabidopsis* seed genomes that are hypomethylated and enriched with TF genes. Seed DMVs resemble in striking ways the DMVs that are present in the cells of many vertebrate animals, including humans (12–14, 21). DMVs in both kingdoms (*i*) are scattered across the genome in thousands of long hypomethylated regions; (*ii*) do not vary significantly with respect to their methylation status across diverse developmental stages, tissues, and cell types; (*iii*) are enriched significantly with developmentally important genes, including large numbers of TF genes; (*iv*) undergo epigenetic changes at the chromatin level that can accompany changes in gene activity; and (*v*) are present within diverse species. Thus, DMVs appear to be a unique feature of eukaryotic genomes that play important roles in cell differentiation and development in the absence of methylation changes at the DNA level. It has been proposed that DMVs arose in animal cells as a result of strong negative selection, or purification, for transposable element insertions that would disrupt critical regulatory genes and have a deleterious effect on development (22). Whether this is the case for the seed DMVs uncovered here remains to be determined.

A large number of genes within both soybean and *Arabidopsis* DMVs are regulated during seed development, and are expressed within specific seed stages, regions, subregions, or tissues. These genes include those encoding TFs, as well as others, such as storage protein genes, that play important roles in seed formation and germination. The remarkable feature of these genes is that they retain their hypomethylated status within DMVs regardless of their state of expression. How then are these genes turned on and off in the absence of methylation changes that have been shown to play, for example, a role in differential activity of maternal and paternal alleles, which is essential for seed endosperm development (7)?

One possibility is that many DMV genes, especially those encoding storage proteins, are activated and repressed by the action of seed-specific regulators such as *LEAFY COTYLEDON1* (*LEC1*), *ABI3*, and *FUSCA3* (*FUS3*) (23), among others. Epigenetic changes at the chromatin level within seeds may also partner with the action of TFs to facilitate development-specific changes in gene activity. Soybean seed DMV genes, like their animal counterparts (12, 13, 15), are enriched for H3K27me3 and bivalent marks, and the histone marks coated on DMV genes, including H3K4me3, can change in parallel with their gene activity levels. DMV genes that are repressed, or expressed at low levels, are coated with repressive H3K27me3 marks, whereas those that are active are marked with either H3K4me3 or bivalent marks. We assume that the bivalent marks on seed DMV genes resemble those that coat animal DMV genes, rather than being caused by a mixture of tissues in which the genes are active and repressed. First, seed DMV genes coated with bivalent marks are enriched with TF genes, similar to their animal counterparts (24). Second, seed DMV genes that contain bivalent marks are expressed at significantly lower levels than their counterparts marked with H3K4me3, which is a feature of animal genes with bivalent marks (24). Finally, seed DMV genes marked with both H3K27me3 and H3K4me3 were identified from soybean embryos that consist primarily of parenchyma storage tissue cells that reside within cotyledon regions (Fig. 3). Thus, we favor a model in which the actions of TFs, coupled with chromatin-level epigenetic changes, regulate genes within seed DMV regions during development in the absence of DNA methylation events. In animals, TF genes have been shown to regulate other genes within the same DMV and are organized into self-contained chromatin domains (15). Whether that is also true of seed DMVs remains to be determined.

What about seed genes that are not present within DMVs? These genes reside within genomic regions with >5% average bulk methylation at CG, CHG, and CHH sites. These regions contain highly methylated transposable elements, genes with body methylation, or both (8). Many soybean and *Arabidopsis* genes outside of DMV regions are also regulated with respect to space and time during seed development (8). We showed previously that many of these genes are activated and shut down in the absence of significant methylation changes similar to genes within DMV regions (8). Whether these genes undergo major epigenetic chromatin changes during seed development remains to be determined.

A major challenge for the future will be to determine the precise mechanisms by which genes residing within both DMV and non-DMV regions are regulated during seed development. This will require identifying the *cis*-control elements that program seed gene activity, cognate TFs that interact with these elements, proteins that drive epigenetic changes at the chromatin level, and how genes expressed during seed development are organized into regulatory networks that are required to form a seed.

Materials and Methods

General Strategy for Identifying DMVs. Soybean (*Glycine max* cv Williams 82) and *Arabidopsis* (Ws-0 and Col-0) seed and postgermination methylome data were taken from our previously published BS-Seq experiments (8). Specific details regarding the BS-Seq libraries used in the experiments reported here, and how they were constructed and characterized, are contained in materials and methods and SI materials and methods of ref. 8.

DMVs were identified using the strategy illustrated in Fig. 1 (12). The methylome at each developmental stage was scanned using a sliding window of 5 kb with smaller 1-kb incremental steps, and regions without cytosine bases were discarded. The bulk methylation levels in CG, CHG, and CHH contexts were calculated for all remaining windows (8, 25). DMVs were defined as genomic regions with either <5% or <0.4% bulk methylation levels in all three cytosine contexts over all developmental stages studied (Table 1). A 0.4% bulk methylation level was used to identify DMVs with no detectable methylation, as this was the lowest level of unmethylated cytosines that were not converted by BS to thymine in our previous experiments (8). Overlapping DMVs were merged and reported as contiguous DMV regions in Dataset S1. Only genes that contained their entire bodies and 1 kb of 5′ and 3′ flanking regions were designated as genes contained within DMV regions (Dataset S2).

Identification and Characterization of Soybean Seed DMVs. Soybean DMVs were identified in the methylomes of globular, cotyledon, early-maturation, midmaturation, late-maturation, early-predormancy, late-predormancy, and dry seeds, as well as in 6-d postgermination seedlings and seedling cotyledons (Fig. 2). Ninety-nine percent of the DMVs identified during seed development were also present as DMVs in postgermination seedlings and seedling

cotyledons. Thus, we refer collectively to these DMVs as soybean seed DMVs (Fig. 2 and Dataset S1).

GO term enrichment analysis of seed DMV genes. The (*i*) GOSeq R Bioconductor package (26), (*ii*) SoyBase GO annotations (https://soybase.org/genomeannotation/index.php), (*iii*) hypergeometric statistical test method, and (*iv*) Benjamini–Hochberg multiple testing correction (FDR < 0.05) were used to identify GO terms enriched in soybean DMV genes (23) (Fig. 4 and Dataset S3).

Expression analysis of seed DMV genes. The dChip software program (27) was used to generate heat maps of soybean up-regulated DMV mRNA levels in whole seeds (Fig. 5) and in specific seed regions, subregions, and tissues (*SI Appendix*, Fig. S3) throughout development. Whole-seed RNA-Seq data were taken from our previously published experiments (8) (GEO accession no. GSE29163). RNA-Seq data for specific seed regions, subregions, and tissues were taken from the Harada–Goldberg LCM datasets (GEO accession no. GSE116036) (23). EdgeR was used to identify DMV mRNAs that were up-regulated greater than fivefold (FDR < 0.001, Benjamini–Hochberg multiple testing correction) in specific seed stages, regions, subregions, and tissues (23).

Identification of histone marks on DMV genes. Embryos at different developmental stages were used to characterize DMV genes for H3K4me3 and H3K27me3 histone marks (Fig. 6). ChIP assays were carried out using anti-H3K4me3, anti-H3K27me3, and anti-H3 antibodies according to our previously published protocol (23). ChIP-Seq library construction, DNA sequencing analysis, and peak calls were carried out as described previously (23). Histone H3 ChIP-Seq results were used as a control to filter out noise during peak calling. MACS2 and SICER programs were used to call H3K4me3 and H3K27me3 peaks, respectively (28–30). DMV genes were considered marked with H3K4me3, H3K27me3, or both H3K4me3 and H3K27me3 (bivalent marks) if peaks had a *P* value <0.05 and were associated with the gene body and/or 1 kb of the upstream region (23). All ChIP-Seq data reported in this paper were deposited in the GEO (accession no. GSE114879).

Identification and Characterization of *Arabidopsis* Seed DMVs. *Arabidopsis* (Ws-0) DMVs were identified in the methylomes of globular, linear cotyledon, mature green, postmature green, and dry seeds, as well as in leaves from 4-wk-old plants (8) (Fig. 7). Ninety-nine percent of the DMVs identified during seed development were also present as DMVs in leaves. Thus, we refer collectively to these DMVs as *Arabidopsis* seed DMVs (Fig. 7 and Dataset S1). *Arabidopsis* (Col-0) methylomes from postmature green and dry seeds (8) and those from sperm, vegetative, and central cells (17, 18) were used to compare DMVs from seeds and gametophytic cells.

GO term enrichment analysis. The VirtualPlant 1.3 program (31) was used for *Arabidopsis* gene GO enrichment analysis with a cutoff FDR value <0.05 (Benjamini–Hochberg multiple testing correction) (Fig. 7 and *SI Appendix*, Figs. S5 and S6).

Gene expression analysis. The dChip software program (27) was used to generate heat maps of *Arabidopsis* DMV mRNA levels in whole seeds (*SI Appendix*, Fig. S7) and in specific seed regions and subregions (*SI Appendix*, Fig. S8) during development. *Arabidopsis* seed transcriptome data were taken from our previously published GeneChip experiments [whole seeds (GEO accession no. GSE680) (32); regions and subregions (GEO accession no. GSE12404) (4)].

ACKNOWLEDGMENTS. 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.).

1. Goldberg RB, de Paiva G, Yadegari R (1994) Plant embryogenesis: Zygote to seed. *Science* 266:605–614.
2. Hands P, Rabiger DS, Koltunow A (2016) Mechanisms of endosperm initiation. *Plant Reprod* 29:215–225.
3. Weijers D (2014) Genetic control of identity and growth in the early Arabidopsis embryo. *Biochem Soc Trans* 42:346–351.
4. Belmonte MF, et al. (2013) Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. *Proc Natl Acad Sci USA* 110: E435–E444.
5. Bauer MJ, Fischer RL (2011) Genome demethylation and imprinting in the endosperm. *Curr Opin Plant Biol* 14:162–167.
6. Gehring M, Satyaki PR (2017) Endosperm and imprinting, inextricably linked. *Plant Physiol* 173:143–154.
7. Satyaki PR, Gehring M (2017) DNA methylation and imprinting in plants: Machinery and mechanisms. *Crit Rev Biochem Mol Biol* 52:163–175.
8. Lin JY, et al. (2017) Similarity between soybean and *Arabidopsis* seed methylomes and loss of non-CG methylation does not affect seed development. *Proc Natl Acad Sci USA* 114:E9730–E9739.
9. An YC, et al. (2017) Dynamic changes of genome-wide DNA methylation during soybean seed development. *Sci Rep* 7:12263, and erratum (2018) 8:7882.
10. Bouyer D, et al. (2017) DNA methylation dynamics during early plant life. *Genome Biol* 18:179.
11. Kawakatsu T, Nery JR, Castanon R, Ecker JR (2017) Dynamic DNA methylation reconfiguration during seed development and germination. *Genome Biol* 18:171.
12. Xie W, et al. (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 153:1134–1148.
13. Jeong M, et al. (2014) Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat Genet* 46:17–23.
14. Long HK, et al. (2013) Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife* 2:e00348.
15. Li Y, et al. (2018) Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biol* 19:18.
16. 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.
17. Hsieh PH, et al. (2016) Arabidopsis male sexual lineage exhibits more robust maintenance of CG methylation than somatic tissues. *Proc Natl Acad Sci USA* 113: 15132–15137.
18. 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.
19. Kawakatsu T, et al.; 1001 Genomes Consortium (2016) Epigenomic diversity in a global collection of Arabidopsis thaliana accessions. *Cell* 166:492–505.
20. Vaughn MW, et al. (2007) Epigenetic natural variation in *Arabidopsis thaliana*. *PLoS Biol* 5:e174.
21. Stadler MB, et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480:490–495, and erratum (2012) 484:550.
22. Simons C, Makunin IV, Pheasant M, Mattick JS (2007) Maintenance of transposon-free regions throughout vertebrate evolution. *BMC Genomics* 8:470.
23. Pelletier JM, et al. (2017) LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development. *Proc Natl Acad Sci USA* 114:E6710–E6719.
24. Bernstein BE, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125:315–326.
25. Hsieh TF, et al. (2009) Genome-wide demethylation of Arabidopsis endosperm. *Science* 324:1451–1454.
26. Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol* 11:R14.
27. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol* 2:RESEARCH0032.
28. Zhang Y, et al. (2008) Model-based analysis of ChIP-seq (MACS). *Genome Biol* 9:R137.
29. Zang C, et al. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-seq data. *Bioinformatics* 25:1952–1958.
30. Bailey T, et al. (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLOS Comput Biol* 9:e1003326.
31. Katari MS, et al. (2010) VirtualPlant: A software platform to support systems biology research. *Plant Physiol* 152:500–515.
32. 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.
33. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139–140.