Supporting Online Material for

Genome-Wide Demethylation of *Arabidopsis* Endosperm

Tzung-Fu Hsieh, Christian A. Ibarra, Pedro Silva, Assaf Zemach,
Leor Eshed-Williams, Robert L. Fischer,* Daniel Zilberman*

*To whom correspondence should be addressed. E-mail: rfischer@berkeley.edu (R.L.F.),
daniel.zilberman@nature.berkeley.edu (D.Z.)

Published 12 June 2009, *Science* **324**, 1451 (2009)
DOI: 10.1126/science.1172417

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/324/5933/1451/DC1)

Supporting Online Material

Materials and Methods

Seed dissection and nucleic acid purification. Seeds at the mid-torpedo stage to early-maturation stage (7-9 days after pollination, ecotype Col-0) were dissected in 0.3 M sorbitol, 5 mM MES (pH 5.7) on a slide under a dissecting microscope. At this stage of development, bilaterally active genes are expressed from both maternal and paternal genomes in embryo and endosperm (*S1*). Embryos were washed to remove contaminating endosperm. Because the Arabidopsis seed coat has considerable tensile strength, it was possible to separate endosperm from the seed coats and extract DNA from pure endosperm, as previously described (*S2*, *3*). We also extracted DNA from aerial tissues of 4-week old Col-0 plants.

Bisulfite sequencing. For our experiments we adapted the protocol published recently by Joseph Ecker and colleagues (*S4*). First, we synthesized custom Illumina adapters in which cytosines were replaced by 5-methylcytosines, so that the adapters would survive bisulfite conversion. We chose to synthesize paired ends (PE) adapters, which allow each molecule to be sequenced from both ends, thus facilitating subsequent alignment to the genomic scaffold. The adapter oligos (5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'; 3'-GAGCCGTAAGGACGACTTGGCGAGAAGGCTAGp-5'; oligonucleotide sequences © 2006 and 2008 Illumina, inc.) were annealed as described by Cronn et al. (*S5*), generating a Y-shaped molecule (fig. S1).

We isolated 0.5-1 micrograms of genomic DNA from endosperm dissected from wild type and *dme* seeds, as well as control wild type embryos and aerial tissues. We did not use *dme* embryos because they abort at a very early stage. DNA was sheared by sonication to fragments

of 100-500 bp, and the adapters were ligated following the Illumina protocol. The library was then treated twice with sodium bisulfite (which converts unmethylated Cs to Us) using the Qiagen EpiTect kit, and amplified by 18 cycles of PCR using PfuTurboC_x DNA polymerase (Stratagene), a proofreading enzyme that tolerates uracil in the template strand. PCR amplification results in a library with distinct adapters at each end, so that the 'forward' Illumina sequencing primer always produces sequence from the 'original' genomic DNA-derived strand (where Cs represent methylated Cs), and the 'reverse' Illumina sequencing primer always produces sequence from the PCR-generated strand (where Gs represent methylated Cs on the opposite strand).

Bands around 300 bp were gel-purified and cloned for validation. Traditional sequencing confirmed that the libraries were properly constructed: the clones were fully converted and had correct adapter sequences on each end. The libraries were sequenced at the UC Berkeley Genomic Sequencing Laboratory, generating 45 bp reads from each end (52.3 million for embryo, 46.7 million for WT endosperm, 40.7 million for *dme* endosperm, and 28.7 million for adult aerial tissues).

Sequence analysis. To align the reads to the TAIR8 *Arabidopsis* genomic scaffold, we adapted the strategy described by Meissner et al. (*S6*). We used custom Perl scripts to convert all the Cs in the 'forward' reads (and in the scaffold) to Ts, and all the Gs in the 'reverse' reads and scaffold to As, and initially aligned the converted reads to the converted scaffold using SeqMap as individual reads (*S7*), allowing up to two mismatches per read, with 75% of the mapped reads aligning without mismatches. We subsequently used a Perl script to insure that paired reads mapped to opposite strands within 300 bp of each other, uniquely mapping a total of 207,060,231 reads (61.5% of all reads), 21.7% percent of which were mapped by using pair information to

resolve non-unique matches. We uniquely aligned 2,460,620,115 bases for embryo, 2,187,031,905 bases for wild-type endosperm, 1,959,340,860 bases for *dme* endosperm, and 1,547,474,265 bases for aerial tissues to the nuclear genome, and 368,429,850 bases for embryo (2,385-fold coverage), 161,853,975 bases for wild-type endosperm (1,048-fold coverage), 215,443,035 bases for *dme* endosperm (1,395-fold coverage), and 135,402,300 bases for aerial tissues (877-fold coverage) to the chloroplast genome (table S1). We then used Perl scripts to recover the original sequence information and, for each C (on either strand), count the number of times it was sequenced as a C or a T. For each sequence context (CG, CHG, CHH) we calculated bulk fractional methylation (#C/(#C+#T)), as well as fractional methylation within a 50 bp sliding window. Chloroplast bulk fractional methylation was 0.10% for embryo, 0.10% for wild-type endosperm, 0.15% for *dme* endosperm, and 0.35% for aerial tissues (table S1), which we used as an aggregate error measure.

Methylation analysis. Our results closely matched both published conventional bisulfite sequencing (table S2) and high-throughput bisulfite sequencing data (fig. S2) (*S2, S8-14*). We used the high-throughput dataset published by Cokus et al. (*S8*) for comparison because it was derived from very similar tissues (aerial tissues of 5-week-old Col-0 plants) to those used for our aerial dataset (4-week-old Col-0 plants), but utilized different methods for library construction, bisulfite conversion and read alignment. We calculated methylation differences for 50 bp windows containing at least 20 informative sequenced cytosines in each dataset. Mean difference for windows with fractional CG methylation of at least 0.7 in either our aerial dataset or the Cokus et al. dataset was 0.0006, over 300 times smaller than the difference between our embryo and WT endosperm datasets. Mean difference for windows with fractional CHG methylation of at least 0.5 in one of the tissues was 0.029 and mean differences for windows with fractional

CHH methylation of at least 0.1 in one of the tissues was -0.006. The same comparison parameters were used to generate the density plots in Fig. 3, fig. S4 and fig. S6.

To identify sequences with significant methylation differences between two tissues, adjacent 50 bp windows containing any methylation in the relevant context in at least one of the tissues being compared were concatenated. Concatenated windows with at least 20 informative sequenced cytosines and a methylation difference of at least 0.1 for CG and CHG contexts and at least 0.05 for CHH context at a $p<0.0001$ (Fisher's exact test) are reported in table S3.

siRNA analysis. We used published siRNA data from *Arabidopsis* flower buds (*S4*) to count the number of siRNAs within 50 bp windows covering the genome. Fig. 2A depicts all windows, or those matching sequences hypermethylated in either the CHG or CHH context in the endosperm.

Expression analysis. We used publically available microarray expression data (GEO accession number GSE12403) for cellularized endosperm and embryo at the linear cotyledon stage (http://estdb.biology.ucla.edu/seed/arabidopsis#microarray). Endosperm and embryo replicates were averaged, normalized, and relative expression scores ($\log_2$ (endosperm/embryo)) were computed for each gene represented on the array. Expression scores for genes with reduced CG methylation in wild-type endosperm within 1 kb of either the 5' or 3' end (table S4) were used in the analysis shown in Fig. 2B.

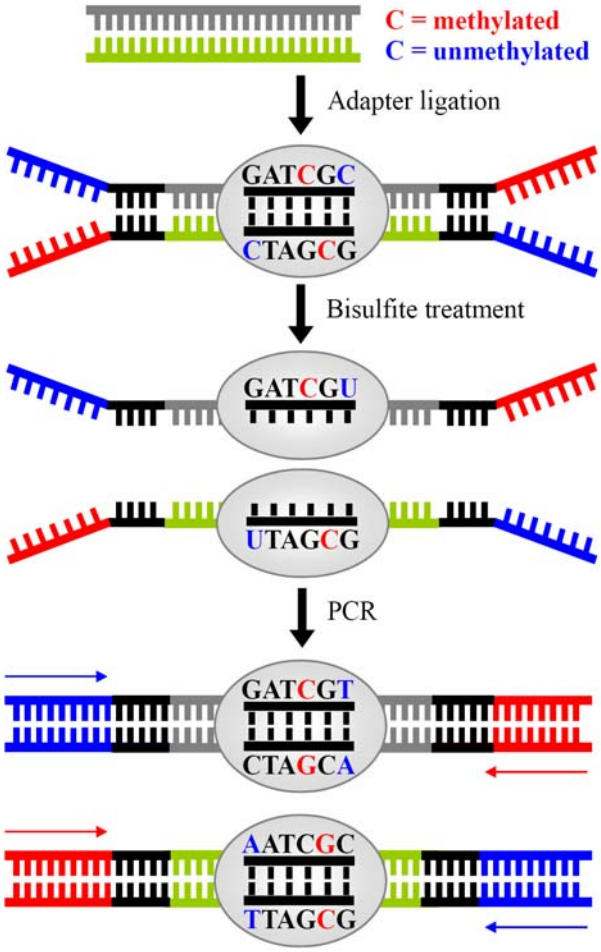

fig. S1. A schematic of library construction for Illumina sequencing of bisulfite-converted DNA.

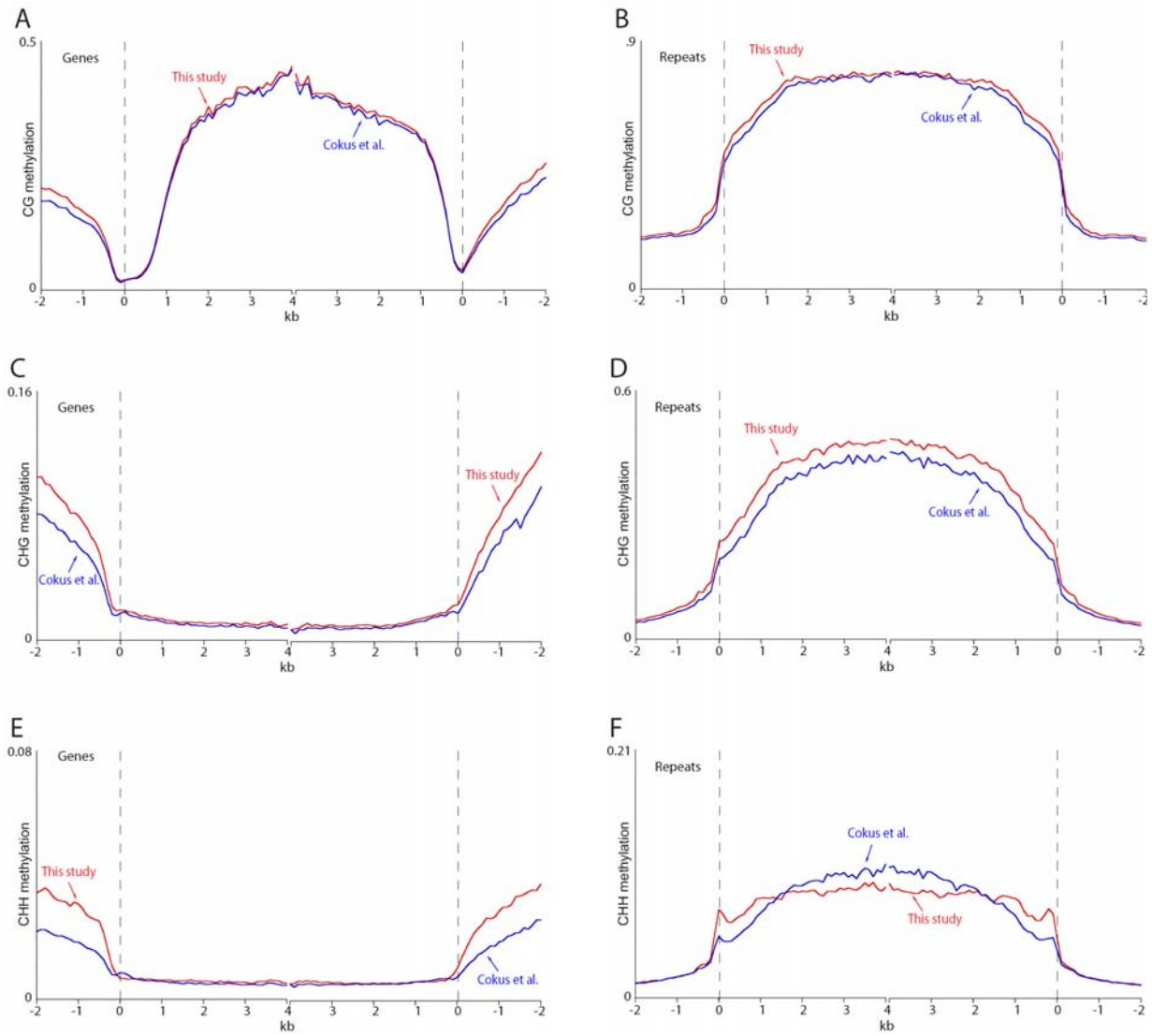

fig. S2. (**A-F**) TAIR8-annoted genes (**A, C, E**) or transposons (**B, D, F**) were aligned at the 5' end (left panel) or the 3' end (right panel) and average methylation levels for each 100-bp interval are plotted from 2 kb away from the gene (negative numbers) to 4 kb into the gene (positive numbers). Methylation of WT aerial tissues as determined in this study is represented by the red trace, and as determined by Cokus et al. (*S8*) by the blue trace. The dashed line at zero represents the point of alignment. CG methylation is shown in (**A-B**), CHG in (**C-D**), CHH in (**E-F**).

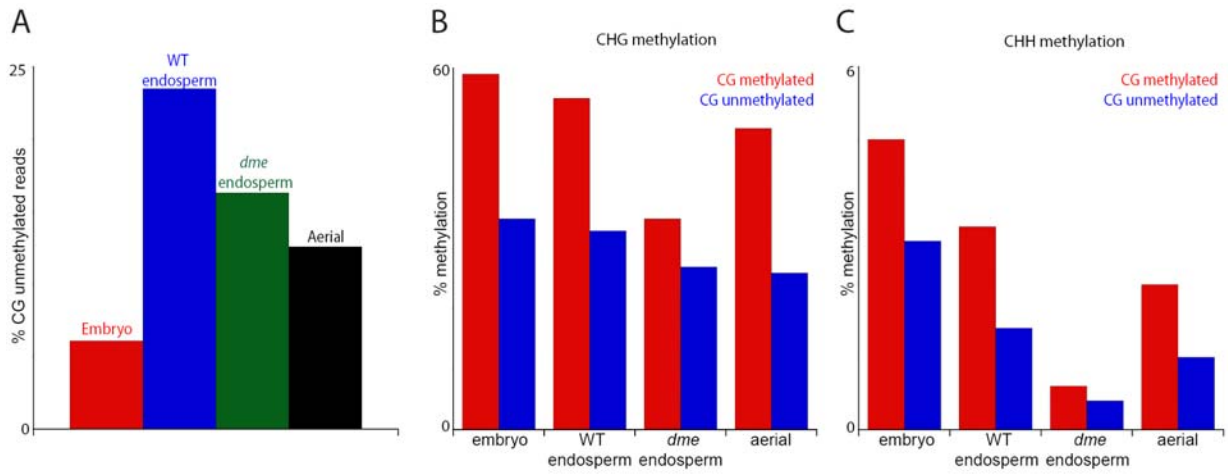

fig. S3. We selected reads from each tissue that overlapped transposable elements with embryo CG methylation above 70% and at least 20 informative sequenced Cs. Reads without CG sites were eliminated, and the rest were separated into two categories: reads that have at least one methylated CG site and reads that have none. Percentage of reads with no CG methylation in each tissue is shown in (**A**); CHG and CHH fractional methylation for each subset is shown in (**B**) and (**C**), respectively. CG methylation in reads with at least one methylated CG is 95% in embryo, 88% in WT endosperm, 92% in *dme* endosperm, and 91% in aerial tissues.

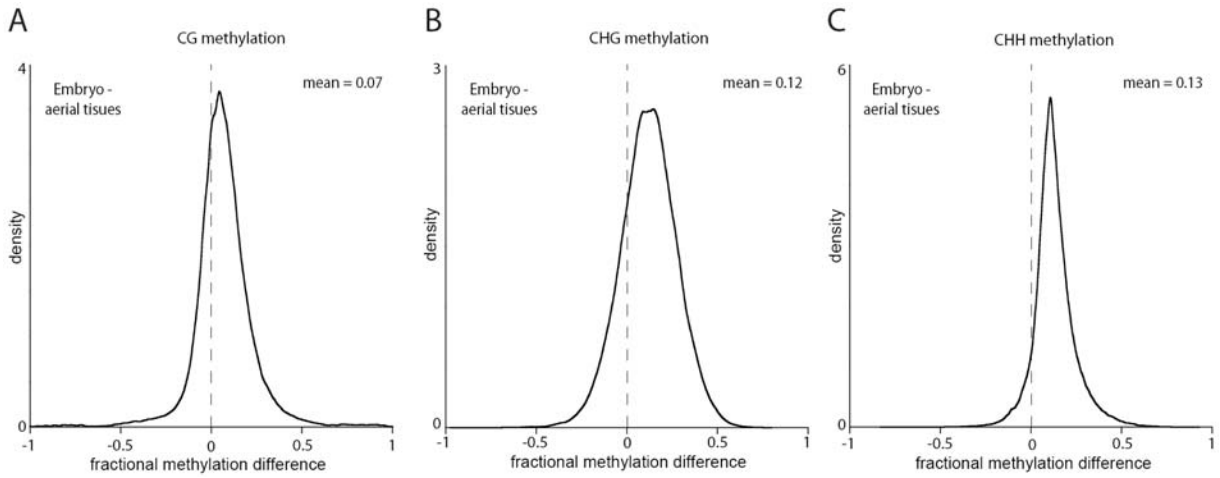

fig. S4. (**A-C**) Kernel density plots, which have the effect of tracing the frequency distribution, of the differences between embryo and aerial tissue methylation. Methylation differences for 50 bp windows containing at least 20 informative sequenced cytosines are shown. Differences for windows with fractional CG methylation of at least 0.7 in one of the tissues are shown in (**A**), differences for windows with fractional CHG methylation of at least 0.5 in one of the tissues are shown in (**B**) and differences for windows with fractional CHH methylation of at least 0.1 in one of the tissues are shown in (**C**).

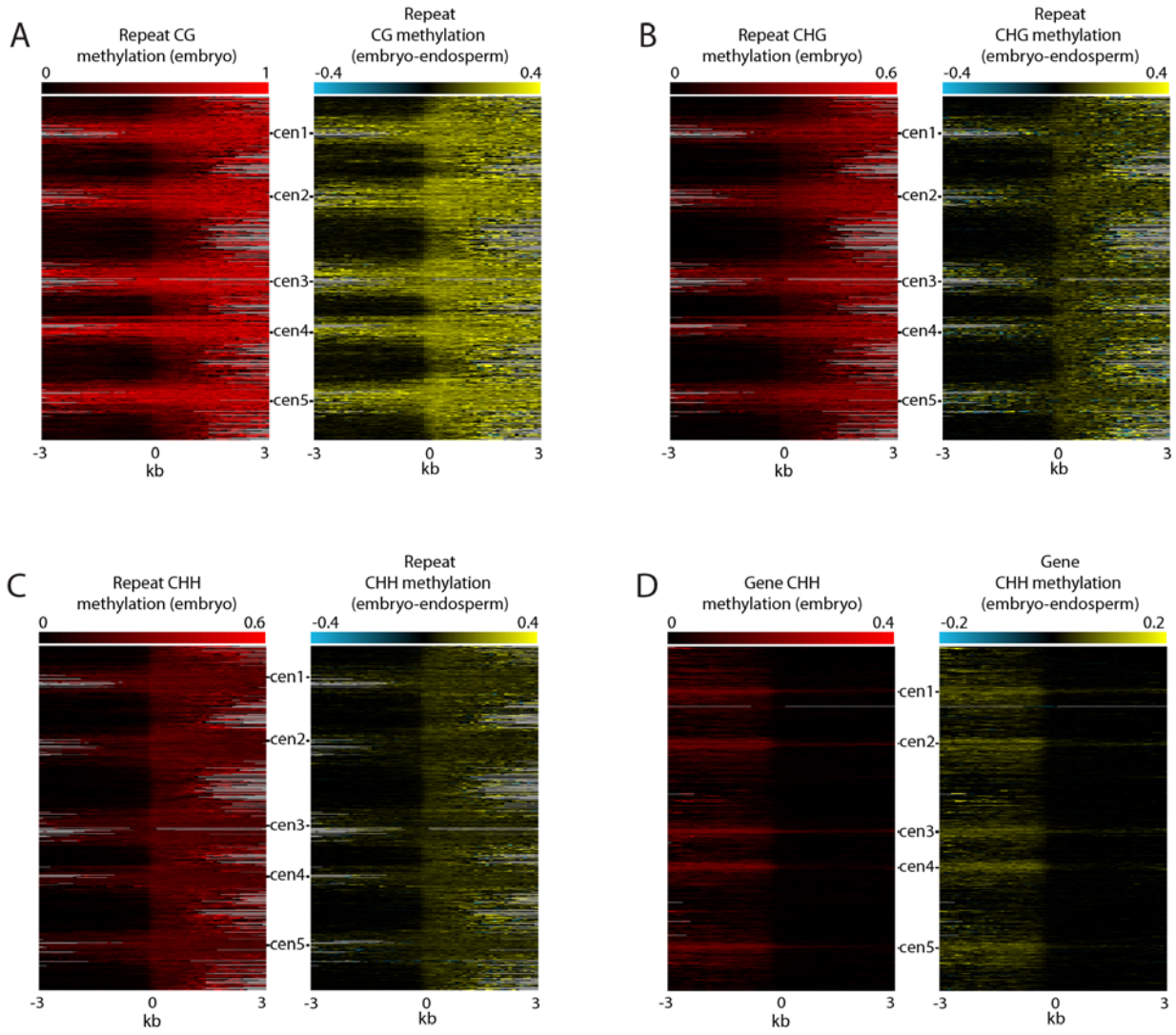

fig. S5. (**A-D**) All TAIR 8 annotated transposable elements and repeats (31,076; **A-C**) or genes (28,244; **D**) were aligned at the 5' end and stacked from the top of chromosome 1 to the bottom of chromosome 5. Embryo methylation is displayed as a heat map in the left panel, differences between embryo and WT endosperm in the right panel. CG methylation of transposable elements is shown in (**A**), CHG methylation of transposable elements is shown in (**B**), CHH methylation of transposable elements is shown in (**C**), and CHH methylation of genes is shown in (**D**).

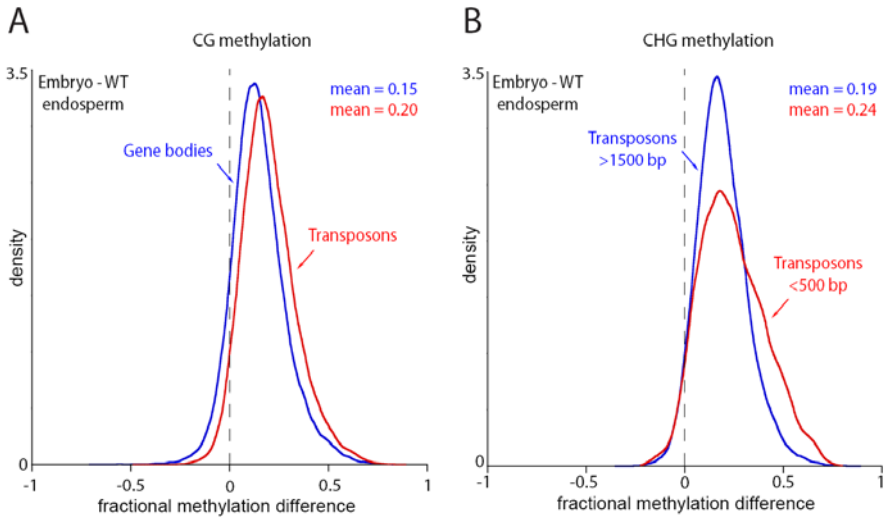

fig. S6. Kernel density plots, which have the effect of tracing the frequency distribution, of the differences between embryo and WT endosperm methylation. CG methylation differences for 50 bp windows containing at least 20 informative sequenced cytosines and with methylation of at least 0.7 are shown. Differences for windows representing transposable elements (red trace) and gene bodies (blue trace) are shown in (**A**) and differences for windows representing transposable elements shorter than 500 bp (red trace) and transposable elements longer than 1500 bp (blue trace) are shown in (**B**).

References

S1. J.-P. Vielle-Calzada, R. Baskar, U. Grossniklaus, *Nature* **404**, 91 (2000).
S2. M. Gehring *et al.*, *Cell* **124**, 495 (2006).
S3. T. Kinoshita, R. Yadegari, J. J. Harada, R. B. Goldberg, R. L. Fischer, *Plant Cell* **11**, 1945 (1999).
S4. R. Lister *et al.*, *Cell* **133**, 523 (2008).
S5. R. Cronn *et al.*, *Nucleic Acids Res* **36**, e122 (2008).
S6. A. Meissner *et al.*, *Nature* **454**, 766 (2008).
S7. H. Jiang, W. H. Wong, *Bioinformatics* **15**, 2395 (2008).
S8. S. J. Cokus *et al.*, *Nature* **452**, 215 (2008).
S9. P. E. Jullien, T. Kinoshita, N. Ohad, F. Berger, *Plant Cell* **18**, 1360 (2006).
S10. G. Makarevich, C. B. R. Villar, A. Erilova, C. Kohler, *J Cell Sci* **121**, 906 (2008).
S11. J. Penterman *et al.*, *Proc Natl Acad Sci U S A* **104**, 6752 (2007).
S12. S. Tiwari *et al.*, *Plant Cell* **20**, 2387 (2008).
S13. R. K. Tran *et al.*, *Genome Biol* **6**, R90 (2005).
S14. W. J. J. Soppe *et al.*, *Mol Cell* **6**, 791 (2000).