Conserved imprinting associated with unique epigenetic signatures in the Arabidopsis genus

Maja Klosinska[1,4], Colette L. Picard[1,2,4], Mary Gehring[1,3]

[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142

[2]Computational and Systems Biology Graduate Program, Massachusetts Institute of Technology, Cambridge, MA 02139

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

[4]Equal Contribution

Key words: genomic imprinting, DNA methylation, endosperm, Arabidopsis, CHG gene body hypermethylation, parental conflict

Abstract

In plants, imprinted gene expression occurs in endosperm seed tissue and is sometimes associated with differential DNA methylation between maternal and paternal alleles [1]. Imprinting is theorized to have been selected for because of conflict between parental genomes in offspring [2], but most studies of imprinting have been conducted in *Arabidopsis thaliana*, an inbred primarily self-fertilizing species that should have limited parental conflict. We examined embryo and endosperm allele-specific expression and DNA methylation genome-wide in the wild outcrossing species *Arabidopsis lyrata*. Here we show that the majority of *A. lyrata* imprinted genes also exhibit parentally-biased expression in *A. thaliana*, suggesting that there is evolutionary conservation in gene imprinting. Surprisingly, we discovered substantial interspecies differences in methylation features associated with paternally expressed imprinted genes (PEGs). Unlike in *A. thaliana*, the maternal allele of many *A. lyrata* PEGs was hypermethylated in the CHG context. Increased maternal allele CHG methylation was associated with increased expression bias in favor of the paternal allele. We propose that CHG methylation maintains or reinforces repression of maternal alleles of PEGs. These data suggest that while the genes subject to imprinting are largely conserved, there is flexibility in the epigenetic mechanisms employed between closely related species to maintain monoallelic expression. This supports the idea that imprinting of specific genes is a functional phenomenon, and not simply a byproduct of seed epigenomic reprogramming.

Genomic imprinting is a form of epigenetic gene regulation in flowering plants and mammals in which alleles of genes are expressed in a parent-of-origin dependent manner. Allele-specific gene expression profiling has identified hundreds of imprinted genes in *A. thaliana*, maize, and rice endosperm, the functions of which are largely unknown [3-10]. Allelic differences in DNA methylation and chromatin modification between maternal and paternal alleles are important for establishing and maintaining imprinted expression [1]. The emerging picture from multiple species is that the paternal allele of PEGs is associated with DNA methylation, while the silent maternal allele is hypomethylated and bears the Polycomb Repressive Complex 2 (PRC2) mark H3K27me3 [11,12].

Several evolutionary theories have been proposed to describe processes that would select for fixation of this unusual pattern of gene expression [13]. The kinship or parental conflict theory posits that imprinting is selected for because of asymmetric relatedness among kin [2,13]. In species where the maternal parent directly provisions growing progeny and has offspring by multiple males, maternally and paternally inherited genomes are predicted to have conflicting interests with regard to the extent of maternal investment. Paternally inherited alleles are expected to favor maternal investment at the expense of half-siblings.

Low conservation of imprinting between *A. thaliana* and monocots[14], limited conservation between rice and maize[14], evidence for intraspecific variation in imprinting [6], and lack of strong phenotypes for some imprinted gene mutants has cast doubt on whether imprinting of particular genes is functionally important. Additionally, although some imprinted genes are associated with differential methylation, it has been suggested that imprinted expression is simply a byproduct of endosperm DNA methylation changes – changes that could have a primary function outside of imprinting regulation [15,16]. We were motivated by these considerations and by predictions of the parental conflict theory to compare imprinting and seed DNA methylation between two closely related species that differ in breeding strategy. *A. lyrata* and *A. thaliana* diverged approximately 13 million years ago [17]. Although *A. thaliana* outcrosses to some extent in the wild, as an obligate outcrosser *A. lyrata* should be subject to a higher degree of parental conflict than *A. thaliana* and should therefore be under greater pressure to maintain imprinting.

To identify *A. lyrata* imprinted genes, we performed mRNA-seq on parental strains and F_1 hybrid embryo and endosperm tissue derived from crosses between the sequenced *A. lyrata* strain MN47 (MN) and a strain from Karhumäki (Kar) (**Supplementary Figure 1, Supplementary Figure 2, Supplementary Tables 1 and 2**). After reannotating *A. lyrata* genes based on our extensive RNA-seq data (see **Supplementary Methods**), sequence polymorphisms between MN and Kar were used to quantify the contributions of each parental genome to gene expression. All possible pairwise comparisons (n=12) of parent-of-origin bias among three MN x Kar and four Kar x MN reciprocal cross replicates were performed to identify imprinted genes using the same criteria we previously applied to *A. thaliana* [6]. Only genes that were defined as imprinted in at least 40% of comparisons were included in the final set (**Figure 1, Supplementary Tables 3 and 4,** see **Supplementary Methods** for details of imprinting criteria). This analysis yielded 49 paternally expressed imprinted genes (PEGs) and 35 maternally expressed imprinted genes (MEGs) in endosperm (**Figure 1A**). Allele-assignment calls for thirteen genes, including both imprinted and non-imprinted genes, were validated by pyrosequencing (**Supplementary Figure 3**). As expected [3,5], there was little evidence for imprinting in embryos (**Figure 1A**).

We compared *A. lyrata* and *A. thaliana* endosperm imprinted genes (**Figure 1, Supplementary Figure 4, Supplementary Table 4**). Of the *A. lyrata* PEGs for which there were sufficient data available in *A. thaliana*, 72% (26/36) were also paternally biased in *A. thaliana* with 50% (18/36) meeting all stringent criteria for being designated as a PEG in both species (**Figure 1B**). Conserved PEGs encoded DNA binding proteins and genes related to chromatin modification, among others (**Supplementary Table 4**). Of the *A. lyrata* MEGs for which there were sufficient data in *A. thaliana*, 70% (12/17) were also significantly maternally biased in *A. thaliana*, with 35% (6/17) meeting all criteria for being called as a MEG in both datasets (**Figure 1B**). The conserved MEGs included the Polycomb group gene *FIS2*, the F-box gene *SDC*, another F-box gene, and three genes encoding DNA binding proteins. While previous research has identified somewhat more imprinted genes in *A. thaliana* than what we describe in *A. lyrata*, these studies involved multiple accessions and assessed imprinting for a greater total number of genes [6]. The majority of genes that were imprinted in *A. thaliana* but not in *A. lyrata* lacked sufficient data to make an imprinting designation in *A. lyrata* (**Supplementary**

Figure 4). Thus, it is presently unclear whether the number of imprinted genes differs significantly between the species. All of the genes that are commonly imprinted among *A. thaliana* and cereals[14] were also imprinted in *A. lyrata*.

Many mammalian imprinted genes are clearly involved in growth regulation, including genes for nutrient uptake and feeding behavior [18]. By contrast, we find that proteins encoded by conserved plant imprinted genes are predicted to regulate or effect the expression of many other genes (chromatin proteins and transcription factors) or protein abundance (F-boxes). We also found that some pathways, rather than orthologous genes, were imprinted in both species, as has been previously noted for imprinting of different subunits of the PRC2 complex among Arabidopsis and cereals[19]. In *A. thaliana*, the large subunit of RNA Polymerase IV, *NRPD1*, which functions in RNA-directed DNA methylation (RdDM) [20], is a PEG [5,6]. Although we did not find evidence for imprinting of the *NRPD1* gene in *A. lyrata*, homologues of two other genes involved in RdDM were PEGs (**Supplementary Table 4**): *NRPD4/NRPE4/RDM2* (AL946699), which encodes a common subunit of Pol IV and Pol V, and *RRP6L1* (AL337734), which encodes an exosomal protein that impacts RdDM. Thus, in both species the function of RdDM in the endosperm is under paternal influence, but this is achieved *via* different genes.

The kinship theory is essentially an argument about optimal total gene expression levels in offspring [13]. We therefore evaluated the expression levels and patterns of imprinted genes. MEGs appear to be primarily endosperm-specific genes; they have much lower than average expression in embryos and flower buds, and much higher than average expression in the endosperm (**Figure 1C**). Conversely, PEGs were more highly expressed in all tissues than genes on average, and showed more modest expression increases in endosperm, suggesting that the expression of MEGs and PEGs is regulated differently. We also compared the percent maternal transcripts for homologous imprinted *A. lyrata* and *A. thaliana* genes (**Figure 1D**). Conserved MEGs and PEGs exhibited similar degrees of parental bias in the two species (**Figure 1D**). However, comparison of the *A. thaliana* and *A. lyrata* gene expression level for individual imprinted genes indicated that the overall expression level of PEGs was higher in *A. lyrata* than in *A. thaliana* (**Figure 1E)**. These findings are consistent with stronger selection for higher expression of PEGs in species with greater parental conflict, such as obligate outcrossers [13].

In *A. thaliana*, active DNA demethylation by the 5-methylcytosine DNA glycosylase DME in the central cell (the female gamete that is the progenitor of the endosperm) before fertilization is essential for establishing gene imprinting at many loci [1]. Imprinting of many *A. thaliana* genes, particularly PEGs, is correlated with maternal allele demethylation of proximal sequences corresponding to fragments of transposable elements [6,21]. *A. lyrata* PEGs were somewhat enriched for the presence of TEs in 5' regions compared to all genes, with 30 out of 49 PEGs (61%) associated with at least one TE within 2 kb 5', compared to 51% of all genes (**Supplementary Table 4**). To test if the relationship between methylation and imprinting was conserved in *A. lyrata*, we profiled methylation genome-wide in MN x MN flower bud, embryo, and endosperm tissue by whole genome bisulfite sequencing. Shared and novel endosperm methylation features were observed compared to *A. thaliana* (**Figure 2, Figure 3, Supplementary Figure 5, Supplementary Table 5**). In plants, DNA methylation is found in CG, CHG, and CHH sequence contexts. CG methylation was strongly decreased in TEs and in the 5' and 3' regions of genes in endosperm relative to other tissues (**Figure 2A, Supplementary Figure 5**). By profiling allele-specific DNA methylation in F_1 embryo and endosperm from Kar females crossed to MN males, we determined that maternally inherited DNA was primarily responsible for endosperm CG hypomethylation (**Figure 2B**). These data suggest that, like in *A. thaliana*, *A. lyrata* maternally-inherited genomes are actively demethylated before fertilization[6,21,22].

By contrast, we were surprised to discover that *A. lyrata* endosperm had a non-CG DNA methylation profile distinct from *A. thaliana*. This was unexpected because DNA methylation patterns in *A. lyrata* vegetative tissues display similar features to *A. thaliana*, although overall methylation levels are higher (**Supplementary Figure 5**). We found that average CHG methylation in gene bodies was increased in endosperm compared to embryo (**Figure 2A**), a phenotype not observed in wild type *A. thaliana* endosperm profiled at similar developmental stages [6,22] (**Supplementary Figure 5**). To determine whether differences in aggregate methylation profiles represented small changes in many regions or larger changes in specific regions of the genome, we compared embryo and endosperm methylation profiles to identify differentially methylated regions (DMRs) [6]. Like in *A. thaliana*, the most abundant class of DMRs were less CG methylated in the endosperm compared to the embryo, with 38% of these falling within 2 kb 5' of genes and 34% within 2 kb 3' of genes (**Supplementary Table 6**). Regions that gained CHG methylation in MN x MN endosperm displayed markedly different characteristics; 84% fell within gene bodies, corresponding to 1606 genes (**Figure 2C, Supplementary Table 6**). CHG endosperm hypermethylated DMRs were also longer than all other DMR types (mean length = 564 bp with 400 bp standard deviation) (**Supplementary Table 6**). CHG gene body hypermethylation was also observed in Kar x MN endosperm, although on fewer genes (n=194). Allele-specific analysis of methylation indicated that endosperm CHG hypermethylation was specific to maternally inherited alleles (**Figure 2D**).

Methylation within gene bodies is usually restricted to the CG context, which is maintained after DNA replication by the maintenance methyltransferase MET1. CHG methylation, normally not found in genes, is maintained by the DNA methyltransferase CMT3, which directly binds to the repressive histone modification H3K9me2 [23]. When accompanied by H3K9me2, CHG gene body methylation is associated with transcriptional repression [24]. We found that gain of gene body CHG methylation in *A. lyrata* endosperm was associated with reduced gene expression (**Supplementary Figure 6**). Of the CHG hypermethylated genes with enough coverage to evaluate differential expression (n=1225), 338 were significantly less expressed in endosperm than in embryo, compared to 159 significantly more highly expressed in endosperm. This represents a significant enrichment of CHG hypermethylated genes among genes less expressed in endosperm than embryo ($P(x \geq 338) = 1.766 \times 10^{-21}$, hypergeometric test) and a significant depletion among genes upregulated in endosperm ($P(x \leq 159) = 2.04 \times 10^{-10}$, see **Supplementary Methods**). The mechanism responsible for CHG gene body hypermethylation in *A. lyrata* endosperm remains unclear. We found significant overlap between *A. thaliana* genes that gain CHG or H3K9me2 in *ibm1* mutants and CHG hypermethylation of orthologous genes in *A. lyrata* endosperm (**Supplementary Figure 7**). *IBM1* encodes a histone lysine demethylase that prevents accumulation of H3K9me2, and thus accumulation of CHG methylation, in genes [24] . *IBM1* transcript abundance was lower in the endosperm than embryo (**Supplementary Figure 7**). In *A. thaliana*, methylation in the long intron of *IBM1* is required for proper transcript splicing and production of an enzymatically active protein[25]. We found that *A. lyrata IBM1* exhibited decreased CG and non-CG methylation and increased accumulation of RNA-seq reads in the long intron in endosperm relative to embryo (**Supplementary Figure 7**). However, *A. thaliana* endosperm also had reduced methylation in the long intron and decreased *IBM1* transcript abundance compared to the embryo (**Supplementary Figure 7**). Thus, differences in *IBM1* expression alone are not sufficient to explain CHG hypermethylation in *A. lyrata* endosperm compared to *A. thaliana*, although reduced IBM1 activity is likely part of the mechanism.

Several of the observed endosperm methylation features were correlated with gene imprinting. More than half of the *A. lyrata* MEGs and approximately one third of PEGs were associated with endosperm CG hypomethylated DMRs in the 2 kb region upstream of the transcriptional start site, whereas only 11% of non-imprinted genes were similarly associated with these DMRs (**Figure 3, Supplementary Table 4, Supplementary Figure 8, Supplementary Figure 9**). CG hypomethylation occurred specifically on the maternally inherited allele (**Supplementary Figure 8**). Thus, reduction of CG methylation by active demethylation is likely also an important component of the *A. lyrata* imprinting mechanism. We found a striking and non-mutually exclusive association between PEGs and endosperm CHG hypermethylation. Almost 60% of PEG gene bodies (n=27) were CHG hypermethylated, and about one-third were also associated with a 5' or 3' CG hypomethylated DMR (**Supplementary Table 4**). The average methylation profile of PEGs containing a CHG endosperm hypermethylated DMR indicated a very strong increase in CHG methylation across the entire gene body, which was specific to the maternally inherited allele (**Figure 3**, **Figure 4**). Results were validated for two PEGs, homologues of AT5G10950 and AT5G26210, by locus-specific BS-PCR (**Figure 4, Supplementary Figure 10**). In both *A. thaliana* and *A. lyrata* these genes were associated with CG or CHH endosperm hypomethylated DMRs in 5' regions, but were additionally associated with gene body CHG hypermethylated DMRs in *A. lyrata*. Interestingly, gain of CHG methylation on the maternal allele was often accompanied by loss of CG gene body methylation, while paternally inherited alleles retained CG gene body methylation and had a similar methylation profile to embryo alleles (**Figure 4, Supplementary Table 4**). For the 22 PEGs lacking a gene body CHG hypermethylated DMR, half had a CG hypomethylated DMR in the flanking regions 2 kb 5' or 3', more like typical *A. thaliana* PEGs (**Supplementary Table 4**). Interestingly, these genes largely lacked CG gene body methylation in all tissues (**Figure 3A**). Thus, there appear to be at least two classes of PEGs in terms of methylation features (**Figure 3A**), which may correspond to different modes of epigenetic regulation. PEGs conserved with *A. thaliana* are found in both classes, although the majority (12/18) are CHG hypermethylated (**Supplementary Table 4**).

To determine if there was a quantitative relationship between gain of CHG methylation and allelic expression bias, we plotted the difference in CHG methylation between maternal alleles in the embryo and endosperm relative to the ratio of maternal to paternal allele transcripts (**Figure 3C**). The degree to which CHG methylation was gained on the maternal allele in endosperm relative to embryo was positively correlated with the extent of paternal allele expression bias in endosperm. In addition, PEGs were clearly distinct from other genes that gained CHG gene body methylation; they tended to exhibit greater gain of CHG methylation (**Figure 3C**) and were also hypermethylated along more of their length than all CHG hypermethylated genes (56% vs. 29%). Thus, a greater extent and amount of maternal allele CHG hypermethylation is correlated with more paternally biased transcription. These data suggest that CHG methylation, perhaps accompanied by gain of H3K9me2, represses the maternal alleles of PEGs. It is unknown whether gene body CHG methylation is established on maternal alleles before or after fertilization. Demethylation of the *IBM1* regulatory intron (**Supplementary Figure 7**) could be initiated before fertilization in the central cell, leading to its downregulation and an increase in CHG methylation specifically on maternal alleles, which would then be maintained after fertilization. Alternatively, if maternal allele CHG methylation occurs post-fertilization, then CMT3 must be able to distinguish maternally and paternally inherited alleles. Retention of CG gene body methylation on the paternal alleles of PEGs (**Figure 4**) could possibly protect them from gain of CHG methylation. Interestingly, gain of gene body CHG methylation was also recently shown to occur in both *A. thaliana* endosperm and embryos when wild type plants were pollinated by diploid hypomethylated pollen[26]. Diploid pollen creates triploid seeds with tetraploid endosperm that usually abort, but seed abortion is suppressed when the pollen is hypomethylated due to mutations in *met1*. Many of the genes that gain CHG methylation and have reduced expression in triploid rescued seeds are PEGs [26]. However, this phenotype appears to be distinct from what we observed; the CHG methylation gain is much more modest than what we have described in wild type *A. lyrata* endosperm, and only one conserved PEG was affected[26]. Our data further suggest that gene body CHG hypermethylation is not a state restricted to mutant tissues, but can occur in a developmentally regulated manner that could be important for maintaining gene expression programs.

This is the first study to compare imprinting between two closely related plant species that differ in breeding strategy. *A. lyrata* and *A. thaliana* homologous imprinted genes are epigenetically modified in a distinct manner despite the close relatedness of the species (**Figure**

4). Allele-specific maintenance of gene repression by the PRC2 complex is an important component of the imprinting mechanism in *A. thaliana* and other species [11,12]. The PRC2 complex silences the hypomethylated maternal allele of PEGs, while the methylated paternal allele is expressed. Several studies have suggested that H3K9me2 and H3K27me3 are repressive marks that can substitute for one another in mutant contexts [26,27]. We suggest that this substitution can also occur in wild type tissues, and favor the hypothesis that in *A. lyrata* endosperm the maternal allele of at least a subset of PEGs is repressed by CHG methylation/H3K9me2. Overall, our results point to high conservation of imprinting accompanied by a distinct epigenetic signature, at least for PEGs. If the mechanism of imprinting is different but the genes that are imprinted are the same, this argues that imprinting is not simply a byproduct of endosperm methylation dynamics, but that imprinted expression of specific genes is under selection. Thus, the means by which monoallelic expression can be achieved are plastic, but the genes subject to this regulation are conserved.

METHODS

Plant material

Arabidopsis lyrata MN47 (MN) seeds were obtained from the Arabidopsis Biological Resource Center (CS22696); seeds from the Karhumäki (Kar) strain were a gift from Dr. Outi Savolainen, University of Oulu, Finland. Plants were grown in a greenhouse with 16 hr of light at 21°C or in a growth chamber with 16 hr of light (120 µMol), 20°C and 50% humidity, vernalized at 4°C for a month after rosettes had formed, and then returned to the greenhouse/growth chamber. With the exception of MN x Kar crosses (female parent in cross is listed first), flowers were not emasculated prior to pollination. MN plants, which are able to self-pollinate at low frequency, were emasculated and pollinated 2 days later. Seeds were dissected into endosperm, embryo, and seed coat portions at seed development stages ranging from torpedo to bent cotyledon, around 14-19 days post pollination, depending on growth temperature, genotype, and age of the maternal parent. In addition, flower bud tissue was collected from MN plants and from Kar x MN F_1 hybrid plants.

mRNA-Seq

RNA was isolated from endosperm, embryo and seed coat samples using the RNAqueous Micro Kit (Ambion). Input for mRNA-seq library construction varied from 120 to 800 ng DNase I-treated RNA. Strand-specific libraries were prepared by the Whitehead Institute Genome Technology Core using the Integenex PolyA prep protocol (Wafergen Biosystems) or using the Illumina TruSeq Stranded mRNA Kit. Libraries were multiplexed and sequenced on an Illumina HiSeq 2000 machine by the Whitehead Institute Genome Technology Core using a paired end protocol. See **Supplementary Table 1** for details of library prep for specific samples. Reads were aligned to the MN47 reference genome[28] using Tophat v2.0.13[29]. See **Supplementary Methods** for details on mRNA-seq analysis parameters, SNP discovery, and updated annotation of the *A. lyrata* genome.

Imprinting analysis

Imprinted genes were identified using our previously described method (**Supplementary Methods**) [5,6]. To assess whether *A. lyrata* imprinted genes were also imprinted in *A. thaliana*, we examined data from Pignatta *et al*. 2014 [6]. If the *A. thaliana* homologue of the *A. lyrata* imprinted gene was not called as an imprinted gene, we took the *A. thaliana* reciprocal cross comparison that showed the strongest degree of parental bias and determined why the gene was not called imprinted (see categories in **Figure 1B**). If any *A. lyrata* imprinted genes had the same *A. thaliana* homologue it was only counted once. To perform the reverse analysis, in which we assessed whether all genes considered imprinted in Pignatta *et al*. 2014 (the union set of MEGs and PEGs) were also imprinted in *A. lyrata*, we determined for each of the 12 comparisons in our *A. lyrata* data whether the gene was called imprinted, and if not, why (**Supplementary Figure 4**). If more than 7 comparisons lacked data, a gene was considered to have insufficient reads. If >40% of comparisons with data called the gene imprinted, it was considered imprinted. If <40% of comparisons with data called the gene imprinted or it did not meet the % maternal cutoff (but met all other imprinting criteria), the gene was considered parentally biased but not meeting the % maternal cutoff. Similarly, if <40% of comparisons called the gene imprinted and it failed the % maternal cutoff and failed the imprinting factor (IF) cutoff but met initial parental bias p-value cutoffs, it was considered parentally biased but with imprinting factor too low. All other genes were considered to have no significant parental bias. For *A. thaliana* imprinted genes with multiple *A. lyrata* homologues that differed in imprinting status, the imprinting status of the most parentally biased homologue was used to obtain counts in **Supplementary Figure 4**.

Whole genome bisulfite sequencing

Genomic DNA was extracted in duplicate from MN flower buds and from seeds dissected into embryo and endosperm from MN x MN and Kar x MN crosses. DNA was isolated from fresh tissue using the DNeasy Plant Mini Kit (Qiagen) (buds) or a CTAB method (embryo and endosperm) and RNase treated. 250-500 ng of DNA was used for bisulfite treatment with the MethylCode Bisulfite Conversion Kit (Invitrogen). Libraries were constructed using the EpiGnome Methyl-seq Kit (Epicentre Biotechnologies) and were sequenced on an Illumina HiSeq 2000 using a 40x40 or 100x100 paired end protocol at the Whitehead Institute Genome Technology Core (see **Supplementary Table 5**). Reads were aligned to the genome using Bismark v.0.13.0[30]. To identify differentially methylated regions, the genome was divided into consecutive 300 bp windows that overlapped by 200 bp. Each window was assessed as a potential DMR between two samples (e.g. embryo and endosperm) using the method described in Pignatta *et al.* [6]. See **Supplementary Methods** for additional details.

Data access

High throughput sequencing data has been deposited in NCBI GEO under accession GSE76076.

REFERENCES

1. Gehring, M. Genomic Imprinting: Insights From Plants. *Annu. Rev. Genet.* **47,** 187–208 (2013).

2. Haig, D. & Westoby, M. Parent-specific gene expression and the triploid endosperm. *Am. Nat.* **134,** 147–155 (1989).

3. Hsieh, T.-F. *et al.* Regulation of imprinted gene expression in Arabidopsis endosperm. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 1755–1762 (2011).

4. Wolff, P. *et al.* High-resolution analysis of parent-of-origin allelic expression in the Arabidopsis Endosperm. *PLoS Genet.* **7,** e1002126 (2011).

5. Gehring, M., Missirian, V. & Henikoff, S. Genomic Analysis of Parent-of-Origin Allelic Expression in Arabidopsis thaliana Seeds. *PLoS ONE* **6,** e23687 (2011).

6. Pignatta, D. *et al.* Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *Elife* e03198 (2014).

7. Luo, M. *et al.* A genome-wide survey of imprinted genes in rice seeds reveals imprinting primarily occurs in the endosperm. *PLoS Genet.* **7,** e1002125 (2011).

8. Waters, A. J. *et al.* Parent-of-Origin Effects on Gene Expression and DNA Methylation in the Maize Endosperm. *Plant Cell* **23,** 4221–4233 (2012).

9. Xin, M. *et al.* Dynamic expression of imprinted genes associates with maternally controlled nutrient allocation during maize endosperm development. *Plant Cell* **25,** 3212–3227 (2013).

10. Zhang, M. *et al.* Extensive, clustered parental imprinting of protein-coding and noncoding RNAs in developing maize endosperm. *Proc. Natl. Acad. Sci. U.S. A.* **108,** 20042–20047 (2011).

11. Zhang, M. *et al.* Genome-wide high resolution parental-specific DNA and histone methylation maps uncover patterns of imprinting regulation in maize. *Genome Res.* **24,** 167–176 (2014).

12. Moreno-Romero, J., Jiang, H., Santos-González, J. & Köhler, C. Parental epigenetic asymmetry of PRC2-mediated histone modifications in the Arabidopsis endosperm.

EMBO J. **35,** 1298–1311 (2016).

13. Patten, M. M. *et al.* The evolution of genomic imprinting: theories, predictions and empirical tests. *Heredity* **113,** 119–128

14. Waters, A. J. *et al.* Comprehensive analysis of imprinted genes in maize reveals allelic variation for imprinting and limited conservation with other species. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 19639–19644 (2013).

15. Berger, F., Vu, T. M., Li, J. & Chen, B. Hypothesis: Selection of Imprinted Genes is Driven by Silencing Deleterious Gene Activity in Somatic Tissues. *Cold Spring Harb. Symp. Quant. Biol.* (2012).

16. Kawashima, T. & Berger, F. Epigenetic reprogramming in plant sexual reproduction. *Nat Rev Genet* **15,** 613–624 (2014).

17. Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 18724–18728 (2010).

18. Morison, I. M. & Reeve, A. E. A catalogue of imprinted genes and parent-of-origin effects in humans and animals. *Hum. Mol. Genet.* **7,** 1599–1609 (1998).

19. Danilevskaya, O. N. *et al.* Duplicated fie genes in maize: expression pattern and imprinting suggest distinct functions. *Plant Cell* **15,** 425–438 (2003).

20. Matzke, M. A. & Mosher, R. A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet* **15,** 394–408 (2014).

21. Gehring, M., Bubb, K. L. & Henikoff, S. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* **324,** 1447–1451 (2009).

22. Ibarra, C. A. *et al.* Active DNA Demethylation in Plant Companion Cells Reinforces Transposon Methylation in Gametes. *Science* **337,** 1360–1364 (2012).

23. Du, J. *et al.* Dual Binding of Chromomethylase Domains to H3K9me2-Containing Nucleosomes Directs DNA Methylation in Plants. *Cell* **151,** 167–180 (2012).

24. Saze, H., Shiraishi, A., Miura, A. & Kakutani, T. Control of genic DNA methylation by a jmjC domain-containing protein in Arabidopsis thaliana. *Science* **319,** 462–465 (2008).

25. Rigal, M., Kevei, Z., Pélissier, T. & Mathieu, O. DNA methylation in an intron of the IBM1 histone demethylase gene stabilizes chromatin modification patterns. *EMBO J.* **31,** 2981–2993 (2012).

26. Schatlowski, N. *et al.* Hypomethylated Pollen Bypasses the Interploidy Hybridization Barrier in Arabidopsis. *Plant Cell* **26,** 3556-3568 (2014).

27. Deleris, A. *et al.* Loss of the DNA Methyltransferase MET1 Induces H3K9 Hypermethylation at PcG Target Genes and Redistribution of H3K27 Trimethylation to Transposons in Arabidopsis thaliana. *PLoS Genet.* **8,** e1003062 (2012).

28. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43,** 476–481 (2011).

29. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

30. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27,** 1571–1572 (2011).

Correspondence and requests for materials should be addressed to M.G. (mgehring@wi.mit.edu).

ACKNOWLEDGEMENTS

M.G. thanks members of the NESCent working group on Testing Theories of Genomic Imprinting for many stimulating discussions. We thank Outi Savolainen for kindly providing *A. lyrata* Karhumäki seeds, the Whitehead Institute Bioinformatics and Research Computing group for assistance, and P.R. Satyaki and Ben Williams for comments on the manuscript. This research was funded by NSF grants MCB 1121952 and 1453459 to M.G. C.L.P. is supported by an NSF graduate research fellowship.

AUTHOR CONTRIBUTIONS

M.G. conceived the project, M.K. performed experiments, C.L.P. developed and implemented computational analyses, M.K., C.L.P., and M.G. analyzed data and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

FIGURE LEGENDS

Figure 1: Identification of imprinted genes in *A. lyrata* and comparison to *A. thaliana*. a) Comparison of maternal to paternal transcript ratio (m/p) from reciprocal crosses for all genes. Counts from biological replicates were pooled for plotting. **b)** Imprinting conservation between *A. lyrata* and *A. thaliana*. Reason for lack of imprinting in *A. thaliana* indicated. **c)** Average gene expression in Kar x Kar tissues. Outliers not shown. **d)** % maternal reads for imprinted genes. Colors as in (b). **e)** Relative gene expression levels in *A. lyrata vs. A. thaliana* endosperm for imprinted genes. Log_2 ratios were calculated using DESeq2.

Figure 2: *A. lyrata* endosperm exhibits an unusual methylation profile. a) Average % methylation for genes (left) and TEs (right) in MN tissues. **b)** Average allelic % CG methylation in Kar x MN embryo and endosperm. **c)** Average % methylation for 1606 genes containing a MN x MN endosperm CHG hypermethylated DMR. Green, embryo; orange, endosperm. **d)** Average % CHG methylation in Kar x MN embryo and endosperm, with separate profiles for the maternal (Kar) and paternal (MN) alleles, over regions CHG hypermethylated in MN x MN endosperm.

Figure 3: A*. lyrata* PEGs are associated with maternal allele CHG gene body hypermethylation. a) Average MN methylation profiles for 27 PEGs containing an endosperm CHG hypermethylated DMR (left) and the 22 PEGs without a CHG DMR (right). **b)** Maternal and paternal allele CHG methylation for 27 CHG DMR PEGs **c)** Average maternal to paternal transcript ratio (m/p) for all genes containing a CHG endosperm hypermethylated DMR plotted as a function of difference in average CHG methylation on the maternal allele (left) or paternal allele (right) between endosperm and embryo. CHG methylation difference calculated within DMRs. Blue dots, PEGs.

Figure 4: Conserved PEGs exhibit distinct methylation profiles between species. a) Bisulfite-seq methylation profile and DMRs in *A. thaliana* (L*er*)[6] and *A. lyrata* (MN x MN and

allele-specific Kar x MN profiles shown) embryo and endosperm around a conserved PEG. Red tracks, CG methylation; blue, CHG methylation; green, CHH methylation. Tick marks below the line indicate Cs with sufficient coverage but no methylation; all tracks shown from 0-100%. **b-c)** Bisulfite-PCR validation of MN x MN DMRs indicated in (a). The first line is the reference sequence. **d)** Allele-specific methylation profiles in region 3 from Kar x MN endosperm.

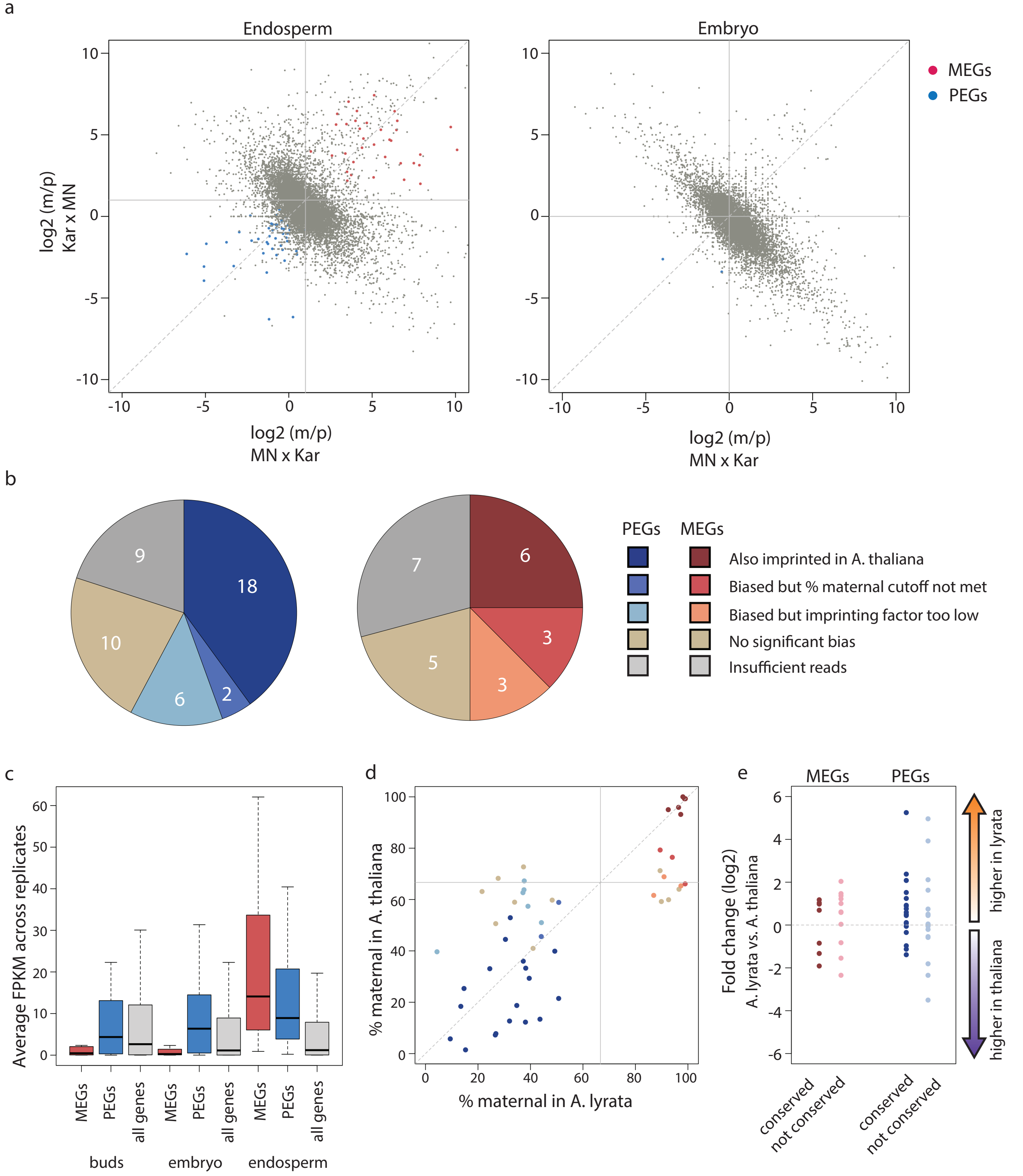

a
Endosperm
Embryo
log2 (m/p)
Kar x MN
log2 (m/p)
MN x Kar
10
5
0
-5
-10
MEGs
PEGs
b
18
2
6
10
9
6
3
3
5
7
PEGs
MEGs
Also imprinted in A. thaliana
Biased but % maternal cutoff not met
Biased but imprinting factor too low
No significant bias
Insufficient reads
c
Average FPKM across replicates
60
50
40
30
20
10
0
MEGs
PEGs
all genes
buds
embryo
endosperm
d
% maternal in A. thaliana
% maternal in A. lyrata
100
80
60
40
20
0
e
MEGs
PEGs
Fold change (log2)
A. lyrata vs. A. thaliana
6
4
2
0
-2
-4
-6
higher in lyrata
higher in thaliana
conserved
not conserved

Figure 1

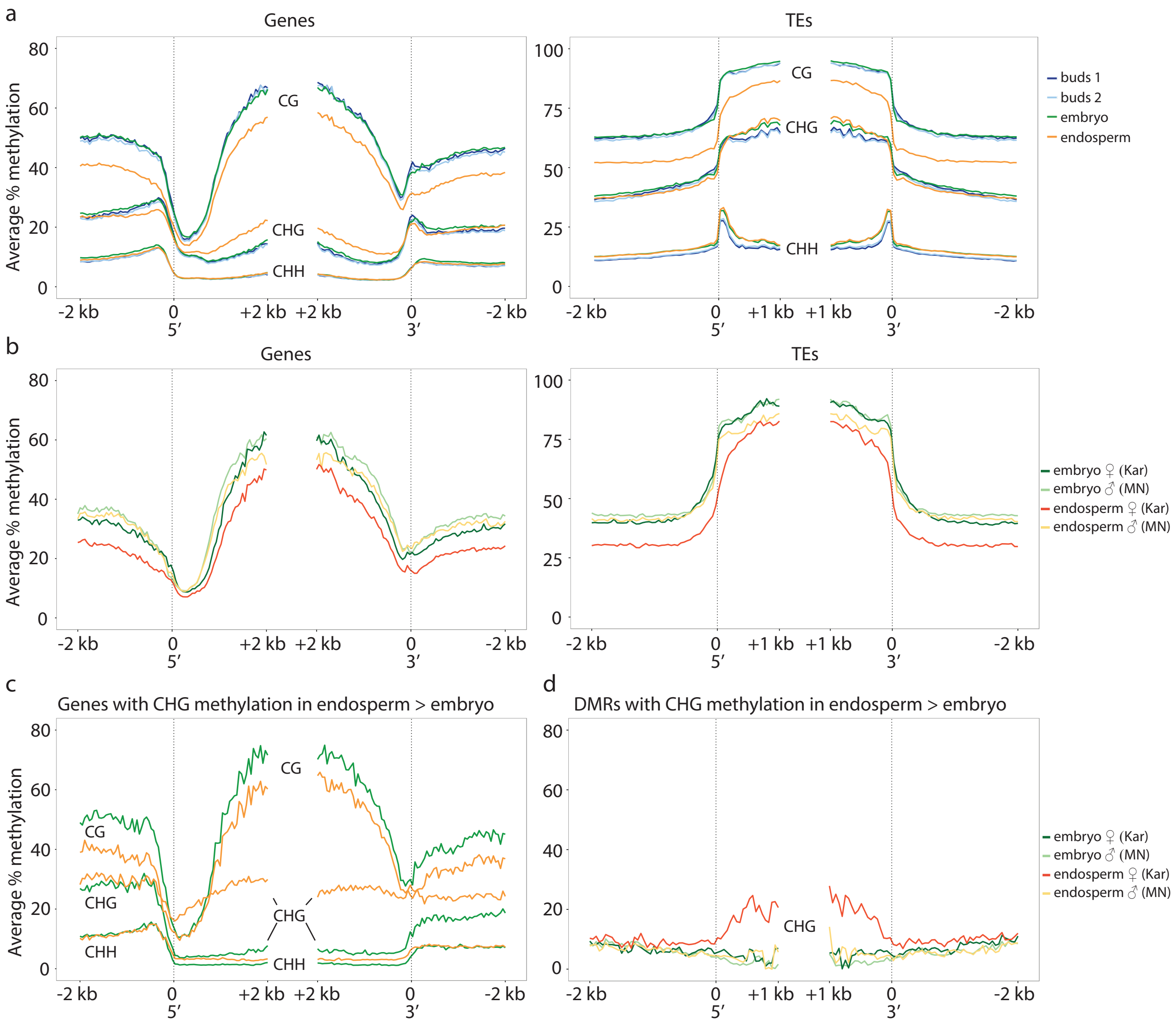

a
Genes
TEs
Average % methylation
80
60
40
20
0
100
75
50
25
CG
CHG
CHH
-2 kb
0
5′
+2 kb
+2 kb
0
3′
-2 kb
+1 kb
+1 kb
— buds 1
— buds 2
— embryo
— endosperm
b
Genes
TEs
— embryo ♀ (Kar)
— embryo ♂ (MN)
— endosperm ♀ (Kar)
— endosperm ♂ (MN)
c
Genes with CHG methylation in endosperm > embryo
CG
CHG
CHH
d
DMRs with CHG methylation in endosperm > embryo
CHG
— embryo ♀ (Kar)
— embryo ♂ (MN)
— endosperm ♀ (Kar)
— endosperm ♂ (MN)

Figure 2

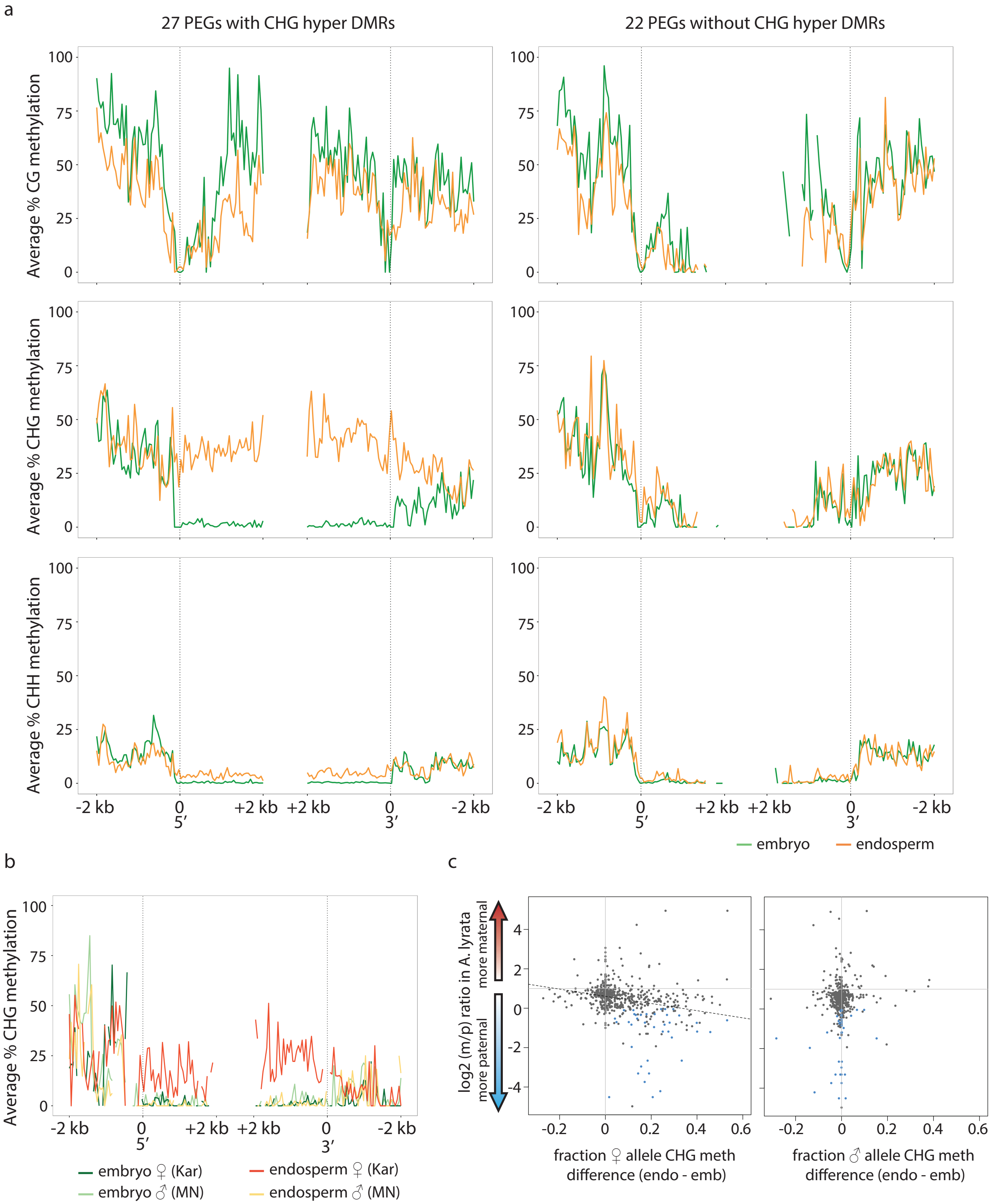

a
27 PEGs with CHG hyper DMRs
22 PEGs without CHG hyper DMRs
Average % CG methylation
Average % CHG methylation
Average % CHH methylation
100
75
50
25
0
-2 kb
0
5′
+2 kb
+2 kb
0
3′
-2 kb
embryo
endosperm
b
Average % CHG methylation
embryo ♀ (Kar)
embryo ♂ (MN)
endosperm ♀ (Kar)
endosperm ♂ (MN)
c
log2 (m/p) ratio in A. lyrata
more maternal
more paternal
4
2
0
-2
-4
-0.2
0
0.2
0.4
0.6
fraction ♀ allele CHG meth difference (endo - emb)
fraction ♂ allele CHG meth difference (endo - emb)

Figure 3

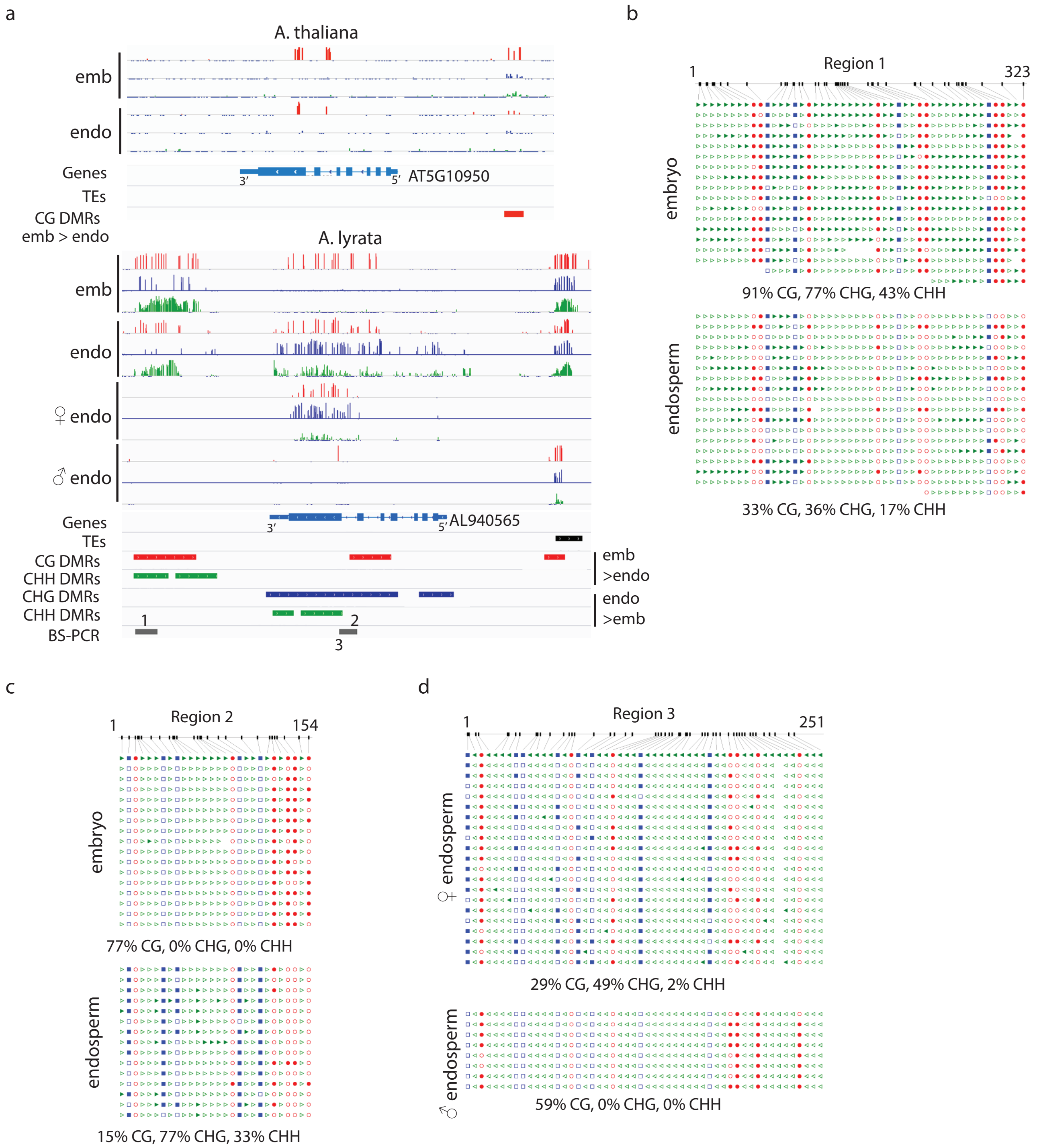

a
A. thaliana
emb
endo
Genes
TEs
CG DMRs
emb > endo
3′
5′
AT5G10950
A. lyrata
emb
endo
♀ endo
♂ endo
Genes
TEs
CG DMRs
CHH DMRs
CHG DMRs
CHH DMRs
BS-PCR
3′
5′
AL940565
emb
>endo
endo
>emb
1
2
3
b
Region 1
1
323
embryo
91% CG, 77% CHG, 43% CHH
endosperm
33% CG, 36% CHG, 17% CHH
c
Region 2
1
154
embryo
77% CG, 0% CHG, 0% CHH
endosperm
15% CG, 77% CHG, 33% CHH
d
Region 3
1
251
♀ endosperm
29% CG, 49% CHG, 2% CHH
♂ endosperm
59% CG, 0% CHG, 0% CHH

Figure 4

Supplementary Information

Supplementary Methods

mRNA-seq analysis

Preliminary filtering and mapping

Paired-end mRNA-seq reads were trimmed to remove adapters and poor quality ends using trim-galore version 0.4.0, with --stringency 3 and quality cutoff -q 25. Quality of reads both before and after trimming was determined using fastqc. Fastqc indicated that the first 8 bases of each read were of poorer quality, and these were also trimmed off using --clip_R1 8 and --clip_R2 8. Only read pairs in which both reads were at least 32 bp long after trimming were kept for mapping. Reads were aligned to the MN47 reference genome [1] using Tophat v2.0.13 [2] with minimum intron length -i 20 and maximum intron length -I 2000. The parameters for maximum mismatches, read edit distance, segment length, and mate inner distance varied based on library read length (see **Supplementary Table 1)**. A GTF file containing known gene annotations was provided using the -G option to improve alignment speed. The resulting SAM file was filtered to remove reads that mapped ambiguously to multiple locations in the genome by removing all alignments with MAPQ < 5, and PCR duplicates were removed using MarkDuplicates from Picard tools. All other alignments were kept for downstream analyses.

Updating gene annotations

After processing the libraries as described above, all reads from the 8 MN x MN libraries were pooled into a single dataset. Cufflinks v2.2.1 [3] was used to obtain a list of predicted transcripts from these pooled alignments, using the existing annotations as a guide (-g option), with maximum intron length -I 50000, and all other parameters left at their default values. In parallel, Trinity v2.0.6 was used in genome-guided mode to assemble transcripts based on these same alignments, with --genome_guided_max_intron 50000 [4]. The FASTA file of assembled transcripts from Trinity and the GTF file of predicted transcripts from Cufflinks were fed into the

PASA pipeline [5], which used this information to predict novel transcripts and update the existing annotations. Three rounds of updating with PASA were used. We also obtained annotations from a recent *A. lyrata* annotation update, which used RNA-seq data from aerial vegetative tissues [6]. We predicted which of the three possible gene models (the original annotation, the PASA updates, and new annotations from Rawat *et al*. [6]) was best supported by our data at each locus. Cuffcompare was used to map the three sets of annotations to each other for each locus. The best supported model for each locus was defined according to the following: (1) if only one of the sets of annotations had an annotated model at a locus, it was chosen automatically, (2) if expression at the locus was very low, defined as (counts^2)/len(CDS) < 10 for all models, the original annotation was chosen, (3) if one model had more than 1.5x higher (counts^2)/len(CDS) than either other model, that model was chosen, and finally if none of the other criteria were satisfied, the model with the longest coding sequence was chosen. This was repeated for all loci to obtain the final set of annotations, representing 36,732 putative protein-coding genes used for all subsequent analysis. Of the 36,732 genes, 19,648 were unchanged from the original version, 8,842 genes had altered UTRs but unchanged coding sequences, while 6,323 genes had altered coding sequences after the update. Finally, 1,919 genes were not found in the original annotations, most of which (1,861) were instead obtained from the annotations by Rawat et al [6]. The remainders were novel genes identified by PASA and were not present in either the original annotations or Rawat *et al*. [6] and likely represent genes specifically expressed in the embryo or endosperm.

Updating gene homology information

Since the updated annotations included a number of novel genes and altered the coding sequences of some existing genes, we also updated a list of putative *A. thaliana* homologues of *A. lyrata* genes obtained from Phytozome [7]. To obtain preliminary new homology information, we performed a reciprocal tblastx of all *A. lyrata* genes (using the updated annotations from above) to the *A. thaliana* genome, and all *A. thaliana* genes to the *A. lyrata* genome. Only alignments with E-values below 0.0001 were reported. We first identified all pairs of *A. lyrata* and *A. thaliana* genes that were reciprocal hits to each other (defined as both alignments passing E-value cutoff, and additionally both alignments covering > 50% of the query gene). Then, for each *A. lyrata* gene ALY, we defined the likely homologue ATH: (1) if ALY and ATH are each other's reciprocal best hits; else if (2) ALY and ATH are each other's only reciprocal hit (even if not both highest scoring); else if (3) there are multiple pairs of reciprocal hits that include ALY, but ALY and ATH are the overall highest scoring according to E-value; else if (4) there are no pairs of reciprocal hits that include ALY, but either ALY has a high scoring alignment to ATH covering > 75% of ALY, or ATH has a high scoring alignment to ALY covering > 75% of ATH (this often occurs if one gene is a fragment of the other). We also kept the Phytozome homologue if the coding sequence of a gene remained unchanged in the new annotations, unless the new homology analysis revealed a different reciprocal best hit. In the final set of *A. thaliana* homologues for the 36,732 *A. lyrata* genes, 32,891 were unchanged from Phytozome (this includes cases where both versions had no identified *A. thaliana* homologue), 996 had a different homologue than before, 530 lost a homologue and 2315 gained a homologue.

Identifying SNPs

To identify SNPs between MN47 and Karhumäki (**Supplementary Table 7**) for allele-specific expression analysis, RNA-seq data from all datasets corresponding to either MN47 or Kar (**Supplementary Table 1**) were pooled into a single dataset containing all MN47-derived reads (the reference strain) and another containing all Kar-derived reads (the alternate strain). PCR duplicates were removed using the Picard tools MarkDuplicates function before pooling. Preliminary SNP info was obtained using SAMtools mpileup with default parameters. SNP calls were refined using vcf-annotate to exclude all sites with fewer than 20 overlapping reads, and with fewer than 10 reads with the alternate allele. SNPs where the MN47 allele was not consistent with the published genome were also removed. To filter out likely heterozygous SNPs, only SNPs with a PLDiff > 20 were kept, where PLDiff is the difference between the Phred-scaled genotype likelihoods (PL) of a homozygous call and a heterozygous call. The remaining 182,256 SNPs were used for all RNA-seq data analyses.

To identify additional non-genic SNPs for analysis of bisulfite-seq data we also sequenced Kar leaf genomic DNA (40 bp single end). Reads were trimmed using trim-galore version 0.4.0, with --stringency 3 and quality cutoff -q 25. Additionally, 8 bp of lower quality at the 5' end were trimmed off using --clip_R1 8. Reads were aligned to the MN47 reference genome using bowtie2 (v. 2.2.5), with –N 0 and –L 22. Reads with mapping quality (MAPQ) greater than 5 were kept, and presumed PCR duplicates were removed using MarkDuplicates from the picard-tools suite. The remaining 44,789,279 reads were used to call SNPs between Kar and MN, requiring at least 10 overlapping reads at the SNP position, at least 10 reads with the alternate allele, and a PLDiff > 20. This resulted in 381,796 SNPs. We also called SNPs after pooling the DNA-seq reads with the RNA-seq reads used previously, using the parameters depth >=20, alt allele >=10, PLDiff > 20, resulting in 190,527 SNPs. The union of these two sets of SNPs, after removing SNPs where the MN47 allele was not consistent with the published genome or which were no longer called after the DNA-seq data was added, resulted in 487,939 SNPs used for all analyses with bisulfite-sequencing data (**Supplementary Table 7**).

Differential expression analysis

Differential expression analysis of *A. lyrata* genes between samples was performed with DESeq2.0 [8]. Genes were considered significantly differentially expressed between the two conditions if abs($\log_2$foldchange) > 1 and the adjusted p-value was < 0.05. The regularized log (rlog) transformation from DESeq2 was used to normalize data before performing PCA.

Identifying imprinted genes

Imprinted genes were identified using our previously described method [9,10]. Briefly, for each library in a pair of reciprocal crosses (MN x Kar compared to Kar x MN), reads overlapping a known SNP were assigned to the parent of origin, and htseq-count [11] was used to count the number of reads from the MN47 allele and the Kar allele found in each gene. Fisher's exact test was used to test the null hypothesis that the proportion of maternal to paternal reads was 1:1 (embryo) or 2:1 (endosperm) in both directions of the cross. Genes were considered imprinted if they had a Benjamini-Hochberg corrected p-value less than 0.01, as well as a minimum imprinting factor of 2 and a maximum cis-effect factor of 10 [9]. In addition, endosperm MEGs were required to have at least 85% maternally-derived reads in each pair of reciprocal crosses, and endosperm PEGs were required to have less than 50% maternally-derived reads in each pair of reciprocal crosses. For embryo, these cutoffs were > 70% maternal (MEGs) and < 30% maternal (PEGs) [10]. This analysis was performed separately for all 12 possible comparisons between the 3 MN x Kar endosperm libraries and the 4 Kar x MN endosperm libraries, and for all 9 possible comparisons between the 3 MN x Kar and 3 Kar x MN embryo libraries. A gene was considered imprinted if at least 5 of the possible comparisons could be evaluated for imprinting (defined as both reciprocal crosses being compared having at least 10 total allele-specific counts), and at least 40% of reciprocal cross comparisons positively identified that gene as imprinted. Under these criteria, 12,633 genes could be evaluated for imprinting. Additionally, we used DEseq2 to identify and filter out 3,449 genes with an estimated $\log_2$ fold change of more than 1.5 in MN x MN seed coat compared to MN x MN endosperm, since these genes could appear as MEGs due to potential seed coat contamination of endosperm samples. Finally, genes with a homologous mitochondrial or chloroplast *A. thaliana* gene were excluded from the final list of imprinted genes.

Validation of allelic bias

Allelic bias was validated by pyrosequencing from MN x MN, Kar x Kar, MN x Kar and Kar x MN embryo and endosperm cDNA. Cloning and Sanger sequencing of genomic DNA confirmed SNPs between MN and Kar. For pyrosequencing, DNase-treated RNA was reverse transcribed using SuperScript III and an oligo(dT) primer (Invitrogen). cDNA was amplified with primers listed in **Supplementary Table 8** using the PyroMark PCR kit (Qiagen). The University of Michigan DNA Sequencing Core performed pyrosequencing with the indicated sequencing primer.

Calculating average % maternal reads

For **Figure 1D** and **Supplementary Figure 4**, the average endosperm % maternal reads in *A. lyrata* was obtained by averaging together the 3 MN x Kar libraries and the 4 Kar x MN libraries separately. These two values (average % maternal in MN x Kar and average % maternal in Kar x MN) were then averaged together to obtain the overall average. This avoids giving undue weight to the cross with more replicates. Since there were an equal number of replicates for the two reciprocal crosses for both *A. thaliana* comparisons (3 replicates each for Col x Cvi and Cvi x Col, same for Col-L*er*), the average was simply taken over all 6 samples. Data was from Pignatta *et al*. 2014 [10]. In **Figure 1d**, average % maternal for *A. thaliana* represents the average over the 3 Col x Cvi and 3 Cvi x Col endosperm replicates.

Differential expression analysis between *A. lyrata* and *A. thaliana*

We used DESeq2 to calculate the $\log_2$ fold change of expression in *A. lyrata* over *A. thaliana* (positive values indicate *A. lyrata* is more highly expressed than *A. thaliana*, negative values indicate the reverse) (see **Figure 1e**). Counts from all 13 *A. lyrata* endosperm samples were used, as were counts from all 6 Col-Cvi and 6 Col-L*er* endosperm libraries from Pignatta *et al.*

(2014)[10]. Estimated $\log_2$ (*A. lyrata/A. thaliana*) values were plotted in **Figure 1e** for all conserved and non-conserved MEGs and PEGs for which there were homologues.

Whole genome bisulfite sequencing data analysis

8 bases were trimmed from the 5' end of the forward and reverse reads, as recommended by the EpiGnome kit protocol. Reads were further trimmed for quality and adapter contamination using trim-galore, with a quality cutoff of -q 25 and --stringency 3. Only read pairs in which both reads were at least 32 bp after trimming were kept for mapping. For samples sequenced from MN47, Bismark was used to align the reads to the "bisulfite treated" *Arabidopsis lyrata* JGI v1.0 (MN47) genome, including the chloroplast and mitochondrial scaffolds [12]. To reduce mapping bias in favor of reads from the MN allele, samples sequenced from Kar x MN47 crosses were aligned to a bisulfite treated "metagenome" containing the full JGI v1.0 genome (MN47) and the Kar "pseudogenome", in which the 487,939 Kar SNPs were introduced into the MN47 sequence. For 40 x 40 bp libraries, 1 mismatch was allowed within the length of the 40 nt seed region, and for 100 x 100 bp libraries, 2 mismatches were allowed within the length the 80 nt seed region. After mapping, PCR duplicates were removed by a script provided with Bismark (deduplicate_bismark_alignment_output.pl), which randomly chooses which single representative of a pool of presumed PCR duplicates to keep in order to avoid biases in methylation calls. Per-site methylation information was extracted using the bismark_methylation_extractor and converted to BED-like format, and cytosines covered by at least 5 reads were kept for further analyses.

Identifying differentially methylated regions

The genome was divided into consecutive 300 bp windows that overlapped by 200 bp. Each window was assessed as a potential DMR between two samples (e.g. embryo and endosperm) using the method described in Pignatta *et al.* [10]. Briefly, the weighted average methylation in each window was computed for each library [13]. A window was required to contain at least 3 (CpG or CHG) or 10 (CHH) cytosines with at least 5 overlapping reads each in both samples in order to perform a comparison. The null hypothesis of no difference in methylation was tested using Fisher's exact test. All windows with a Benjamini-Hochberg corrected p-value < 0.01, and with a difference in average percent methylation of at least 35 (CpG and CHG) or 10 (CHH), were considered DMRs. For Kar x MN allele-specific DMRs (comparing the maternal allele of embryo to the maternal allele of endosperm, for example), all parameters were the same except we required an average % methylation difference of at least 20 (CpG and CHG) or 10 (CHH). Overlapping DMRs were merged together into single intervals using bedtools merge.

Plots of average methylation profiles across features

For each gene or TE, 50 bp windows were created beginning 2 kb 5' of the gene/TE start site, and ending 2 kb into genes and 1 kb into TEs. The same number of 50 bp windows were also created symmetrically around the gene/TE transcriptional start site (TSS), starting 2 kb 5' of the gene TSS or 1 kb 5' of the TE TSS, and ending 2 kb 3' of the gene/TE TSS. Neither set of windows was permitted to go beyond the midpoint of a particular gene/TE. Average methylation levels[13] were calculated for each window in each gene separately using the per-site methylation information, and then these values were averaged for each 50 bp window across all genes or TEs. For plots over imprinted genes, if a 50 bp window only had data for a single gene, that point was omitted from the plot.

Locus-specific bisulfite PCR

Loci of interest were identified based on DMR data. MN x MN and Kar x MN seeds at the walking stick/bent cotyledon stage were dissected into embryo and endosperm, DNA was extracted using a CTAB method, RNase-treated, and genomic DNA from multiple dissections was pooled. 214-500 ng of gDNA was bisulfite treated with the MethylCode Bisulfite Conversion kit (Invitrogen), according to manufacturer's instructions. PCR products were amplified using ExTaq for 40 cycles, with annealing temperatures varying from 50 – 55°C, and were cloned into either pJET1.2/blunt (Thermo Fisher) or pCR2.1-TOPO TA (Thermo Fisher) vectors. Primer sequences are in **Supplementary Table 8**. Individual bisulfite clones were Sanger sequenced and DNA methylation analysis was performed using CyMATE [14].

Testing the relationship between endosperm CHG methylation and gene expression

Of the 36,732 genes in our *A. lyrata* annotations, 23,809 had sufficient coverage in both MN x MN embryo and endosperm to evaluate differential expression using DESeq2 (all replicates of MN x MN embryo and endosperm RNA-seq data were used). 4057 genes were significantly less expressed in endosperm than embryo (defined as having a DESeq2 adjusted p-value < 0.05 and $\log_2$ fold change (endosperm/embryo) < -1). Of the 1606 genes that gained endosperm CHG methylation, 1225 had sufficient coverage to evaluate differential expression; 338 were significantly less expressed in endosperm than embryo. We evaluated whether an overlap of 338 genes between the 4057 genes less expressed in endosperm than embryo and the 1225 endosperm CHG methylated genes was significantly greater than expected by chance using a hypergeometric test in R:

```
phyper(337,4057,23809-4057,1225,lower.tail = FALSE, log.p = FALSE)
```

This is the probability of observing 338 or more genes shared between these two groups. We obtained $P(x \geq 338) = 1.766 \times 10^{-21}$, suggesting that genes that gain CHG methylation in endosperm are significantly more likely to be less expressed in endosperm compared to embryo than would be expected by chance. Similarly, we identified 4671 genes that were significantly more highly expressed in endosperm than in embryo (adjusted p-value < 0.05 and $\log_2$ fold change (endosperm/embryo) > 1), 159 of which also gained CHG methylation:

```
phyper(158,4671,23809-4671,1225,lower.tail = FALSE, log.p = FALSE)
```

We obtained $P(x \geq 159) \approx 1$, or equivalently $P(x \leq 159) = 2.04 \times 10^{-10}$ (phyper(159,4671,23809-4671,1225,lower.tail = TRUE, log.p = FALSE)), suggesting that there are significantly fewer genes more expressed in endosperm compared to embryo among the endosperm CHG hypermethylated genes than would be expected by chance.

RT-qPCR

RNA was isolated from dissected embryo and endosperm using the RNAqueous Micro Kit (Ambion) and treated with DNase I (Invitrogen). RNA was reverse transcribed with SuperScript III using an oligo(dT) primer (Invitrogen). qPCR was performed with Fast SYBR Green Master Mix (Applied Biosystems) using primers listed in **Supplementary Table 8**. Relative enrichment was calculated using the ΔΔCt method [15].

Supplementary References

1. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43,** 476–481 (2011).

2. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

3. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (2012).

4. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8,** 1494–1512 (2013).

5. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31,** 5654–5666 (2003).

6. Rawat, V. *et al.* Improving the Annotation of Arabidopsis lyrata Using RNA-Seq Data. *PLoS ONE* **10,** e0137391 (2015).

7. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40,** D1178–86 (2012).

8. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

9. Gehring, M., Missirian, V. & Henikoff, S. Genomic Analysis of Parent-of-Origin Allelic Expression in Arabidopsis thaliana Seeds. *PLoS ONE* **6,** e23687 (2011).

10. Pignatta, D. *et al.* Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *Elife* e03198 (2014).

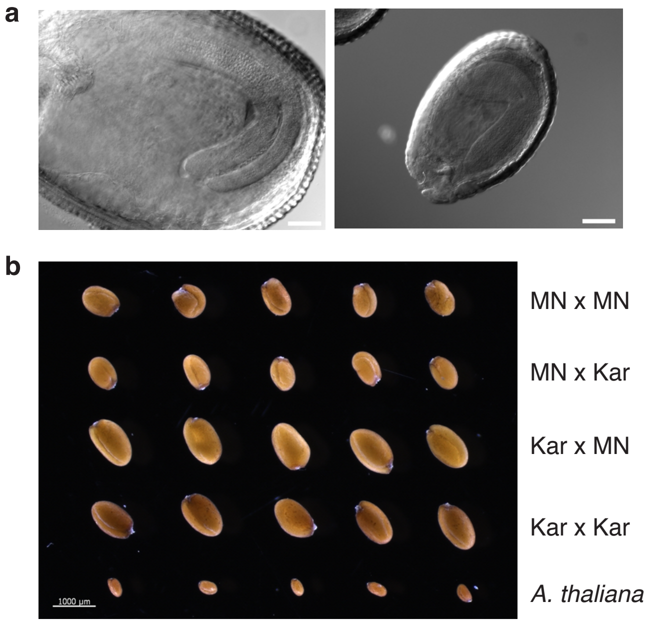
11. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

12. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27,** 1571–1572 (2011).

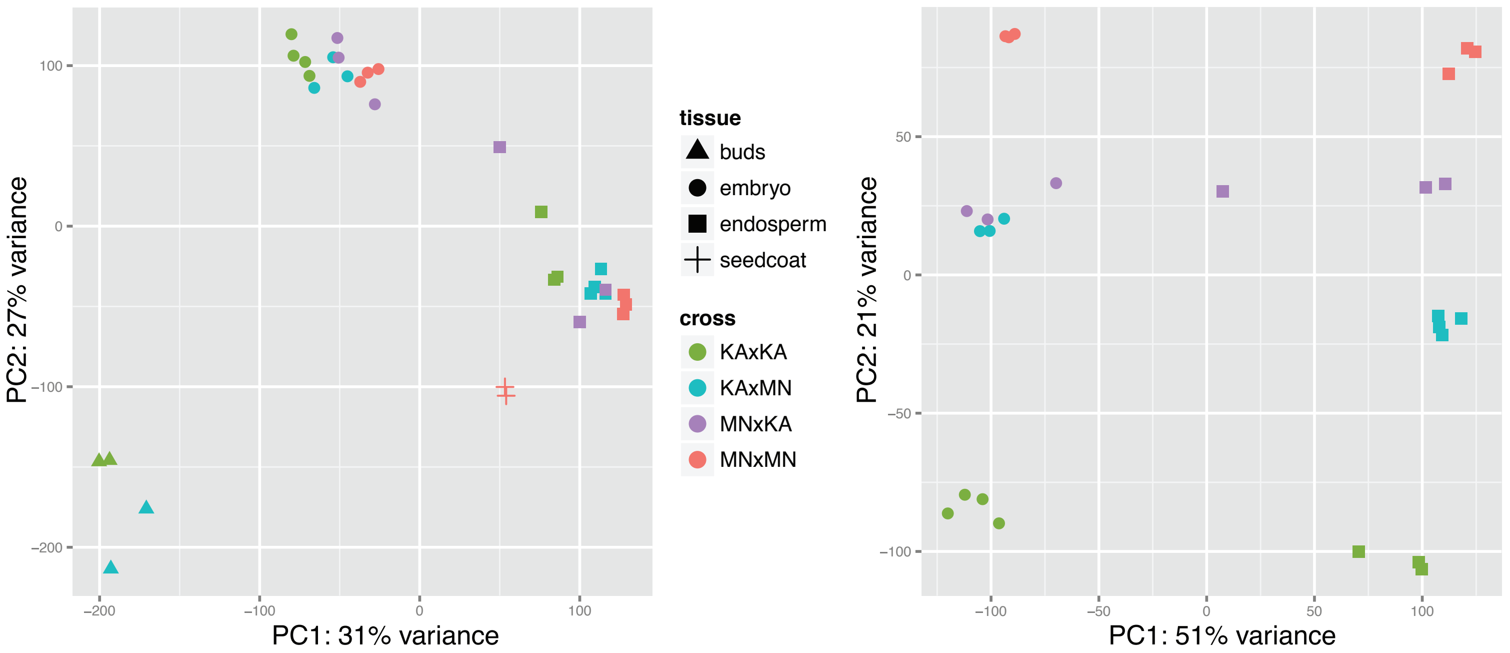
13. Schultz, M. D., Schmitz, R. J. & Ecker, J. R. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.* **28,** 583–585 (2012).

14. Hetzl, J., Foerster, A. M., Raidl, G. & Mittelsten Scheid, O. CyMATE: a new tool for methylation analysis of plant genomic DNA after bisulphite sequencing. *Plant J.* **51,** 526–536 (2007).

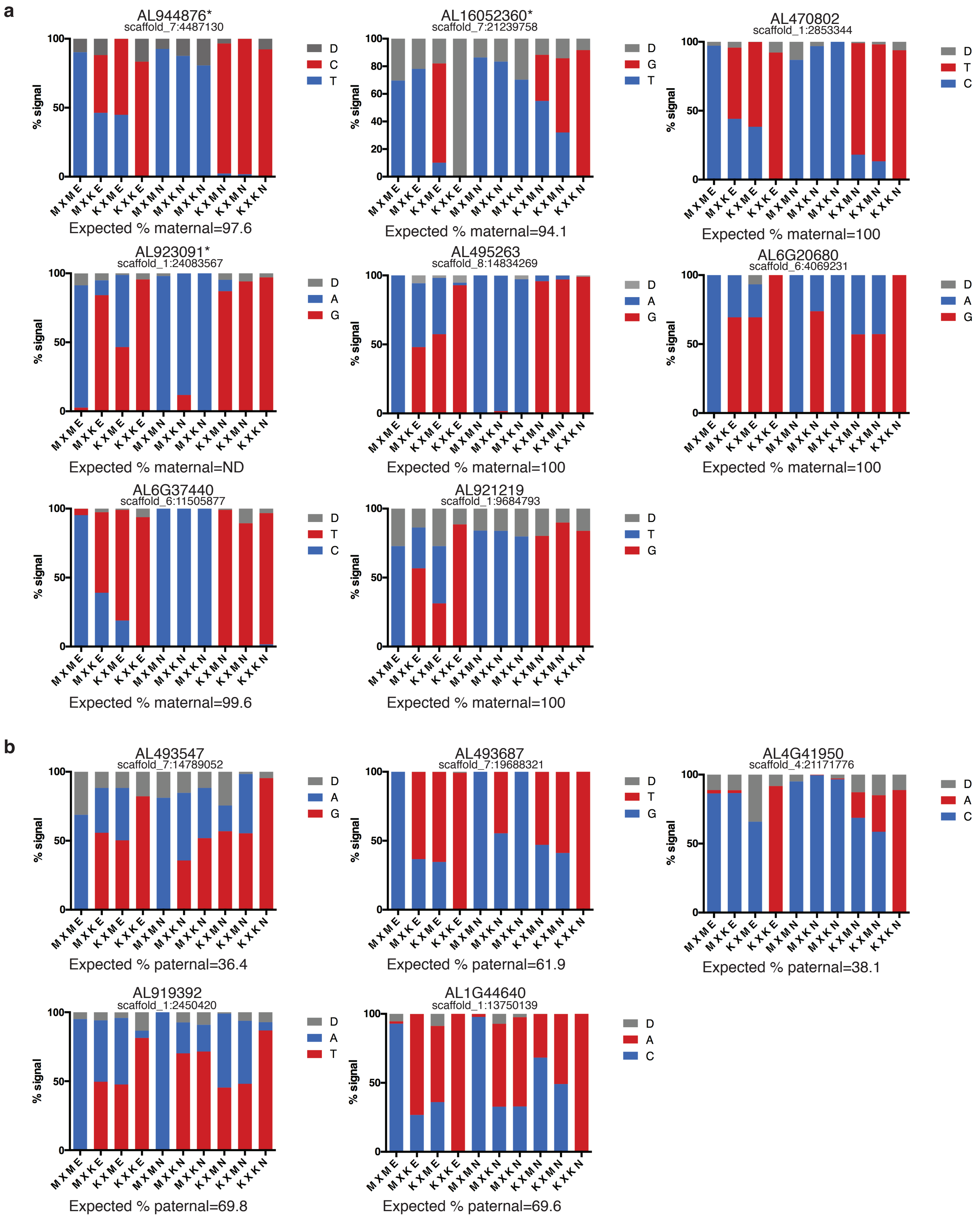
15. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25,** 402–408 (2001).



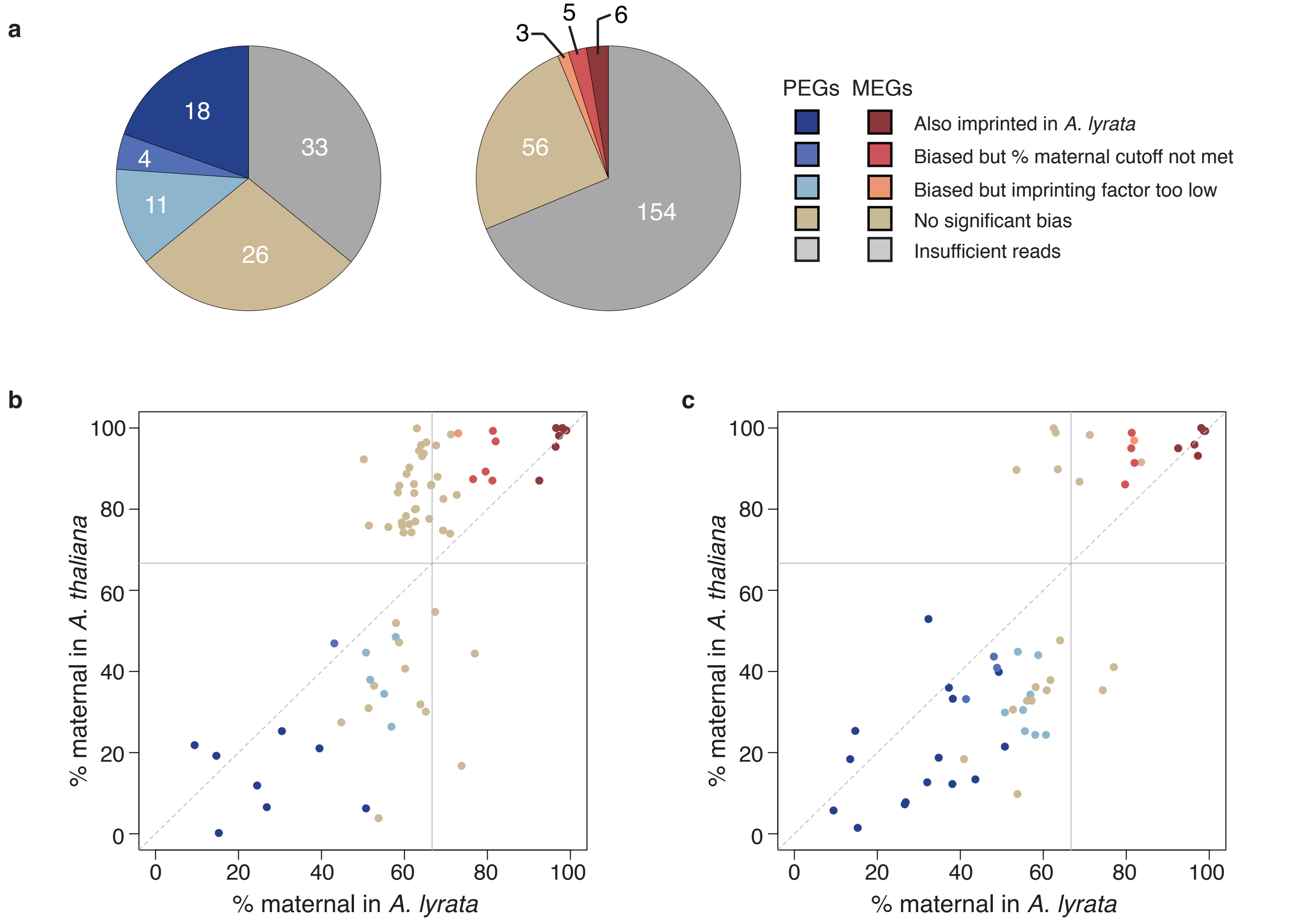
Supplementary Figure 1: *A. lyrata* seed structure. **a)** *A. lyrata* MN47 seed at the bent cotyledon stage, ~18 days after pollination (DAP) (left). *A. thaliana* Col-0 seed at the bent cotyledon stage, ~7 DAP (right). Scale bar is 100 *µ*m. **b)** Mature dry seed of *A. lyrata* parental strains, their hybrids, and *A. thaliana* Col-0. Scale bar is 1000 *µ*m.



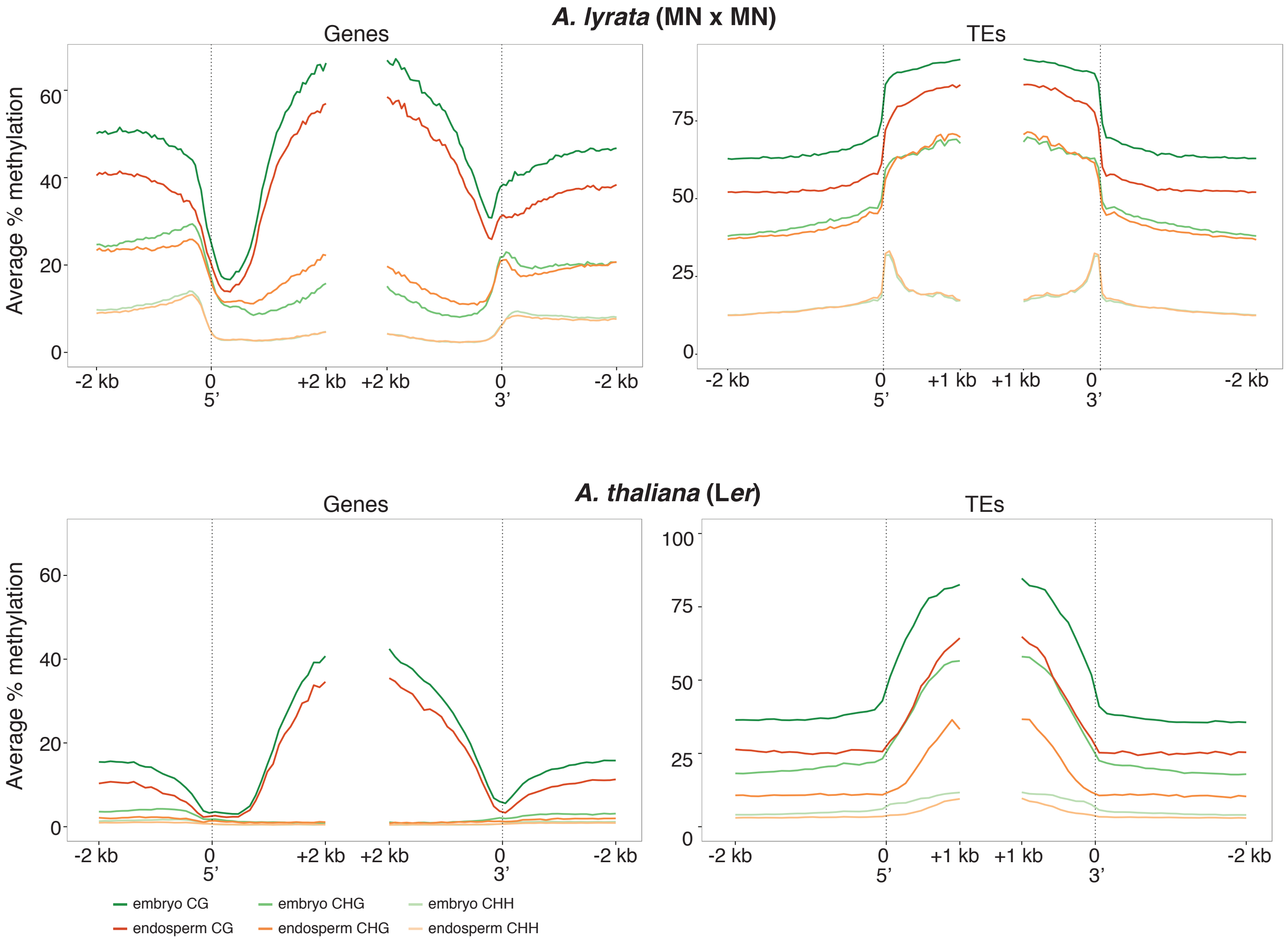
Supplementary Figure 2: Principal component analysis of *A. lyrata* gene expression. PCA of all mRNA-seq samples (left) or embryo and endosperm samples only (right). Female parent in the cross is listed first.



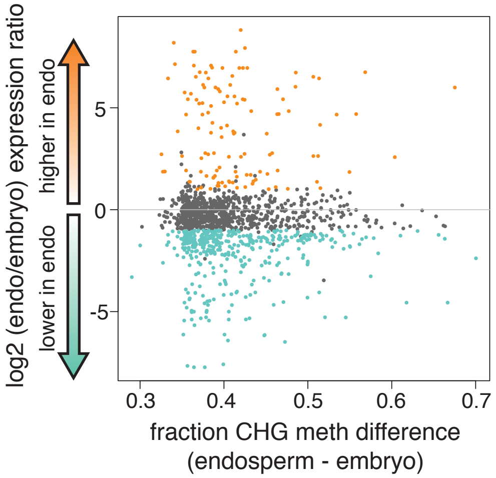
Supplementary Figure 3: Validation of mRNA-seq data by pyrosequencing. Validation of parent-of-origin results for selected SNPs with varying degrees of maternal (**a**) or paternal (**b**) bias. Signal for each SNP from parental or F1 embryo and endosperm is shown. Expected percent maternal or paternal expression was calculated using endosperm mRNA-seq data. Imprinted genes are marked with an asterisk. Blue, MN sequence; Red, Kar sequence; Gray, undetermined signal. M, MN; K, Kar; E, Embryo; N, Endosperm.



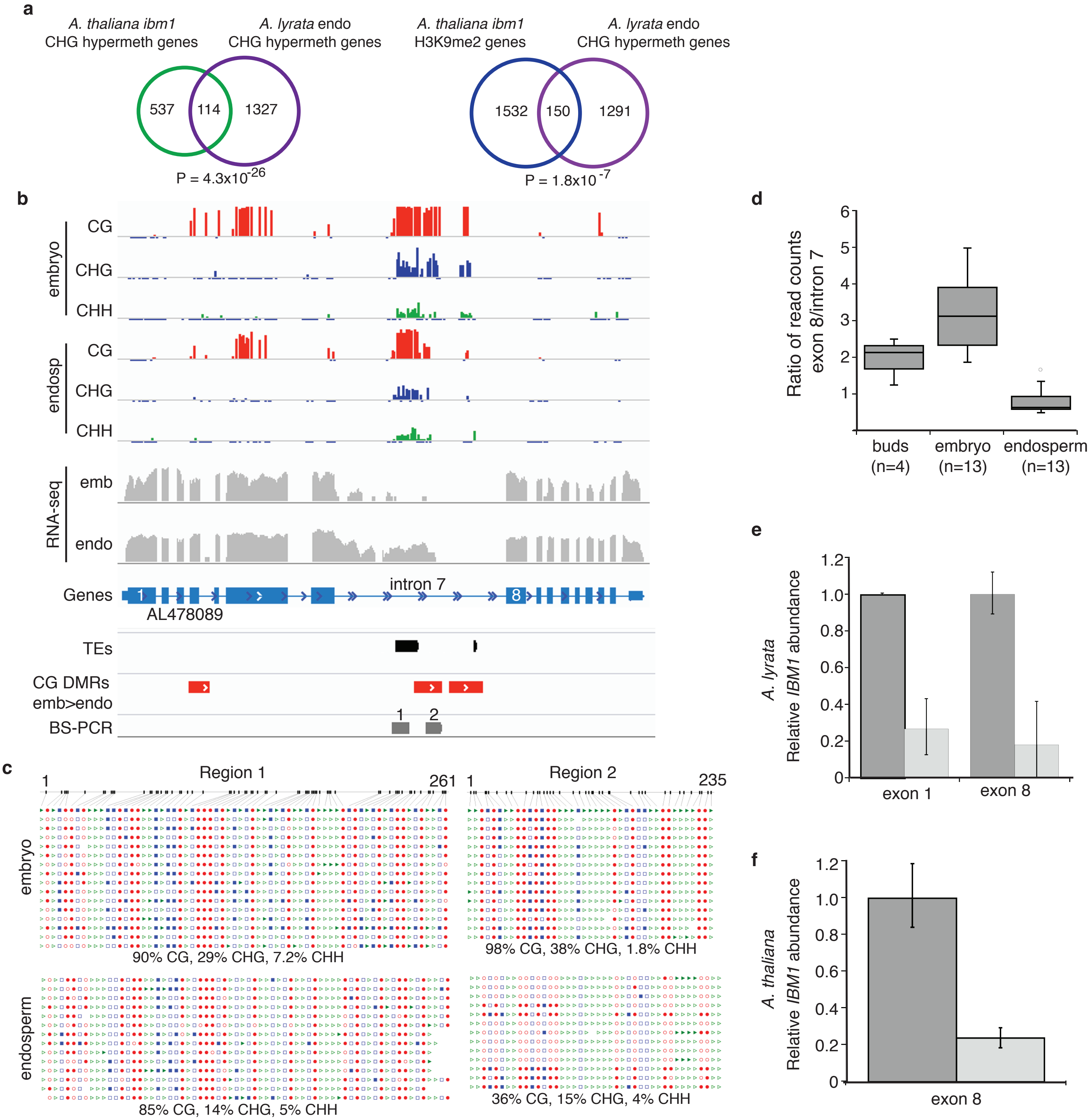
Supplementary Figure 4: Comparison of *A. thaliana* imprinted genes to *A. lyrata*. a) Conservation of imprinting between *A. thaliana* and *A. lyrata* endosperm. Each gene in the union set of *A. thaliana* imprinted genes in Pignatta *et al*. (2014) with an *A. lyrata* homologue was examined for imprinting or parental bias in *A. lyrata*. Reason for lack of imprinting in *A. lyrata* is given in the legend. **b)** Comparison of endosperm parental bias in *A. thaliana* and *A. lyrata* for all genes considered imprinted in Col-L*er* reciprocal crosses in *A. thaliana* from Pignatta *et al.* (2014). To obtain representative % maternal values for MN-Kar and Col-L*er*, reads for all replicates were pooled and % maternal was calculated using pooled counts for both reciprocal crosses in a pair separately. The average % maternal for the two reciprocal crosses was used as the final representative value for MN-Kar or Col-L*er*. Colors of points correspond to categories in (a). **c)** Same as (b), but % maternal for *A. thaliana* was calculated from Col-Cvi reciprocal cross data and plotted for all genes considered imprinted in Col-Cvi crosses in *A. thaliana* from Pignatta *et al.* (2014).



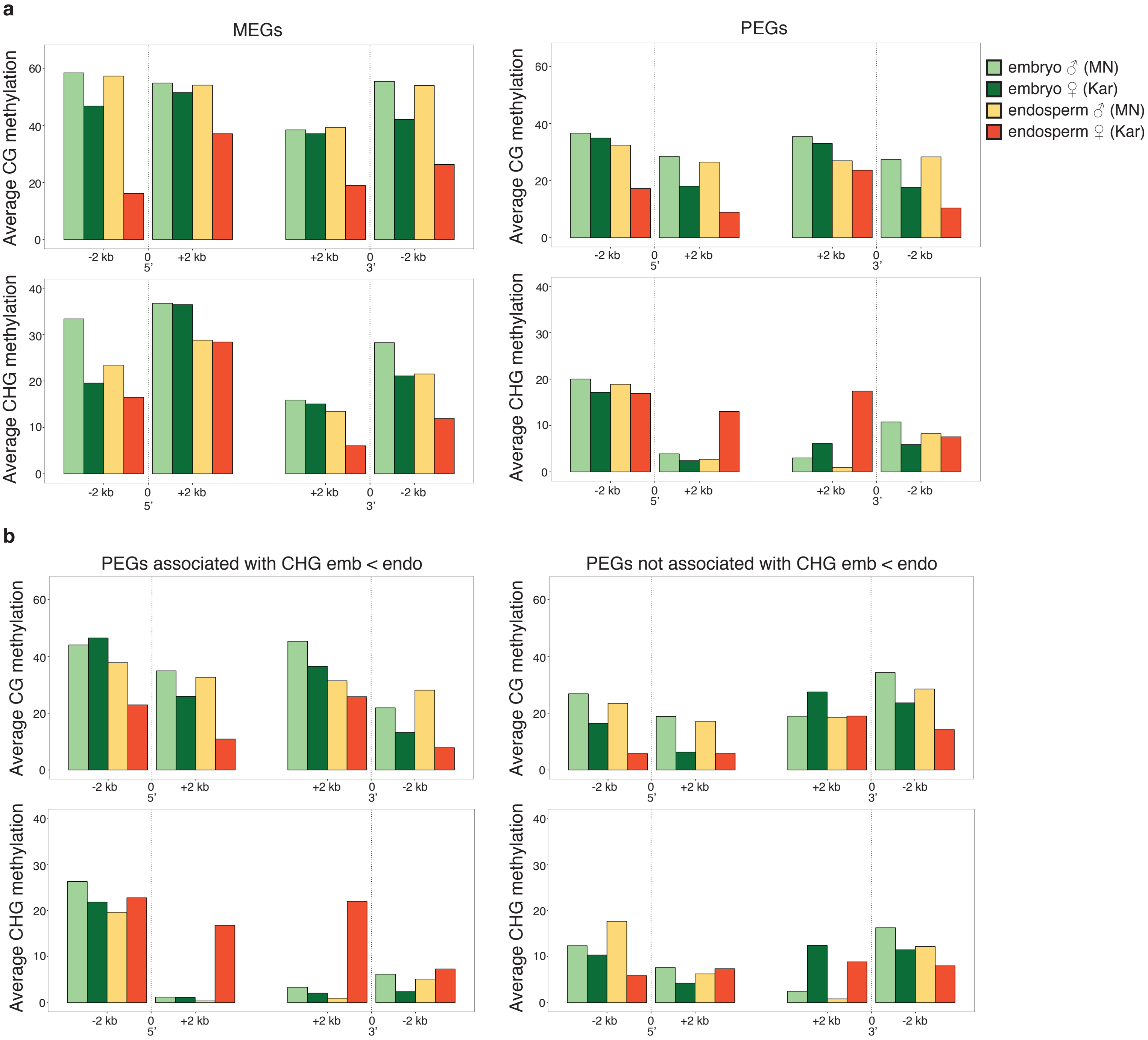
Supplementary Figure 5: Comparison of average methylation profiles in *A. lyrata* and *A. thaliana* seed tissues. Comparison of methylation over genes and TEs in *A. lyrata* (top) and *A. thaliana* (bottom) embryo and endosperm. Methylation data for *A. thaliana* were taken from Pignatta *et al.* 2014 for the L*er* strain. *A. thaliana* has lower levels of methylation in all contexts, and *A. thaliana* gene bodies do not gain CHG methylation in endosperm, unlike what is observed in MN x MN *A. lyrata*.



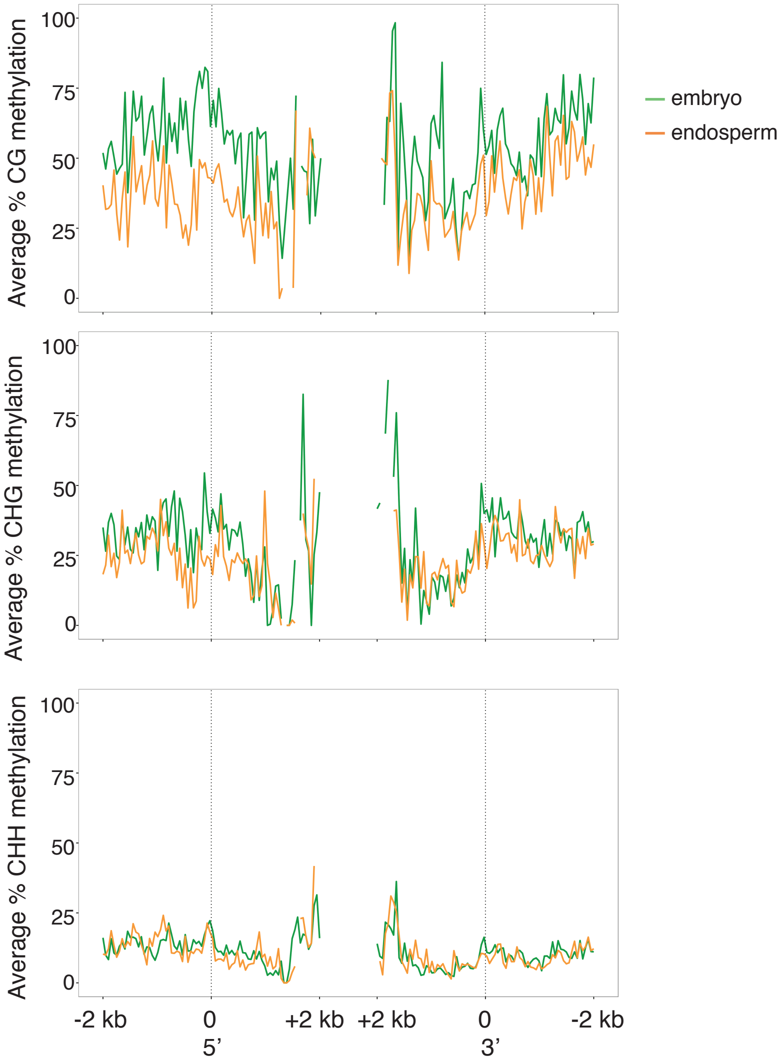
Supplementary Figure 6: Expression bias of genes that gain CHG methylation in endosperm relative to embryo. Scatterplot showing the average MN x MN endosperm *vs*. embryo gene expression ratio, calculated using DESeq2, for genes that overlap an endosperm CHG hypermethylated DMR. X-axis is the CHG methylation difference between endosperm and embryo for each gene. Genes with signficantly different expression are highlighted. Orange; higher expression in endosperm; cyan, lower expression in endosperm.



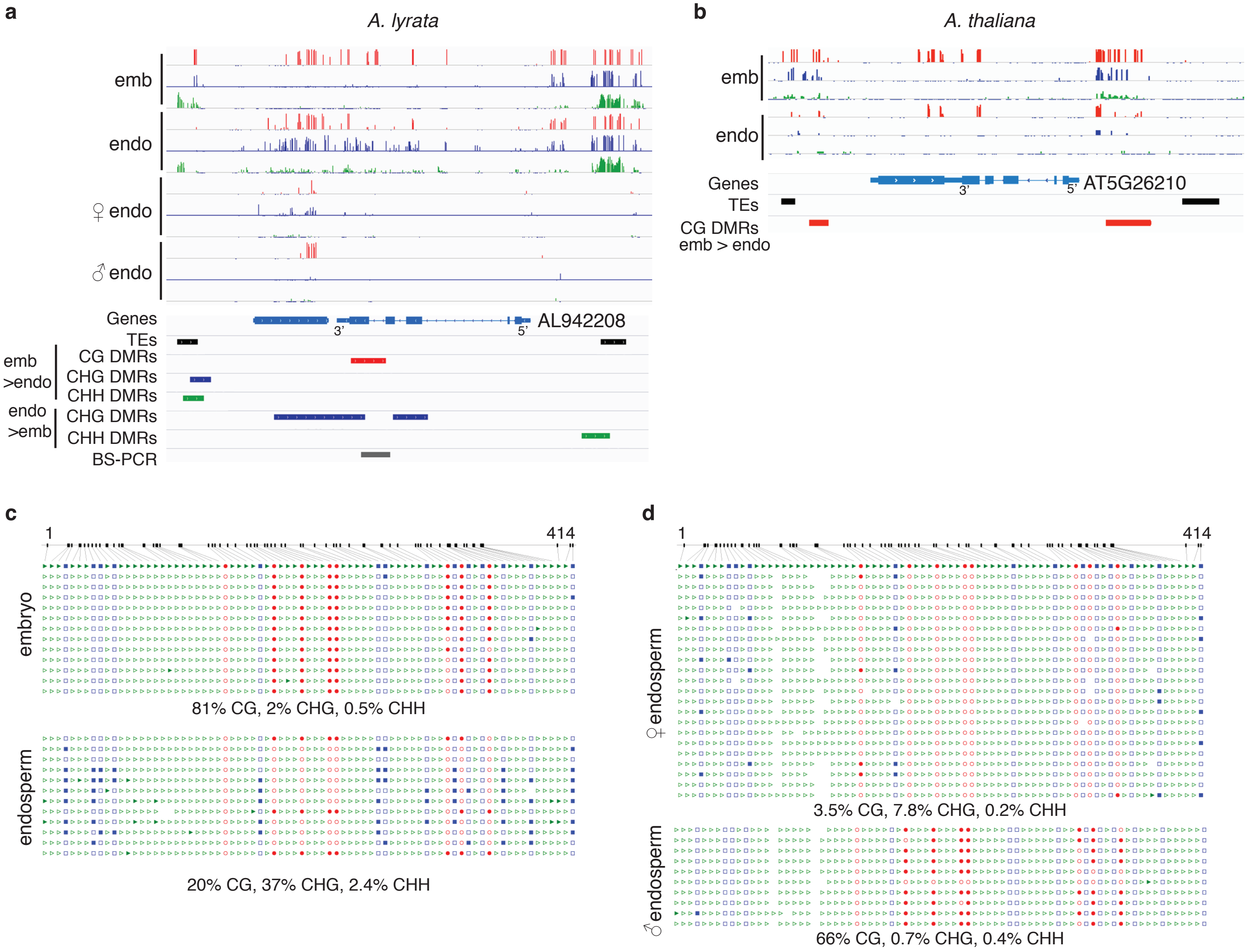
Supplementary Figure 7: Reduced *IBM1* mRNA in *A. lyrata* endosperm is correlated with reduced intronic methylation. **a)** Overlap between *A. thaliana* orthologues of *A. lyrata* endosperm CHG hypermethylated genes and genes identified in Miura *et al*. (2009) as strongly DNA hypermethylated in *ibm1* (left) and genes identified as gaining H3K9me2 in *ibm1* from Deleris *et al*. 2012 (right). P values for significance of overlap determined using the hypergeometric test. **b)** BS-seq and mRNA-seq of the MN x MN *A. lyrata IBM1* locus. RNA accumulates in intron 7 in the endosperm, correlated with reduced intronic methylation in that tissue. **c)** Locus-specific bisulfite PCR validation of embryo-endosperm CG DMR (region 2) in *IBM1* intron and lack of a DMR in region 1. **d)** Box plot of read count ratios in *IBM1* exon 8 (long transcript form) and intron 7 (short transcript form) in *A. lyrata* buds, embryo, and endosperm mRNA-seq datasets. **e)** Abundance of *IBM1* transcripts relative to actin in *A. lyrata* embryo (dark gray) and endosperm (light gray) determined by RT-qPCR. Data is from 2 biological replicates. Error bars show standard deviation. **f)** Abundance of *IBM1* transcripts relative to AT1G58050 in *A. thaliana* embryo (dark gray) and endosperm (light gray). Data is from 2 biological replicates of Col torpedo stage seeds. Error bars show standard deviation.



Supplementary Figure 8: Average methylation levels within and around imprinted genes. a) Average % methylation levels on the maternal (Kar) and paternal (MN) alleles of *A. lyrata* embryo and endosperm in the CG and CHG contexts across all MEGs and PEGs. Average methylation values are calculated over 4 separate regions (2 kb flanking the TSS and TTS). **b)** Average % methylation calculated over the 27 PEGs associated with endosperm CHG hypermethylation (left) and the 22 PEGs not associated with endosperm CHG hypermethylation (right).



Supplementary Figure 9: Methylation patterns of MEGs. Average methylation profiles in the CG, CHG and CHH contexts in MN x MN embryo and endosperm for MEGs aligned at their 5’ and 3’ ends.



Supplementary Figure 10: Distinct methylation profiles at the conserved PEG AT5G26210. a) *A. lyrata* bisulfite-seq methylation profile at AL942208, a conserved PEG homologous to AT5G26210. Methylation data from MN x MN embryo and endosperm are shown, along with allele-specific endosperm methylation data from Kar x MN. All MN x MN embryo-endosperm DMRs are indicated. Red tracks, CG methylation; blue, CHG methylation; green, CHH methylation. Tick marks below the line indicate Cs with sufficient coverage but no methylation. All tracks set at 100%. **b)** Methylation of the corresponding region in *A. thaliana*. Data is from L*er* x L*er*, published in Pignatta *et al.* (2014). **c)** BS-PCR validation of a CG hypomethylated and CHG hypermethylated DMR indicated in (a) in MN x MN embryo and endosperm. **d)** Allele-specific BS-PCR of the same region from Kar x MN endosperm.