Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population

Robert J. Schmitz,[1,2,11] Yupeng He,[2,3,11] Oswaldo Valdés-López,[4,12] Saad M. Khan,[4,5] Trupti Joshi,[4,5,6,7] Mark A. Urich,[2] Joseph R. Nery,[2] Brian Diers,[8] Dong Xu,[4,5,6,7] Gary Stacey,[4,7,9,13] and Joseph R. Ecker[1,2,10,13]

[1]Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA; [2]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA; [3]Bioinformatics Program, University of California at San Diego, La Jolla, California 92093, USA; [4]Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA; [5]Informatics Institute, University of Missouri, Columbia, Missouri 65211, USA; [6]Department of Computer Science, University of Missouri, Columbia, Missouri 65211, USA; [7]National Center for Soybean Biotechnology, University of Missouri, Columbia, Missouri 65211, USA; [8]Department of Crop Sciences, University of Illinois, Urbana, Illinois 61801, USA; [9]Divisions of Plant Science and Biochemistry, University of Missouri, Columbia, Missouri 65211, USA; [10]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, California 92037, USA

Cytosine DNA methylation is one avenue for passing information through cell divisions. Here, we present epigenomic analyses of soybean recombinant inbred lines (RILs) and their parents. Identification of differentially methylated regions (DMRs) revealed that DMRs mostly cosegregated with the genotype from which they were derived, but examples of the uncoupling of genotype and epigenotype were identified. Linkage mapping of methylation states assessed from whole-genome bisulfite sequencing of 83 RILs uncovered widespread evidence for local methylQTL. This epigenomics approach provides a comprehensive study of the patterns and heritability of methylation variants in a complex genetic population over multiple generations, paving the way for understanding how methylation variants contribute to phenotypic variation.

[Supplemental material is available for this article.]

Phenotypic variation results from a combination of genetic variation, environment, and interactions among the two. The contribution of natural epigenetic variation to phenotypic variation still remains enigmatic due to the relatively few characterized natural epigenetic alleles (epialleles) (Bender and Fink 1995; Cubas et al. 1999; Manning et al. 2006; Rangwala et al. 2006; Hitchins et al. 2007; Woo et al. 2007; Becker et al. 2011; Schmitz et al. 2011; Durand et al. 2012). Epialleles are classified into three major groups, which are defined by their dependence on an underlying genetic variant (Richards 2006). Briefly, obligate epialleles are completely dependent on a genetic variant, whereas pure epialleles are maintained independently of genetic variants. The dependence on genetic variants for the third group, facilitated epialleles, breaks down because the genetic variant can influence the epiallelic state but not as reliably as they do for obligate epialleles (Richards 2006).

In *Arabidopsis thaliana,* there is extensive evidence for the involvement of epialleles in creating phenotypic diversity (Johannes et al. 2009; Reinders et al. 2009; Roux et al. 2011). Outside of *Arabidopsis thaliana,* these are most evident for the *peloric, colorless non-ripening*, and *B′* epialleles from *Linaria vulgaris, Solanum lycopersicum*, and *Zea mays,* respectively (Patterson et al. 1993; Cubas et al. 1999; Manning et al. 2006). Still, these are rare events and appear to be the exception rather than the rule. Work in *Arabidopsis thaliana* has led to the most comprehensive analyses of natural epigenetic variation and uncovered a variety of modes to the formation of epialleles (Schmitz and Ecker 2012). These include genetic variants that can exert their influence on epiallelic states both locally and distantly to other chromosomes (Bender and Fink 1995; Rangwala et al. 2006; Woo et al. 2007; Durand et al. 2012; Schmitz et al. 2013).

The RNA-directed DNA methylation pathway (RdDM) (for review, see Law and Jacobsen 2010) provides a molecular basis for the formation and maintenance of epiallelic states of many of the identified epialleles in *Arabidopsis* and likely other flowering plant species. This pathway generates a feedback loop between small RNAs (smRNAs) and DNA methylation that represses gene expression and enables propagation of epiallelic states through both mitotic and meiotic cell divisions. The presence of smRNAs also provides sequence-specific guides that facilitate silencing at distant loci, even on different chromosomes.

Because most characterized epialleles contain distinct molecular signatures, usually smRNAs in combination with DNA methylation, it is possible to systematically determine how extensive natural epigenomic variation is in the wild. Pioneering efforts using epigenomic techniques (for review, see Schmitz and Zhang 2011) revealed extensive natural variation in methylation of gene bodies compared to smRNA-associated transposon and repetitive sequences between two accessions of *Arabidopsis thaliana* (Vaughn et al. 2007; Zhang et al. 2008). Similar epigenomic approaches in maize uncovered hundreds of differentially methylated regions (DMRs), some of which were subsequently found unlinked to genetic variants using near isogenic lines derived from the two profiled

[11]These authors contributed equally to this work.
[12]Present address: Laboratorio de Bioquímica, Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México, Tlalnepantla, Estado De México, Mexico 54090.
[13]Corresponding authors
E-mail ecker@salk.edu
E-mail staceyg@missouri.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.152538.112.

23:1663–1674 © 2013, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

parental lines revealing the presence of heritable pure epialleles (Eichten et al. 2011).

A major challenge in understanding natural epigenetic variation is determining the dependence of methylation variants on genetic variants. Recent studies addressed one aspect of this challenge by using a population of mutation accumulation lines (Shaw et al. 2000), which reduced genetic variation to the spontaneous mutation rate (Ossowski et al. 2010), enabling a better understanding of pure epigenetic variation. These studies uncovered single methylation polymorphisms (SMPs) occurring at a much higher rate than DNA mutations and found that they primarily occurred in gene bodies (Becker et al. 2011; Schmitz et al. 2011). Larger regions of differential DNA methylation that resembled loci targeted by RdDM were also identified and some were even found to affect gene expression levels, although the rate of occurrence of these types of DMRs was similar to the spontaneous DNA mutation rate (Becker et al. 2011; Schmitz et al. 2011). Therefore, it is clear that natural epigenetic variation can be uncoupled from genetic variation in the laboratory, but in nature, these two types of variants coevolve.

Soybean (*Glycine max* L. merr.) is a major crop providing an important source of protein and oil. A high-quality reference soybean genome is available (Schmutz et al. 2010), which supports that this plant has experienced at least two polyploid events, the most ancient being 59 Mya. Soybean is considered an allopolyploid (Gill et al. 2009), which resulted from the merger of two genomes that diverged ∼13 Mya and reunited ∼5–10 Mya when the genus *Glycine* was formed (Doyle et al. 2003; Straub et al. 2006; Innes et al. 2008; Stefanovic et al. 2009). Roughly 75% of all soybean-coding sequences are present in two or more copies in the genome. Therefore, to understand the role of DNA methylation in this species and its impact on gene expression, we sequenced genomes, DNA methylomes, and transcriptomes in the parents and RILs. This also enabled us to understand how DNA methylation patterns are established, inherited, and maintained as they segregate through a complex genetic population. The vast majority of identified DMRs cosegregated with the genetic background from which they were identified, which enabled population-wide identification of methylQTL for >90% of the DMRs. Rare examples of DMRs were identified that did not show evidence for linkage to a particular genomic region, which could be indicative of pure epigenetic variants.

The findings of this study have broad implications for the fields of crop epigenomics, epigenetics, inheritance of methylation variants, and plant breeding. There is a growing interest about the potential role for epigenetics to explain phenotypic diversity that cannot be attributed to genetics in a variety of systems, but the evidence is still limiting at the population level. This study clearly demonstrates that the majority of methylation variants adheres to Mendelian modes of inheritance but also demonstrates rare examples of epigenetic variation that do not follow the standard laws of inheritance.

Results

Single-base resolution DNA methylome of *Glycine max*

To understand the contribution of cytosine DNA methylation to the soybean genome, whole-genome bisulfite sequencing (MethylC-seq) (Lister et al. 2008) was performed on DNA isolated from leaves of the LD00-2817P germplasm (hereafter referred to as "LD"). In total, greater than 162 million 101-bp reads were sequenced that only aligned to unique regions of the genome, which represents approximately eightfold coverage per strand of the genome (Supplemental Table 1). Briefly, methylated cytosines were determined by applying a binomial test to data from reads covering each cytosine and using the unmethylated chloroplast genome as a control (see "Methods" for a more detailed description). The LD methylome contains 15,444,227 methylated CGs (mCG) (51% of all CGs), 14,942,676 mCHG (39% of all CHGs), and 13,628,219 mCHH (0.05% of all CHHs), which represents a greater proportion of methylated cytosines compared with a recently reported DNA methylome for *Arabidopsis thaliana* (Fig. 1A; Schmitz et al. 2011). Of the cytosines that are methylated in the LD genome, there are almost equal numbers of mCG and mCHG, which contrasts to the *Arabidopsis thaliana* methylome (Fig. 1B) and could indicate that RdDM targets a greater proportion of the soybean genome. Of the detected methylcytosines, the distribution of methylation levels at each site in each context was similar to the levels found in *Arabidopsis thaliana*, with the exception of mCHG (Supplemental Fig. 1A,B). In general, mCG and mCHG are methylated at higher levels as compared to mCHH.

The distribution of mCG, mCHG, and mCHH sites genome-wide revealed that gene-rich and transposon-poor euchromatic regions of each chromosome contain lower bulk methylation compared with the gene-poor and transposon-rich heterochromatic regions in the pericentromeres of the chromosomes (Fig. 1C; Supplemental Fig. 2A–F). Using previously published small RNA sequencing data (Tuteja et al. 2009), the relative abundance of 21–24 nucleotide (nt) smRNAs were plotted along each chromosome, which revealed a higher density of 24-nt smRNAs in regions of the genome that contain abundant non-GC methylation (Fig. 1C; Supplemental Fig. 2G), as well as for 21–23-nt smRNAs (Supplemental Fig. 3).

Two whole genome duplications have occurred in the diploid ancestor of soybean, an early duplication ∼59 million years ago (Mya) and a recent duplication ∼13 Mya. (Schmutz et al. 2010). A comparison of the DNA methylation profiles between these duplicated regions revealed that younger sequences are more likely to contain greater amounts of DNA methylation typical of the RdDM pathway (Fig. 1C,D), indicating that these sequences are actively being silenced. In fact, CG, CHG, and CHH methylation was, on average, ∼10%, ∼20%, and ∼10% higher for recently duplicated regions, genes, or exons compared to early duplicated regions, respectively (Fig. 1D). The increase in DNA methylation of the recent duplications also was significantly differentially associated with distance from the pericentromeric regions when compared to the early duplication events (Mann-Whitney-Wilcoxon test, *P*-value < 2.2×10^{-16}) (Fig. 1E,F; Supplemental Fig. 4). Collectively, these results indicate that DNA methylation is one potential mechanism that plants use to cope with duplicated DNA and could potentially explain, in part, why the soybean genome contains greater amounts of DNA methylation compared to *Arabidopsis thaliana*.

Patterns of DNA methylation in genes and transposons

CG gene-body methylation appears to be conserved in plants and animals (Feng et al. 2010; Zemach et al. 2010b); and soybean is no exception (Fig. 2A), although its exact function is still unknown and not all genes in plant genomes contain CG gene-body methylation. The density of both mCHG and mCHH is lowest throughout the gene body when compared to transposons (Fig. 2A; Supplemental Fig. 5A–C), which is consistent with the lack of 24-nt smRNA-directed DNA methylation targeting most genes (Fig. 2B). The density of all three types of DNA methylation is higher at

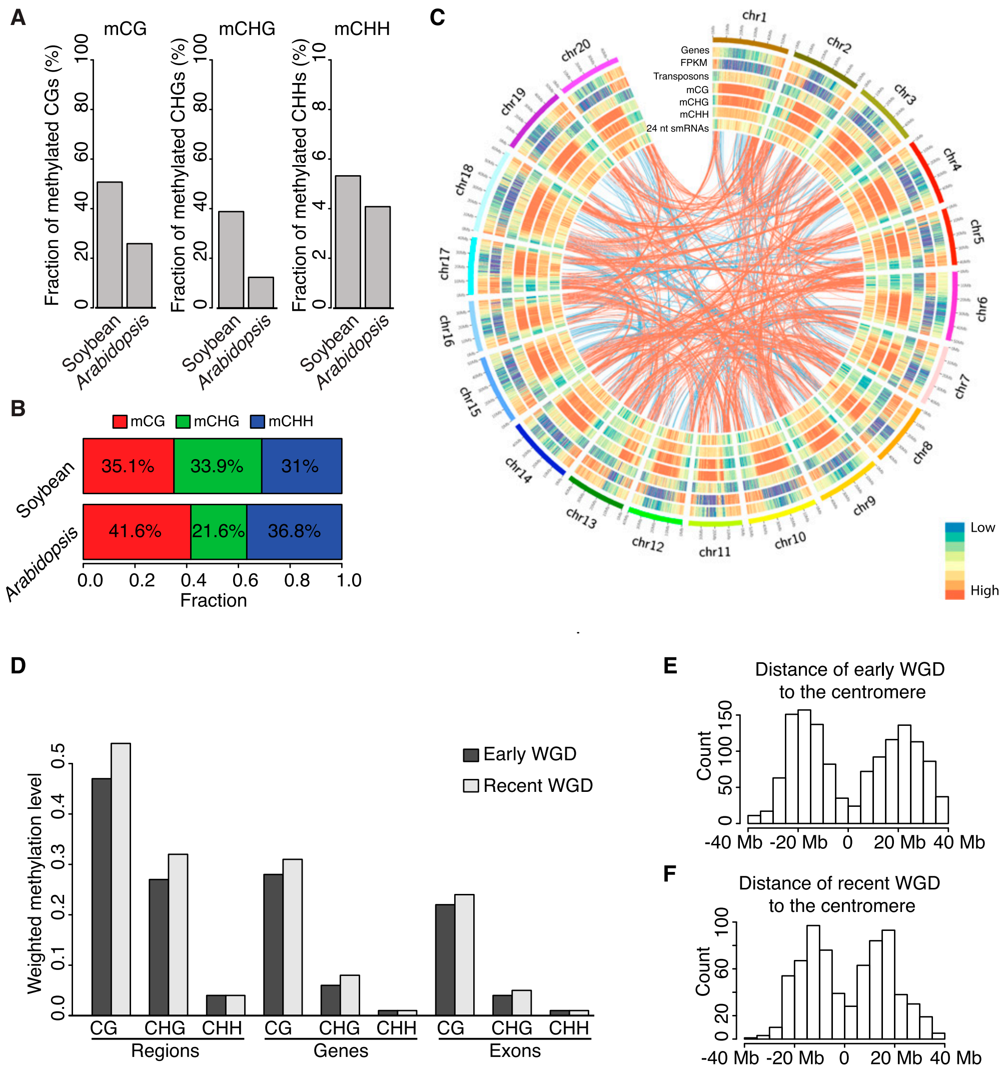

Figure 1. Characteristic features of the DNA methylomes between *Arabidopsis thaliana* and soybean. (*A*) Fraction of methylated cytosines for each context as a proportion of that context genome wide. (*B*) The DNA methylation present in the soybean methylome is highly enriched for CHG and CHH methylation. (*C*) A circle plot of gene density, transposon density, FPKMs, mCG, mCHG, mCHH, and 24-nt smRNAs for LD. (Red lines) Regions from the 59 Mya whole-genome duplication. (Blue lines) Regions from the recent whole-genome duplication 13 Mya. (*D*) Weighted methylation levels for early and recent whole-genome duplications (WGD) for the entire duplicated regions and genes and exons within those duplicated regions. (*E*,*F*) Distance in Mb of duplicated regions from the centromeres.

sequences both upstream and downstream from the transcriptional start and stop sites increases (Fig. 2A; Supplemental Fig. 5D–F).

In contrast to CG gene-body methylation, transposons are targeted by the RdDM pathway, resulting in enriched levels of all types of DNA methylation and an abundance of 24-nt smRNAs (Fig. 2A,B; Supplemental Fig. 5G). Although the levels of all types of DNA methylation between the major classes of soybean retrotransposons (long terminal repeat [LTRs] and LINEs) and DNA transposons (terminal inverted repeats [TIRs] and helitrons) are similar, there are interesting characteristics that distinguish them from one another. The LTR and LINE retrotransposons contain the highest levels of DNA methylation near the actual repeat structures that define the 5′ and 3′ ends of retrotransposons (Fig. 2A), which coincides with the location containing the greatest abundance of 24-nt smRNAs (Fig. 2B; Supplemental Fig. 6A,B). This contrasts to the DNA methylation levels at the 5′ and 3′ ends of TIR transposons, which are depleted relative to the transposon bodies where 24-nt smRNAs are most abundant (Fig. 2B; Supplemental Fig. 6C). The higher density of methylation upstream and downstream from LTRs also distinguishes them from the other classes of transposons (Fig. 2A), which reflects their distribution along the chromosomes (Fig. 2C). LTR transposons are located in the heterochromatic pericentromeres, whereas TIRs and LINEs are located throughout the chromosome arms (Fig. 2C).

Effects of DNA methylation on gene expression

To better understand the role of DNA methylation and its association with gene expression, RNA was isolated from LD leaves and RNA-seq data were generated, aligned, and quantified (Methods; Supplemental Table 2). The levels of CG gene-body methylation were positively correlated with gene expression levels (Fig. 3A), whereas the levels of CHG or CHH methylation in gene bodies were negatively associated with gene expression levels (Fig. 3B,C),

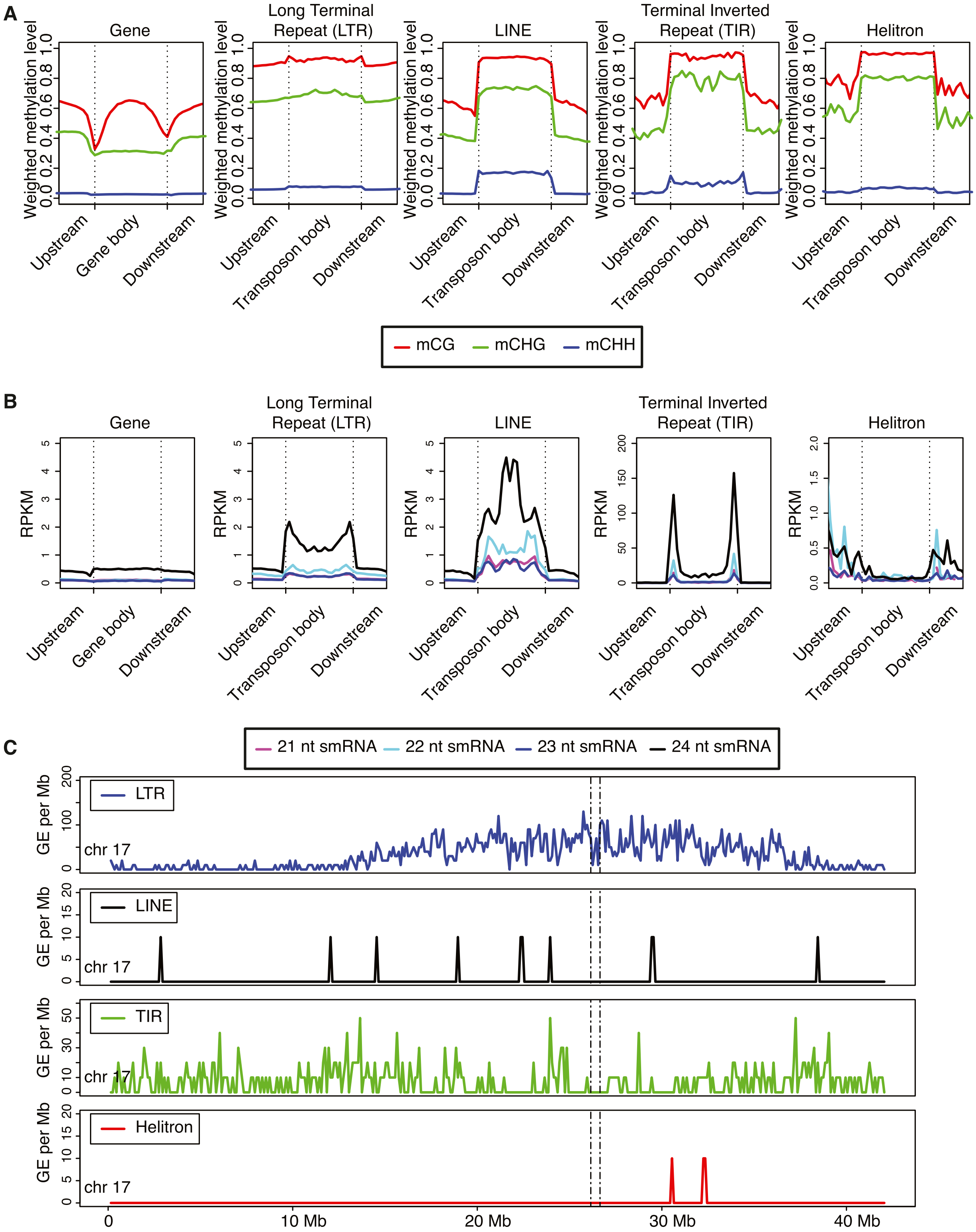

Figure 2. Characterization of gene body and transposon DNA methylation in soybean. (*A*) The distribution of mCG, mCHG, and mCHH densities and (*B*) the distribution of 21–24 nt smRNA levels in gene bodies, Long terminal repeats (LTRs) retrotransposons, LINE retrotransposons, terminal inverted repeat (TIR) DNA transposons, and Helitron DNA transposons including ±4 kb from the start and stop codons. (*C*) Chromosome-wide density of LTR, LINE, TIR, and Helitron transposons. Only data points between 1% and 99% quintiles were used to generate values for each bin in the plots presented in *A* and *B*.

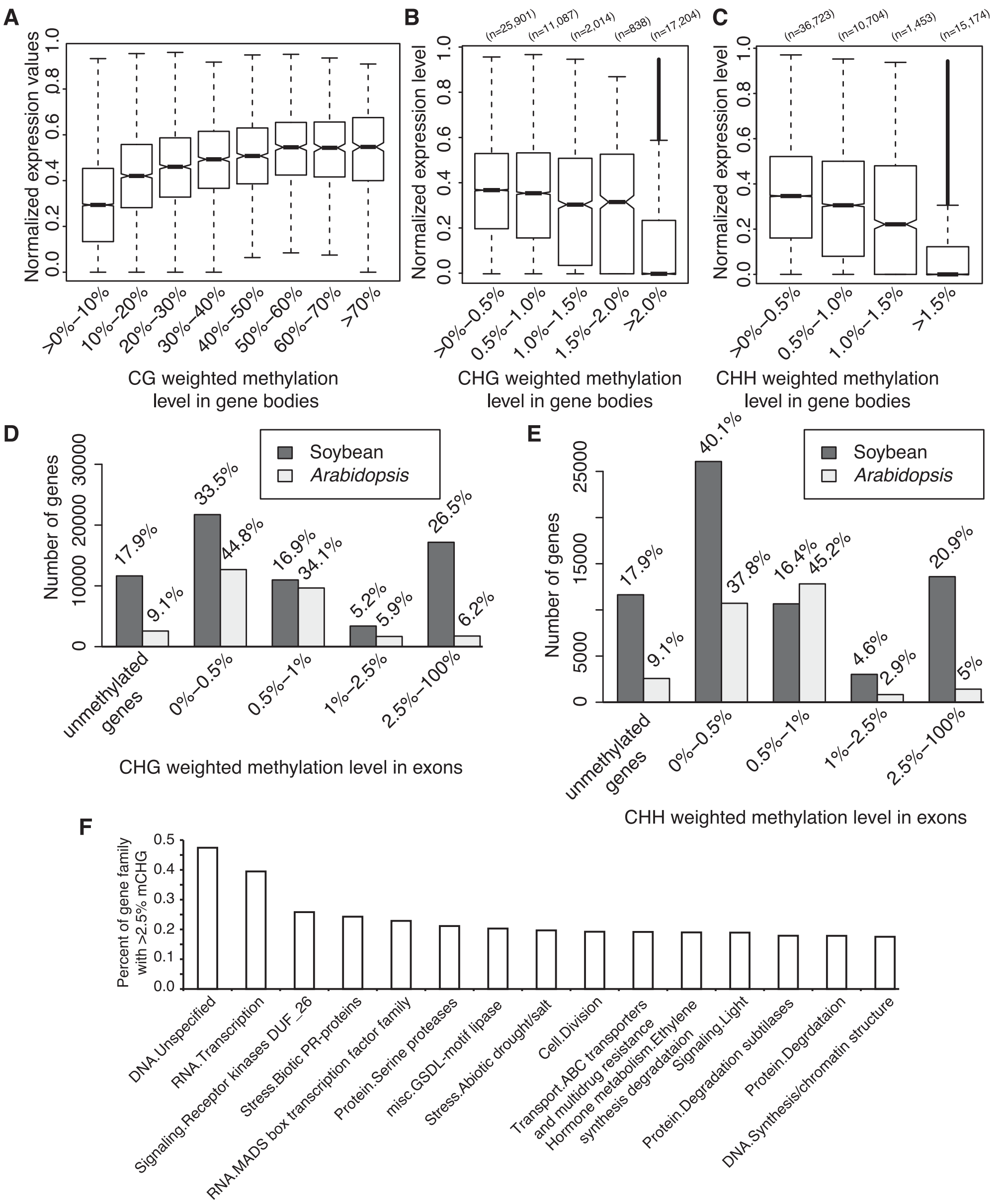

Figure 3. The association between DNA methylation and gene expression levels in soybean. (*A*) Increasing levels of CG gene-body methylation is correlated with increasing levels of gene expression. Box plot representation of different levels of CG gene-body methylation is displayed along the *x*-axis, whereas normalized gene expression levels are plotted on the *y*-axis. Genes containing >0.5% non-GC methylation were filtered from this analysis. Increasing methylation levels of both CHG (*B*) and CHH (*C*) sites are associated with decreasing levels of gene expression levels. (*D,E*) The soybean genome contains a higher proportion of RdDM-targeted loci compared to the *Arabidopsis thaliana* genome. Fraction of genes (*y*-axis) targeted by varying levels of CHG (*D*) and CHH (*E*) DNA methylation (*x*-axis). Unmethylated genes were defined as loci containing <0.5% CG, CHG, and CHH methylation. (*F*) Gene families containing the highest fraction of members containing >2.5% weighted CHG methylation. Only gene families with more than 100 members were considered in this analysis and only the top fifteen classes are displayed.

which is consistent with the ability of the RdDM pathway to actively repress subsets of soybean genes.

The soybean genome contains ~66,000 protein-coding genes (Schmutz et al. 2010) in contrast to the ~27,000 protein-coding genes present in the *Arabidopsis thaliana* Col-0 genome (The Arabidopsis Genome Initiative 2000), which reflects recent whole genome duplications (Schmutz et al. 2010). The soybean LD genome contains higher amounts of all three types of DNA methylation (Fig. 1A) and greater proportions of CHG methylation when compared to *Arabidopsis thaliana* (Fig. 1B), which is likely a result of the RdDM pathway more actively targeting genes in soybean (Fig. 3D,E). We defined genes as possible targets of RdDM that contained >2.5% of either CHG or CHH weighted methylation levels as these levels had a measurable effect on gene expression. Although ~6% and ~5% of the *Arabidopsis thaliana* genes are targeted, respectively, by CHG and CHH methylation, ~26% and ~20% are targeted in soybean, representing an approximately fourfold increase (Fig. 3D). A closer inspection of the top 15 classes of genes that are most frequently targeted by CHG and CHH methylation (Fig. 3F) revealed enrichment for biotic pathogen response proteins, MADS-box transcription factors and protein degradation machinery, which is similar to the three gene families most targeted by CHG and CHH methylation in *Arabidopsis thaliana* (Schmitz et al. 2013).

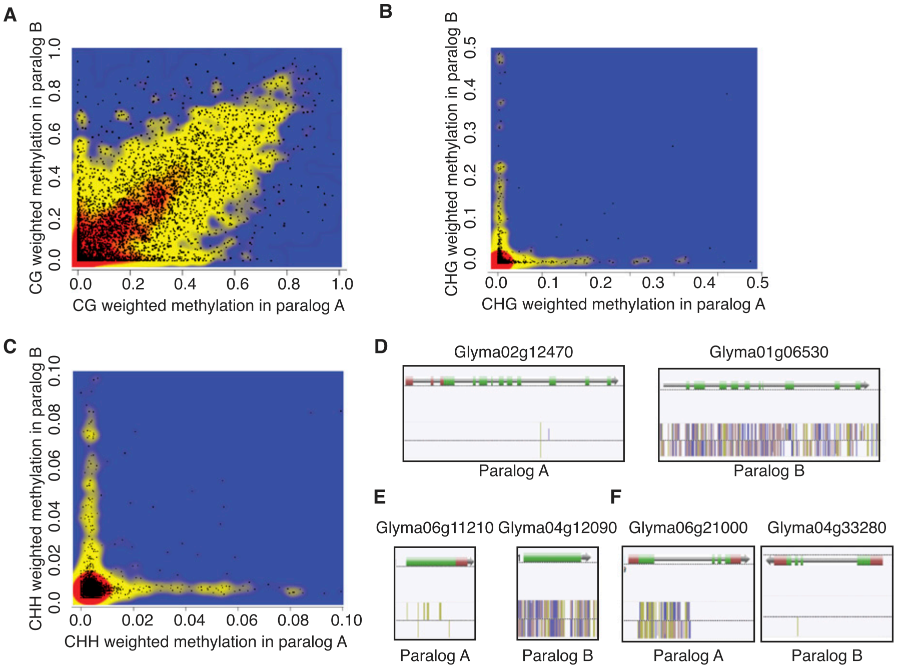

Figure 4. Pairwise methylation levels for paralog gene pairs. (*A*–*C*) Pairwise plots of CG (*A*), CHG (*B*), and CHH (*C*) methylation levels for all paralog pairs of genes. The A form of the paralog is plotted along the *x*-axis and the B form is plotted along the *y*-axis. (*D*–*F*) Examples of variation in RdDM-like methylation between paralog pairs in the LD methylome. (Gold lines) mCG; (purple lines) mCHG; (pink lines) mCHH. Loci containing >0.5% non-GC methylation levels were excluded from the plot in *A*, and genes containing >2.5% mCHG in *B* and *C* were considered targets of RdDM.

RdDM targets recently duplicated paralogs

Recently duplicated genes in *Arabidopsis thaliana* show a strong preference for maintenance of methylation states (Widman et al. 2009), but analysis of the effects of DNA methylation on gene expression in this species is limited by a relatively low number of paralogs. In contrast, the whole-genome duplications in soybean have resulted in almost 10,000 identified pairs of paralogs, and it is plausible that these recent duplications underlie the increased number of genes targeted by CHG and CHH methylation detected above. To determine if paralogs are more likely to contain one paralog that is enriched for non-CG methylation, the CG, CHG, and CHH methylation levels of these pairs were plotted against one another (Fig. 4A–C). For CG methylation, the density plot revealed that most pairs are methylated at relatively equal levels, as the diagonal of the plot is most dense (Fig. 4A). This pattern reflects that CG gene-body methylation, which is not repressive in nature, is largely maintained between paralogs similar to previous reports for orthologs (Takuno and Gaut 2013). This pattern contrasts with CHG and CHH methylation in which the vast majority of paralogs are unmethylated or methylated at very low levels and present near the lower left corner of the plot (Fig. 4B,C). Interestingly, clear examples of differentially methylated paralogs are present along the zero plane of the *x*- or *y*-axis (Fig. 4B–F; Supplemental Table 3). A total of 602/9793 paralogs in soybean were differentially targeted by non-CG methylation, which represents a significant enrichment (P-value < 6.648×10^{-7}) (Methods) when compared to *Arabidopsis thaliana* paralogs (4/497). Furthermore, the methylated forms of the paralogs were expressed at significantly lower levels (P-value < 2×10^{-16}, Wilcoxon signed-rank test) and resided closer to transposons when compared to their unmethylated counterparts (Supplemental Fig. 7). In fact, 56/602 differentially methylated paralogs were strictly defined by the presence of a transposon that was targeted by CG, CHG, and CHH methylation overlapping the gene space. These results indicate that one potential route to gene expression variation in recently duplicated genomes among paralogs is through the actions of DNA methylation and in some cases nearby transposon sequences.

Variation in DNA methylation among soybean parental and recombinant inbred lines

To explore the potential for natural variation of DNA methylation patterns in soybean, MethylC-seq and RNA-seq were performed on the LDX01-1-165 germplasm (hereafter referred to as "LDX"). Additionally, the methylomes and transcriptomes of two recombinant inbred lines (RILs) that were derived from the LD and LDX parental germplasms (Kim et al. 2011) were profiled to enable determination of the heritability of DNA methylation states upon combination of newly introduced genetic variants. Over 100 million 101-bp aligned reads were recovered for LDX and the two RILs (R-11268 and R-11272) representing greater than ~11× coverage for each sample (Supplemental Table 1). RNA-seq data was acquired for all three lines in biological triplicates and had at least 30 million aligned reads per sample (Supplemental Table 2).

The single-base resolution bisulfite-sequencing data enabled identification of single methylation polymorphisms (SMPs) (Schmitz et al. 2011), differentially methylated regions only in the CG context (CG-DMRs), and differentially methylated regions in all types of DNA methylation (C-DMRs) present between the four sequenced lines. In total, 280,712 CG-SMPs, 703,685 CHG-SMPs, and 9,819,894 CHH-SMPs were identified between the parental and RIL methylomes. CG-SMPs were more abundant in genes and more specifically in introns compared to transposon and intergenic sequences (Supplemental Fig. 8A), whereas CHG- and CHH-SMPs were more abundant in transposon sequences (Supplemental Fig. 8A). The patterns of CG-SMP variability are similar to the patterns observed for CG-DMRs (Supplemental Fig. 8B).

A total of 3241 CG-DMRs were identified among the four lines sequenced, and 61% of these overlapped gene bodies and were found in similar distributions across gene bodies (Supplemental Fig. 8C), similar to the patterns of CG gene-body methylation (Fig. 2A). To determine the potential impact of these CG-DMRs on gene expression, the methylation levels of each CG-DMR were plotted against the gene expression level of the locus overlapping the position of each CG-DMR within each sample (Supplemental Fig. 8D). Regardless of the methylation level present within each of the

CG-DMRs, the gene expression levels remained constant, indicating that in these limited samples no clear correlation between CG-DMR methylation levels and gene expression levels was detected (Supplemental Fig. 8D). Although CG-DMRs are an average size of 431 bp and preferentially found in gene bodies, C-DMRs, of which there are 1416, are an average size of 1162 bp, also abundant in gene bodies (Supplemental Fig. 8E,F), and are most abundant within 1 kb of the transcriptional start site of genes (Supplemental Fig. 8G). However, for C-DMRs that overlapped genes, increasing levels of methylation within each C-DMR are correlated with decreasing levels of gene expression (Supplemental Fig. 8H), indicating that these types of DMRs can contribute to the variation in gene expression observed between different genotypes.

Additionally, whole-genome sequencing data were obtained for the LD and LDX parental lines and SNPs were identified using the SHORE analysis pipeline (Ossowski et al. 2008). As expected, most of the SNPs identified in each genotype sequenced were located in intergenic regions, but significant fractions were identified in protein-coding genes (Supplemental Table 4). Major effect mutations were identified and defined as SNPs that abolished known start and stop codons, as well those SNPs that created premature stop codons. We hypothesized that loci that are targeted by non-CG methylation may accumulate major effect mutations at a higher rate than non-unmethylated loci because they are not frequently expressed in sporophytic tissues, but were unable to find any significant correlation to support this claim (χ^2 test, P-value = 0.93). Therefore, it is likely that the repressive DNA methylation at these loci has evolved for other purposes, some of which may be important for plant development (Zemach et al. 2010a; Martínez and Slotkin 2012; Schmitz et al. 2013), germ line maintenance (Slotkin et al. 2009; Calarco et al. 2012; Ibarra et al. 2012) and/or responses to biotic stresses (Dowen et al. 2012).

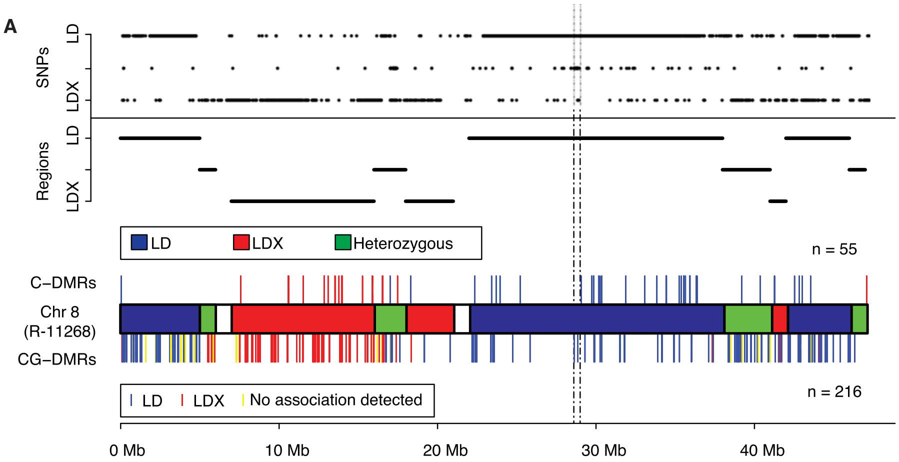

Figure 5. Heritability of CG- and C-DMRs in two soybean recombinant inbred lines. (*A*) Reconstitution of recombinant inbred line genotypes using SNPs between LD and LDX from bisulfite sequencing data. SNPs were combined to determine homozygous LD or LDX and heterozygous regions of each chromosome in R-11268 and R-11272. An example is shown for the patterns of heritability for both CG-DMRs and C-DMRs from chromosome 8 from R-11268. The entire data set can be found in Supplemental Tables 5 and 6. (Blue shaded areas) LD homozygous regions; (red shaded areas) LDX homozygous regions; (green shaded regions) heterozygous regions; (white shaded regions) ambiguous. (Blue bars) LD; (red bars) LDX; (yellow bars) no association detected.

Cosegregation analysis of DMRs and genotype

Although there is extensive methylation variation within and between plant species, the heritability of methylation variants has not been extensively explored in a population on a genome-wide scale. To understand the stability and heritability of methylation variants, we examined the methylation levels of CG- and C-DMRs in homozygous regions of R-11268 and R-11272 and compared them to their parental states in LD and LDX (Fig. 5A; Methods). In total, 3670/4474 and 1924/2048 of the methylation levels of CG-DMRs and C-DMRs, respectively, in R-11268 and R-11272 cosegregated with the parental state (Supplemental Tables 5, 6), whereas 254/4981 for CG-DMRs and 122/2048 for C-DMRs were found to contain the methylation state of the other parent. This would suggest that methylation states of some DMRs are due to distant loci or are epigenetically unstable, as has been observed in *Arabidopsis thaliana* and maize (Becker et al. 2011; Eichten et al. 2011; Schmitz et al. 2011), although other possible explanations could include incorrect assignment of DMRs to their genotype and low sequencing coverage of DMRs.

Population-wide identification of methylQTL

The epigenomics approach undertaken in this study enabled identification of methylation variants that are both linked and unlinked to genotype (although the latter is much rarer), but the low sample size of only two RILs makes understanding the population dynamics of methylation states difficult. However, because the methylation status of the majority of C-DMRs cosegregated with their genotype in the two RILs, it should be possible to map potential causal variants for the methylation variation in this population. DNA methylome data were acquired for an additional 81 lines from the RIL population and QTL mapping for each C-DMR was performed, which revealed evidence for a methylQTL for 1293/1416 (91%) C-DMRs (Fig. 6A; Supplemental Table 7). Of the identified methylQTL, 1260/1293 mapped locally to the C-DMR (Fig. 6B,C; Supplemental Fig. 9), whereas 33 mapped to a different chromosome from where the C-DMR was located (Fig. 6D,E). Lastly, heritability estimates for each methylQTL were calculated, which revealed that many methylQTL could explain a large proportion of the methylation variation of their associated C-DMR (Fig. 6F). The methylQTL with lower heritability estimates could be reflective of methylation variants that display higher epimutation rates, possibly because these variants are not directly linked to a genetic variant. Future efforts to identify causal genetic variants will be necessary to understand the stability of different classes of methylation variants.

Discussion

Studies in plants have led to major advances in the field of epigenetics, especially with regard to natural epigenetic variation (Weigel and Colot 2012). Plant genomes contain cytosine DNA methylation that occurs not only in the CG context but also in CHG and CHH contexts (Cokus et al. 2008; Lister et al. 2008), and these specific signatures are often indicative of the type of regulation occurring at the methylated locus. Epigenomic techniques have revealed widespread natural variation in DNA methylation in a range of plant species (Vaughn et al. 2007; Zhang et al. 2008; He et al. 2010; Becker

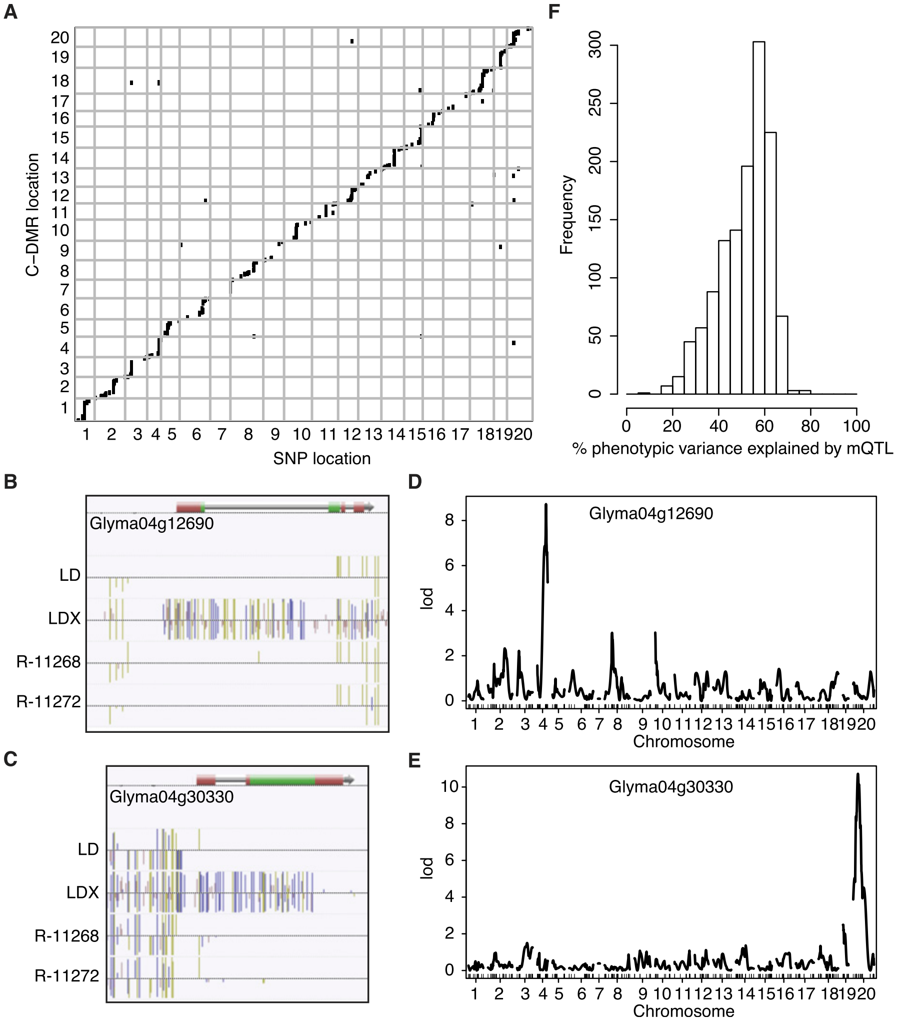

Figure 6. Population-level analysis of methylation variants and identification of methylQTL. (*A*) Scatter plot of the location (*x*-axis) of all single methylQTL identified for C-DMRs (*y*-axis). (*B,C*) DNA methylation profiles for LD, LDX, R-11268 and R-11272 of two different C-DMRs. (Gold lines) mCG; (purple lines) mCHG; (pink lines) mCHH. (*D*) An example QTL map of a local methylQTL and (*E*) a distant methylQTL. (*F*) Broad-sense heritability estimates for all single methylQTL/C-DMR pairs.

et al. 2011; Eichten et al. 2011; Groszmann et al. 2011; Schmitz et al. 2011; Greaves et al. 2012; Shen et al. 2012), but we are only beginning to understand the role of DNA methylation and modes of inheritance for different methylation variants.

Although whole-genome bisulfite sequencing data revealed natural epigenomic variation between soybean germplasms, it revealed a greater proportion of the methylome that was invariably methylated. In fact, although the soybean genome is approximately eightfold larger than the *Arabidopsis thaliana* genome, it contained proportionally more DNA methylation, which was disproportionally present in CHG and CHH sites, indicating that RdDM is more active in the soybean genome. A closer inspection of the regions of the genome targeted by non-CG methylation revealed that approximately fourfold more protein-coding regions are actively silenced. Given the recent genome duplications present in the soybean genome (Schmutz et al. 2010), this additional targeting could indicate that these genes are being purged from the genome or expressed at very low levels until subfunctionalization occurs (Roulin et al. 2012). If this were the case, it would be expected that these loci accumulate higher frequencies of major effect mutations compared to loci that do not contain non-CG methylation, but we were unable to find any evidence to support this possibility. Therefore, these genes might be present within the genome in a transcriptionally inert state, which could allow them to function during situations that result in global reactivation of loci that contain non-CG methylation, which can occur upon biotic infection or during certain developmental stages (Zemach et al. 2010a; Dowen et al. 2012; Martínez and Slotkin 2012; Schmitz et al. 2013; Yu et al. 2013) similar to reports for transposons (Ohtsu et al. 2007; Slotkin et al. 2009; Li et al. 2010; Calarco et al. 2012; Ibarra et al. 2012).

One additional role of non-GC methylation in soybean genome found in this study is differential targeting of paralogs, which drives gene expression variation. Whether or not this is restricted to soybean or widespread among additional crop genomes is unknown. Although, it was recently reported that differential accumulation of H3K27me3 between maize paralogs preferentially occurs in recently duplicated regions of the genome (Makarevitch et al. 2013). Therefore, it is likely that there are multiple epigenomic mechanisms that can lead to expression variation and will undoubtedly be an interesting topic for future investigations.

In this study, we have revealed that the majority of C-DMR methylation variants identified cosegregated with the genetic background from which they were derived, but there were rare examples of uncoupling between methylation states and genotype, which potentially provide an additional source for natural epigenetic variation. One possible mechanism to explain the methylation variants that did not cosegregate with their genotype could include paramutation, as has been observed in maize (Patterson et al. 1993; Arteaga-Vazquez and Chandler 2010), but analysis of these C-DMRs did not reveal such events. For the C-DMR methylation variants that do follow standard laws of inheritance, their stability is likely a result of being targeted by non-GC methylation, a process that would enact a double-hit mechanism by taking advantage of the activities of maintenance methyltransferases at CG and CHG sites (Ronemus et al. 1996; Mathieu et al. 2007; Du et al. 2012) in addition to small RNA directed methylation by de novo methyltransferases at all cytosines (Cao et al. 2003; Teixeira et al. 2009).

The heritability of methylation states of C-DMRs suggested that some of these C-DMRs might have arisen as a consequence of genetic variants, as has been observed with the *PAI* gene family and the *AtFOLT1* paralogs in *Arabidopsis thaliana* (Bender and Fink 1995; Durand et al. 2012), whereas others could have arisen and segregated independently of genetic variants like *QQS* (Silveira et al. 2013). In fact, QTL mapping of this population uncovered a methylQTL for ~91% of the C-DMRs, indicating that these C-DMRs are either linked to a genetic variant or are stably inherited, whereas the methylation states of the remaining ~9% of C-DMRs likely do not follow standard laws of inheritance. The vast majority of these methylQTL mapped to the C-DMR, but there was also clear evidence for methylQTL located on different chromosomes than the C-DMR. It should be noted that genome rearrangements in the parental lines or misassemblies in the reference genome could explain some of these distant methylQTL. For example, there are four C-DMRs on chromosome 18 that are all significantly associated with a single marker on chromosome 4. In any case, these methyQTL are candidate regions that in many cases likely harbor a causal genetic variant(s) underlying the methylation variation of the respective C-DMR. Identifying the types of causal variants that lead to methylation variation in plants will require large-scale epigenomic projects using natural plant populations, which will enable higher-resolution association mapping. In fact, a number of phenotyping and sequencing projects are already underway that will advance the use of quantitative genetic approaches to understanding natural variation of morphological or molecular phenotypes of interest (Lam et al. 2010; Cao et al. 2011; Gan et al. 2011; Huang et al. 2012).

Methods

Plant material

The two parental lines, LD00-2817P (Diers et al. 2010) and LDX01-1-65 (Brucker et al. 2005), were used to create the studied RIL population (see Supplemental Methods for additional information).

Construction of sequencing libraries

DNA sequencing libraries for LD and LDX were constructed as reported in Johnson et al. (2012). MethylC-seq libraries were constructed according to Schmitz et al. (2011). RNA-seq libraries were constructed using the Illumina TruSeq Kit v2 according to the manufacturer's guidelines.

Sequencing

gDNA-seq, MethylC-seq, and RNA-seq libraries were sequenced using an Illumina HiSeq 2000 according to the manufacturer's instructions. gDNA-seq and MethylC-seq libraries were sequenced for 101 cycles, and RNA-seq libraries were sequenced for 51 cycles. An additional run of paired-end 2 × 101 bp sequencing was performed for the gDNA libraries.

RNA-seq analysis

All RNA samples were performed as biological triplicates for each genotype. Illumina HiSeq2000 output files in the FASTQ format were aligned to the *Glycine max* reference genome version 1.0 (Schmutz et al. 2010) (Gm1.0 ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v8.0/Gmax_v1.0/) using Bowtie version 0.12.7 (Langmead et al. 2009) and TopHat version 1.3.3 (Trapnell et al. 2009) (flags = -g 1, –F 0). Gene expression values were calculated using Cufflinks version 1.1.0 (flags = -F 0, -b, -N) (Trapnell et al. 2010).

MethylC-seq analysis

MethylC-seq analysis was performed similarly to Lister et al. (2011) with some modifications (see Supplemental Methods).

Identification of DMRs

To identify DMRs, a root mean square test (Perkins et al. 2011) was applied to all cytosines, which required building a contingency table where the rows indicated a particular sample and the columns indicated the number of reads that supported a methylated cytosine or an unmethylated cytosine at each position in a given sample. Using 10,000 permutations, the *P*-values were simulated; and for each new permutation, a contingency table was generated by randomly assigning reads to cells with a probability equal to the product of the row marginal and column marginal divided by the total number of reads squared. To increase the efficiency of this process, if a *P*-value returned 100 permutations with a statistic greater than or equal to the original test statistic, permutations were discontinued (i.e., we used adaptive permutation testing). To determine a *P*-value cutoff that would control the false discovery rate (FDR) at a rate of 1%, the procedure in Bancroft et al. (2013) was applied. Briefly, this method first generates a histogram of the *P*-values and calculates the expected number of *P*-values to fall in a particular bin under the null. This expected count is computed by multiplying the width of the bin by the current estimate for the number of true null hypotheses (m_0), which is initialized to the number of tests performed. It then looks for the first bin (starting from the most significant bin and working its way toward the least significant) where the expected number of *P*-values is greater than or equal to the observed value. The differences between the expected and observed counts in all the bins up to this point are summed, and a new estimate of m_0 is generated by subtracting this sum from the current total number of tests. This procedure was iterated until convergence, which we defined as a change in the m_0 estimate less than or equal to 0.01. With this m_0 estimate, we were able to estimate the FDR of a given *P*-value by multiplying the *P*-value by the m_0 estimate (the expected number of positives at that cutoff under the null hypothesis) and dividing that product by the total number of significant tests we detected at that *P*-value cutoff. We chose the largest *P*-value cutoff that still satisfied a 1% FDR requirement. Once this *P*-value cutoff was chosen, significant sites were combined into blocks if they were within 500 bases of one another and had methylation changes in the same direction (e.g., sample A was hypermethylated and sample B was hypomethylated at both sites). Three different types of DMRs were identified from the data set—C-DMR (a change in all three contexts), CG-DMR (a change only in the CG context), and CH-DMR (a change in either the CHG or the CHH contexts). Furthermore, C-DMR, CG-DMR, and CH-DMR blocks that contained fewer than 10, 5, and 5 differentially methylated sites were discarded, respectively. Final lists of C-DMRs required an overlap with both a CG-DMR and a CH-DMR, and the final list of CG-DMRs were only retained if they did not overlap a C-DMR or a CH-DMR.

Weighted methylation levels

Weighted methylation levels were computed as described in Schultz et al. (2012).

Identification of "early" and "recent" whole genome duplications (WGD) and paralogs

Synteny blocks were identified with DAGchainer (Haas et al. 2004), based on anchor points determined using the NCBI *blastp* program (*E*-value ≤ 1×10^{-10}), filtered to the top reciprocal best matches per chromosome pair. Synteny blocks from *Glycine* ("recent") WGD were identified as those with median K_s values ≤ 0.35 per block, and blocks from the legume ("old") WGD were identified as those with median K_s values > 0.35 and ≤ 1.5 per block. K_s values per gene were determined using the codeml program from the PAML package, version 4.4 (Yang 2007). The paralogs used in this study and the descriptions of how they were identified were obtained from Supplemental Table 4 in a previously published study (Libault et al. 2010). Furthermore, our analysis on soybean paralogs was strictly focused on genes harboring two copies in the genome that were strictly duplicated genes from the recent WGD. Differentially methylated paralogs were identified by searching for pairs in which one paralog contained >2.5% CHG methylation and the other paralog had <0.5% CHG methylation. *Arabidopsis thaliana* paralogs were obtained from Supplementary Material in a previously published study (Ganko et al. 2007). The "prop.test" function in R was used to estimate if the proportion of significantly differentially methylated paralogs in soybean was greater than in *Arabidopsis thaliana*.

Small RNA analysis

smRNA data were downloaded from the National Center for Biotechnology for Information SRA012752 (Tuteja et al. 2009). These small RNAs were isolated from young cotyledons from the Williams accession. Raw smRNA data were preprocessed by removing the 3′ adapter sequence and any sequencing reads under 16 bp. Reads passing these filters were aligned to the Gm1.0 reference genome using the Bowtie (v0.12.7) and the following parameters: -e 1 -l 20 -n 0 -a -m 1000–best–nomaqround. Only reads that contained perfect matches within the genome and that did not have more than a thousand locations were retained for further data analysis.

Identification of SNPs

SNPs were identified using the SHORE variant identification software package (Ossowski et al. 2008) using the BWA aligner (Li and Durbin 2009), allowing up to 5% errors per read and a max of three gaps. Any SNP with a quality score of 25 or above was used for further analysis.

SNP effects

The impact of SNPs on coding regions were determined using the SnpEff tool ("SnpEff: Variant effect prediction"; http://snpeff.sourceforge.net) (Cingolani et al. 2012) using a *Glycine max* reference file.

Genetic reconstruction of RILs based on bisulfite sequencing reads

Only SNPs that distinguished the LD and LDX parental lines were used to determine the genotypes of the RILs. All SNP pairs containing C-T, T-C, A-G, or G-A changes were excluded because we were unable to distinguish those SNPs due to bisulfite conversion of reads. Next, the number of reads in each RIL that matched the LD or LDX alleles was determined using the bisulfite converted reads, and any position containing at least four reads matching a parent was considered for further analysis. A position was determined heterozygous if at least four reads were identified that supported each parent. Using these data, a score was assigned as 1.0 for the LD genotype, 0 for heterozygous positions, and -1.0 for the LDX genotype. Next, the genome was divided into 100-kb bins, and the score for each bin was computed by averaging the scores of each position within it. Only bins with greater than 10 SNPs were included in the calculation. Next, we assigned tags to bins based on the score. Bins with a score greater than 0.5, were tagged as LD, whereas bins with a score of less than -0.5 were assigned as LDX. Bins without a score were kept untagged, and the rest (with score between -0.5 and 0.5) were labeled as heterozygous. Lastly, large regions were formed by concatenating adjacent bins with the same tag (LD or LDX) or bins that had the same tag but were spaced by untagged bins.

Assignment of DMRs to genotype

DMRs were assigned to genotypes based on their overlap with the genetic reconstruction of the R-11268 and R-11272 RILs. If a DMR was within one reconstructed region, then its genotype was the same as that region. To determine the parental methylation state of each DMR, the weighted methylation level was computed for each genotype (LD, LDX, R-11268, and R-11272). If the value of RIL was within 20% of either one of the parents, then the DMR was assigned to that parent. If the value of RIL was more extreme or between either parent-weighted methylation level, then it was labeled "No association detected." To compare the methylation level of DMRs in the offspring and the parents, we computed the ratio of the weighted methylation level of each region in the RILs and then subtracted the lower weighed methylation level from each parent to determine the absolute difference between the weighted methylation levels of two parents. The ratio can be represented using the following equation:

$$\begin{aligned} Ratio = {} & Met - \min(Met_{LD}, Met_{LDX}) \max Met_{LD}, Met_{LDX} \\ & - \min(Met_{LD}, Met_{LDX}) \end{aligned}$$

where *Met* is the weighted methylation level of this region in a RIL, and Met_{LD} *and* Met_{LDX} are the weighted methylation levels in LD and LDX, respectively.

QTL mapping of C-DMRs

The R/qtl package (Broman et al. 2003) was used to map QTL for each C-DMR. First, missing genotypes were imputed using the "fill.geno" function. Next, genotypes between SNP markers were simulated and imputed using the "sim.geno" function with the following parameters: "step=1, error.prob=0.01, n.draws=20." Then, for each C-DMR, the "scanone" function (option: "model='np' ") was used to compute a LOD score for each SNP marker across the genome. Permutation testing (1000 times) was used to estimate the significance of each LOD peak(s). methylQTL was defined as the closest significant SNP marker (*P*-value < 0.01) to the summit of the highest peak. Only the single highest LOD score was reported for each C-DMR. The broad-sense heritability of each QTL was estimated by doing an ANOVA analysis using the "fitqtl" function.

Additional analyses

For analysis of SMPs, transposons, and for information regarding gene annotations *Arabidopsis thaliana* data used in this study, see Supplemental Methods.

Data access

The data generated for this work have been deposited in the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) and are accessible through accession number GSE41753. Genome sequencing data have been deposited in the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number SRA060034. Processed data can be visualized at http://neomorph.salk.edu/soybean_RIL_methylomes/browser.html.

Acknowledgments

We thank Mathew Lewsey and Matthew Schultz for their comments on this manuscript; Steve Cannon for identification of the "early" and "recent" whole genome duplications; and Justin Borevitz and Riyan Cheng for discussions about QTL analysis. R.J.S. was supported by a National Institutes of Health postdoctoral fellowship (K99GM100000). This research was funded by the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation (through Grant GBMF3034); the Mary K. Chapman Foundation; and the National Science Foundation (MCB-0929402 and MCB1122246) to J.R.E. This work was also supported by the Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences of the U.S. Department of Energy through Grant DE-FG02-08ER15309, as well as funding from the United Soybean Board and North Central Soybean Research Program to G.S.

Author contributions: R.J.S., J.R.E., and G.S. conceived and designed experiments. R.J.S., O.V.L., M.A.U., and J.R.N. performed experiments. R.J.S., Y.H., S.M.K., T.J., and B.D. analyzed the data. D.X. provided guidance and valuable insight. R.J.S. wrote the paper.

References

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Arteaga-Vazquez MA, Chandler VL. 2010. Paramutation in maize: RNA mediated *trans*-generational gene silencing. *Curr Opin Genet Dev* **20:** 156–163.

Bancroft T, Du C, Nettleton D. 2013. Estimation of false discovery rate using sequential permutation *p*-values. *Biometrics* **69:** 1–7.

Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D. 2011. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480:** 245–249.

Bender J, Fink GR. 1995. Epigenetic control of an endogenous gene family is revealed by a novel blue fluorescent mutant of *Arabidopsis*. *Cell* **83:** 725–734.

Broman KW, Wu H, Sen S, Churchill GA. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19:** 889–890.

Brucker E, Carlson S, Wright E, Niblack T, Diers B. 2005. *Rhg1* alleles from soybean PI 437654 and PI 88788 respond differentially to isolates of *Heterodera glycines* in the greenhouse. *Theor Appl Genet* **111:** 44–49.

Calarco JP, Borges F, Donoghue MT, Van Ex F, Jullien PE, Lopes T, Gardner R, Berger F, Feijó JA, Becker JD, et al. 2012. Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell* **151:** 194–205.

Cao X, Aufsatz W, Zilberman D, Mette MF, Huang MS, Matzke M, Jacobsen SE. 2003. Role of the *DRM* and *CMT3* methyltransferases in RNA-directed DNA methylation. *Curr Biol* **13:** 2212–2217.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43:** 956–963.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}*; iso-2; iso-3*. *Fly (Austin)* **6:** 80–92.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452:** 215–219.

Cubas P, Vincent C, Coen E. 1999. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401:** 157–161.

Diers BW, Cary T, Thomas D, Colgrove A, Niblack T. 2010. Registration of LD00-2817P soybean germplasm line with resistance to soybean cyst nematode from PI 437654. *J Plant Registrations* **4:** 141–144.

Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, Dixon JE, Ecker JR. 2012. Widespread dynamic DNA methylation in response to biotic stress. *Proc Natl Acad Sci* **109:** E2183–E2191.

Doyle JJ, Doyle JL, Harbison C. 2003. Chloroplast-expressed glutamine synthetase in *Glycine* and related Leguminosae: Phylogeny, gene duplication, and ancient polyploidy. *Syst Bot* **28:** 567–577.

Du J, Zhong X, Bernatavichute YV, Stroud H, Feng S, Caro E, Vashisht AA, Terragni J, Chin HG, Tu A, et al. 2012. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151:** 167–180.

Durand S, Bouche N, Perez Strand E, Loudet O, Camilleri C. 2012. Rapid establishment of genetic incompatibility through natural epigenetic variation. *Curr Biol* **22:** 326–331.

Eichten SR, Swanson-Wagner RA, Schnable JC, Waters AJ, Hermanson PJ, Liu S, Yeh CT, Jia Y, Gendler K, Freeling M, et al. 2011. Heritable epigenetic variation among maize inbreds. *PLoS Genet* **7:** e1002372.

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci* **107:** 8689–8694.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477:** 419–423.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. *Mol Biol Evol* **24:** 2298–2309.

Gill N, Findley S, Walling JG, Hans C, Ma J, Doyle J, Stacey G, Jackson SA. 2009. Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol* **151:** 1167–1174.

Greaves IK, Groszmann M, Ying H, Taylor JM, Peacock WJ, Dennis ES. 2012. Trans chromosomal methylation in *Arabidopsis* hybrids. *Proc Natl Acad Sci* **109:** 3570–3575.

Groszmann M, Greaves IK, Albertyn ZI, Scofield GN, Peacock WJ, Dennis ES. 2011. Changes in 24-nt siRNA levels in *Arabidopsis* hybrids suggest an epigenetic contribution to hybrid vigor. *Proc Natl Acad Sci* **108:** 2617–2622.

Haas BJ, Delcher AL, Wortman JR, Salzberg SL. 2004. DAGchainer: A tool for mining segmental genome duplications and synteny. *Bioinformatics* **20:** 3643–3646.

He G, Zhu X, Elling AA, Chen L, Wang X, Guo L, Liang M, He H, Zhang H, Chen F, et al. 2010. Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell* **22:** 17–33.

Hitchins MP, Wong JJ, Suthers G, Suter CM, Martin DI, Hawkins NJ, Ward RL. 2007. Inheritance of a cancer-associated *MLH1* germ-line epimutation. *N Engl J Med* **356:** 697–705.

Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, et al. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490:** 497–501.

Ibarra CA, Feng X, Schoft VK, Hsieh TF, Uzawa R, Rodrigues JA, Zemach A, Chumak N, Machlicova A, Nishimura T, et al. 2012. Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science* **337:** 1360–1364.

Innes RW, Ameline-Torregrosa C, Ashfield T, Cannon E, Cannon SB, Chacko B, Chen NW, Couloux A, Dalwani A, Denny R, et al. 2008. Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. *Plant Physiol* **148:** 1740–1759.

Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, et al. 2009. Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5:** e1000530.

Johnson DB, Wang C, Xu J, Schultz MD, Schmitz RJ, Ecker JR, Wang L. 2012. Release factor one is nonessential in *Escherichia coli*. *ACS Chem Biol* **7:** 1337–1344.

Kim M, Hyten DL, Niblack TL, Diers BW. 2011. Stacking resistance alleles from wild and domestic soybean sources improves soybean cyst nematode resistance. *Crop Sci* **51:** 934–943.

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* **42:** 1053–1059.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11:** 204–220.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li H, Freeling M, Lisch D. 2010. Epigenetic reprogramming during vegetative phase change in maize. *Proc Natl Acad Sci* **107:** 22184–22189.

Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G. 2010. An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *Plant J* **63:** 86–99.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133:** 523–536.

Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, et al. 2011. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471:** 68–73.

Makarevitch I, Eichten SR, Briskine R, Waters AJ, Danilevskaya ON, Meeley RB, Myers CL, Vaughn MW, Springer NM. 2013. Genomic distribution of maize facultative heterochromatin marked by trimethylation of H3K27. *Plant Cell* **25:** 780–793.

Manning K, Tör M, Poole M, Hong Y, Thompson AJ, King GJ, Giovannoni JJ, Seymour GB. 2006. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat Genet* **38:** 948–952.

Martínez G, Slotkin RK. 2012. Developmental relaxation of transposable element silencing in plants: Functional or byproduct? *Curr Opin Plant Biol* **15:** 496–502.

Mathieu O, Reinders J, Caikovski M, Smathajitt C, Paszkowski J. 2007. Transgenerational stability of the *Arabidopsis* epigenome is coordinated by CG methylation. *Cell* **130:** 851–862.

Ohtsu K, Smith MB, Emrich SJ, Borsuk LA, Zhou R, Chen T, Zhang X, Timmermans MC, Beck J, Buckner B, et al. 2007. Global gene expression analysis of the shoot apical meristem of maize (*Zea mays* L.). *Plant J* **52:** 391–404.

Ossowski S, Schneeberger K, Clark R, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* **18:** 2024–2033.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327:** 92–94.

Patterson GI, Thorpe CJ, Chandler VL. 1993. Paramutation, an allelic interaction, is associated with a stable and heritable reduction of transcription of the maize *b* regulatory gene. *Genetics* **135:** 881–894.

Perkins W, Tygert M, Ward R. 2011. $\chi 2$ and classical exact tests often wildly misreport significance; the remedy lies in computers. http://arxiv.org/abs/1108.4126.

Rangwala SH, Elumalai R, Vanier C, Ozkan H, Galbraith DW, Richards EJ. 2006. Meiotically stable natural epialleles of *Sadhu*, a novel *Arabidopsis* retroposon. *PLoS Genet* **2:** e36.

Reinders J, Wulff BB, Mirouze M, Mari-Ordóñez A, Dapp M, Rozhon W, Bucher E, Theiler G, Paszkowski J. 2009. Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes Dev* **23:** 939–950.

Richards EJ. 2006. Inherited epigenetic variation—revisiting soft inheritance. *Nat Rev Genet* **7:** 395–401.

Ronemus MJ, Galbiati M, Ticknor C, Chen J, Dellaporta SL. 1996. Demethylation-induced developmental pleiotropy in *Arabidopsis*. *Science* **273:** 654–657.

Roulin A, Auer PL, Libault M, Schlueter J, Farmer A, May G, Stacey G, Doerge RW, Jackson SA. 2012. The fate of duplicated genes in a polyploid plant genome. *Plant J* **73:** 143–153.

Roux F, Colomé-Tatché M, Edelist C, Wardenaar R, Guerche P, Hospital F, Colot V, Jansen RC, Johannes F. 2011. Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. *Genetics* **188:** 1015–1017.

Schmitz RJ, Ecker JR. 2012. Epigenetic and epigenomic variation in *Arabidopsis thaliana*. *Trends Plant Sci* **17:** 149–154.

Schmitz RJ, Zhang X. 2011. High-throughput approaches for plant epigenomic studies. *Curr Opin Plant Biol* **14:** 130–136.

Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR. 2011. Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334:** 369–373.

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, et al. 2013. Patterns of population epigenomic diversity. *Nature* **495:** 193–198.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463:** 178–183.

Schultz MD, Schmitz RJ, Ecker JR. 2012. "Leveling" the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet* **28:** 583–585.

Shaw RG, Byers DL, Darmo E. 2000. Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. *Genetics* **155:** 369–378.

Shen H, He H, Li J, Chen W, Wang X, Guo L, Peng Z, He G, Zhong S, Qi Y, et al. 2012. Genome-wide analysis of DNA methylation and gene expression changes in two *Arabidopsis* ecotypes and their reciprocal hybrids. *Plant Cell* **24:** 875–892.

Silveira AB, Trontin C, Cortijo S, Barau J, Del Bem LE, Loudet O, Colot V, Vincentz M. 2013. Extensive natural epigenetic variation at a *de novo* originated gene. *PLoS Genet* **9:** e1003437.

Slotkin RK, Vaughn M, Borges F, Tanurdzić M, Becker JD, Feijó JA, Martienssen RA. 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136:** 461–472.

Stefanovic S, Pfeil BE, Palmer JD, Doyle JJ. 2009. Relationships among phaseoloid legumes based on sequences from eight chloroplast regions. *Syst Bot* **34:** 115–128.

Straub SC, Pfeil BE, Doyle JJ. 2006. Testing the polyploid past of soybean using a low-copy nuclear gene—is *Glycine* (Fabaceae: Papilionoideae) an auto- or allopolyploid? *Mol Phylogenet Evol* **39:** 580–584.

Takuno S, Gaut BS. 2013. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci* **110:** 1797–1802.

Teixeira FK, Heredia F, Sarazin A, Roudier F, Boccara M, Ciaudo C, Cruaud C, Poulain J, Berdasco M, Fraga MF, et al. 2009. A role for RNAi in the selective correction of DNA methylation defects. *Science* **323:** 1600–1604.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.

Tuteja JH, Zabala G, Varala K, Hudson M, Vodkin LO. 2009. Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats. *Plant Cell* **21:** 3063–3077.

Vaughn MW, Tanurdzić M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, et al. 2007. Epigenetic natural variation in *Arabidopsis thaliana*. *PLoS Biol* **5:** e174.

Weigel D, Colot V. 2012. Epialleles in plant evolution. *Genome Biol* **13:** 249.

Widman N, Jacobsen SE, Pellegrini M. 2009. Determining the conservation of DNA methylation in *Arabidopsis*. *Epigenetics* **4:** 119–124.

Woo HR, Pontes O, Pikaard CS, Richards EJ. 2007. VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization. *Genes Dev* **21:** 267–277.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591.

Yu A, Lepère G, Jay F, Wang J, Bapaume L, Wang Y, Abraham AL, Penterman J, Fischer RL, Voinnet O, et al. 2013. Dynamics and biological relevance of DNA demethylation in *Arabidopsis* antibacterial defense. *Proc Natl Acad Sci* **110:** 2389–2394.

Zemach A, Kim MY, Silva P, Rodrigues JA, Dotson B, Brooks MD, Zilberman D. 2010a. Local DNA hypomethylation activates genes in rice endosperm. *Proc Natl Acad Sci* **107:** 18729–18734.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010b. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328:** 916–919.

Zhang X, Shiu SH, Cal A, Borevitz JO. 2008. Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. *PLoS Genet* **4:** e1000032.

Received November 22, 2012; accepted in revised form June 5, 2013.