

Patterns of population epigenomic diversity

Robert J. Schmitz^{1,2*}, Matthew D. Schultz^{2,3*}, Mark A. Urlich^{1,2}, Joseph R. Nery², Mattia Pelizzola^{2†}, Ondrej Libiger⁴, Andrew Alix¹, Richard B. McCosh¹, Huaming Chen², Nicholas J. Schork⁴ & Joseph R. Ecker^{1,2,5}

Natural epigenetic variation provides a source for the generation of phenotypic diversity, but to understand its contribution to such diversity, its interaction with genetic variation requires further investigation. Here we report population-wide DNA sequencing of genomes, transcriptomes and methylomes of wild *Arabidopsis thaliana* accessions. Single cytosine methylation polymorphisms are not linked to genotype. However, the rate of linkage disequilibrium decay amongst differentially methylated regions targeted by RNA-directed DNA methylation is similar to the rate for single nucleotide polymorphisms. Association analyses of these RNA-directed DNA methylation regions with genetic variants identified thousands of methylation quantitative trait loci, which revealed the population estimate of genetically dependent methylation variation. Analysis of invariably methylated transposons and genes across this population indicates that loci targeted by RNA-directed DNA methylation are epigenetically activated in pollen and seeds, which facilitates proper development of these structures.

DNA methylation is a covalent base modification of plant nuclear genomes that is accurately inherited through both mitotic and meiotic¹ cell divisions. However, similarly to spontaneous mutations in DNA, errors in the maintenance of methylation states result in the accumulation of single methylation polymorphisms (SMPs) over an evolutionary timescale^{2,3}. The rates of SMP formation are orders of magnitude greater than those of spontaneous mutations, which are in part, probably due to the lower fidelity of maintenance DNA methyltransferases and accompanying silencing machinery^{2–5}. Epiallele formation in the absence of genetic variation can result in phenotypic variation, which is most evident in the plant kingdom, as exemplified by the *peloric* and *colorless non-ripening* variants from *Linaria vulgaris* and *Solanum lycopersicum*, respectively^{6,7}. Although rates of spontaneous variation in DNA methylation and mutation can be decoupled in the laboratory^{8–11}, in natural settings, these two features of genomes co-evolve to create phenotypic diversity on which natural selection can act. In plant genomes, DNA methylation is present in the symmetrical CG and CHG contexts (where H = A, C or T) as well as the asymmetrical CHH context. CG gene-body methylation is a common feature of animal and plant genomes^{12,13}. Regions of plant genomes that contain methylation in the CG, CHG and CHH contexts are indicative of loci that are under control of RNA-directed DNA methylation (RdDM)¹⁴.

Similarly to the limited examples of pure epialleles (methylation variants that form independent of genetic variation), few examples of DNA methylation variants linked to genetic variants are known^{15–17}. Previous studies comparing two accessions of *A. thaliana* or *Zea mays* revealed genome-wide natural variation in DNA methylation^{18–21}, but the dependence of these methylation variants on genetic variants at the population level remains unaddressed. To understand the types and extent of natural DNA methylation variants in *A. thaliana*, epigenomes for genotypically distinct, wild accessions, isolated from throughout the Northern Hemisphere, were determined using MethylC-sequencing¹⁸, (152 methylomes, Supplementary Table 1), RNA-sequencing (144 transcriptomes, Supplementary Table 2) and genomic DNA-sequencing (217 genomes, Supplementary Table 3)¹⁸. Integration of genomic and epigenomic data allowed investigation

into variable methylation states of both CG gene-body methylation and loci targeted by RdDM along with their interactions with genetic variants at the population level.

Population-wide patterns of SMPs

Recent reports of SMPs in a population of essentially isogenic plants indicated that they are major contributors to epigenomic variation^{2,3}. Therefore, we assessed SMP diversity to understand their frequency and patterns throughout a population of genetically distinct accessions. A median of 390,255 SMPs ranging from 92,646 to 527,393 (Supplementary Table 4) were found in the sequenced accessions when compared to the Col-0 reference methylome. On average, CG-, CHG- and CHH-SMPs accounted for 23%, 13% and 64% of all SMPs, respectively. These newly identified SMPs were used to construct an epigenome-based phylogeny and then were compared to a genome-based (single nucleotide polymorphism (SNP)) phylogeny (Supplementary Figs 1–4). A high correlation in the tree structures was specifically observed between CG-SMPs and SNPs as compared to CHG-SMPs or CHH-SMPs and SNPs (Supplementary Table 5).

To determine patterns of SMP diversity, chromosome-wide conservation of methylation states at each SMP was examined by computing a conservation score (Fig. 1a and Supplementary Fig. 5). The methylation state of SMPs in the CG and CHG contexts is biased towards the methylated form at the pericentromere and biased towards the unmethylated form in gene-rich regions (Fig. 1a and Supplementary Fig. 5). Next, the distribution of conservation scores across different features and methylation contexts were plotted genome-wide (Fig. 1b–d). Like the pericentromeric regions, CG- and CHG-SMPs in transposable elements tend to be faithfully methylated throughout this population, whereas CHH-SMPs are largely unmethylated. Unlike CHG and CHH-SMPs, CG-SMPs have a significantly larger amount of methylation at single-copy genes (Fig. 1b–d). Because CG gene-body methylation is associated with moderately expressed genes¹⁹, we postulated that these genes are more active because of the lack of other genes redundant in function. We tested this hypothesis by examining RNA-seq data for 144 of these accessions

¹Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA. ²Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA.

³Bioinformatics Program, University of California at San Diego, La Jolla, California 92093, USA. ⁴The Scripps Translational Science Institute and the Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, California 92037, USA. ⁵Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA.

[†]Present address: Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (ITT), Via Adamello 16, 20139 Milan, Italy.

*These authors contributed equally to this work.

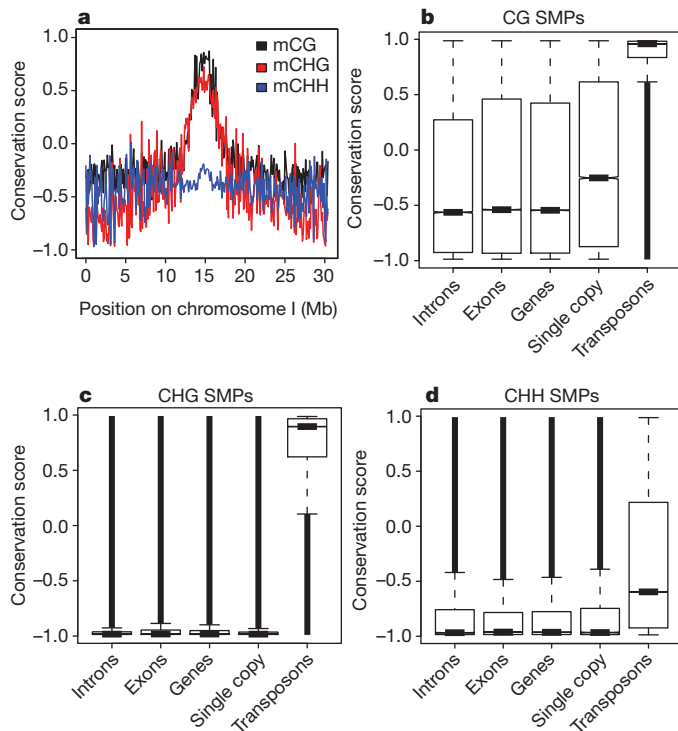


Figure 1 | Population-wide analyses of SMPs. **a**, A plot of the genome-wide distribution of methylation conservation across chromosome I. Mb, megabases. **b–d**, The distributions of SMP conservation scores across various genomic features. Notches in the boxplots represent bootstrap estimated 95% confidence intervals around the medians. mCG, methylated cytosine in CG context; mCHG, methylated cytosine in CHG context; and mCHH, methylated cytosine in CHH context.

at these loci, which revealed the fraction of transcripts where expression was detected—that is, FPKMs > 0 , where FPKM indicates fragments per kilobase of exon per million fragments mapped—was higher in single-copy genes than non-single-copy genes (85% versus 71.8%). Moreover, the median expression level of single-copy genes was also significantly greater (361,814.50 FPKMs of single-copy genes versus 56,107.85 FPKMs of non-single-copy genes), supporting the finding that single-copy genes across the population are more transcriptionally active.

Population-wide variation of DMRs

Spontaneous formation of SMPs represents one form of natural epigenetic variation, but variation also exists in the form of differentially methylated regions (DMRs)^{2,3}. Therefore, we scanned this population for DMRs in the CG context (CG-DMRs) typically found in gene bodies or in the CG, CHG and CHH contexts (C-DMRs) typical of regions targeted by RdDM. Because CG-DMRs and C-DMRs are not mutually exclusive, only CG-DMRs that did not overlap with a C-DMR were defined as CG-DMRs. Hierarchical clustering of accessions based on weighted methylation levels²⁰ (Supplementary Information)—referred to as methylation levels throughout the rest of the paper—of CG-DMRs or C-DMRs revealed patterns across the population that were coincident with certain genomic features (Fig. 2a and b). For example, CG-DMRs are enriched in gene bodies and are present in both unmethylated and methylated states equally throughout the population (Fig. 2a), whereas C-DMRs occur in both gene bodies and transposons (Fig. 2b). Additionally, the C-DMRs in genes are largely unmethylated, which contrasts with the heavy methylation levels that occur in transposons (Fig. 2b). In total, 40,269 CG-DMRs (Supplementary Table 6), with an average size of 321 base pairs (bp; Supplementary Fig. 6) were identified across the population that were enriched in gene bodies and depleted in transposons (Fig. 2a,

Supplementary Fig. 7 and Supplementary Table 7). We also identified 13,485 C-DMRs (Supplementary Table 8), with an average size of 221 bp (Supplementary Fig. 6), that showed enrichment in transposons and depletion in genes (Fig. 2b, Supplementary Fig. 7 and Supplementary Table 7).

This distribution of both CG- and C-DMRs reflects the distribution of genes and transposons along each chromosome and the type of DNA methylation primarily associated with these features, namely CG gene-body methylation versus RdDM. Furthermore, the distribution of methylation levels of CG-DMRs is skewed towards lower levels when the CG-DMR overlaps a gene and towards higher levels when it overlaps a transposon (Fig. 2c, d). The distribution of methylation levels in CG-DMRs resembles the patterns of CG-SMPs for genes versus transposons, as the transposon sequences often contained highly methylated sites or DMRs when compared to genes, supporting the observation that these regions are faithfully repressed by methylation across the population. A comparison of the distribution of methylation levels of the C-DMRs revealed that genes are infrequently methylated at high levels in the population when compared to C-DMRs overlapping transposons (Fig. 2c, d). In this regard, C-DMRs overlapping genes are rare variants in the population, whereas most transposon sequences are almost invariably methylated. Clustering these accessions based on their methylation levels of C-DMRs revealed that accessions that are geographically separated are less likely to cluster together, indicating the potential for underlying genetic structure (Fig. 2e, f). Alternatively, these results could also be obtained for methylation variants that are not dependent on genetic variants if they are stable. Therefore, the observation that accessions cluster based on their geographical distribution is probably due to a combination of both genetic structure and stable methylation variants.

For a subset of accessions examined, methylation data were produced for two tissue types: leaf and mixed-stage inflorescence. Regardless of the tissue used for methylome analysis, when hierarchical clustering was performed using methylation levels of either CG-DMRs (Fig. 2g) or C-DMRs (Fig. 2h), these accessions grouped by their genotype not their tissue type. When the same analysis was applied to RNA-seq data from the same tissues of six accessions, samples clustered on the basis of their tissue type not their genotype (Fig. 2i). Collectively, these data indicate that DNA methylation is less dynamic than gene expression patterns in plants and only plays a role during specific stages of development or cell types^{1,21,22}. Although DNA methylation is more static than transcription, it varies appreciably over an evolutionary timescale, significantly affecting the transcriptional output of specific genes (Fig. 2j, k). Using CG-DMRs that overlap with genes, a positive correlation (Spearman correlation; $P < 2.2 \times 10^{-16}$) between their methylation levels and gene expression levels were found (Fig. 2j), whereas the opposite was true for C-DMRs that overlapped genes, supporting a role for RdDM in transcriptionally silencing these loci (Spearman correlation; $P < 2.2 \times 10^{-16}$; Fig. 2k, Supplementary Figs 8 and 9, and Supplementary Information). Although the role of CG gene-body methylation is still unclear, these data indicate that CG-DMRs that are heavily methylated are associated with higher gene expression levels and can possibly give rise to transcriptional variation.

Linking genetic and methylation variants

Genome sequencing was performed for 217 individuals of which 152 had a matching sequenced DNA methylome. We used the SHORE analysis pipeline²³ to identify SNPs between each accession and the Col-0 genome (Supplementary Information). The identification of SMPs and SNPs that were variable between at least two accessions was used to determine the population-level frequency of these variants, which revealed approximately 70% of CG-SMPs and 41% of SNPs are present at $<1\%$ allele frequency (Supplementary Table 9). These results indicate that a large fraction of SMPs and SNPs are

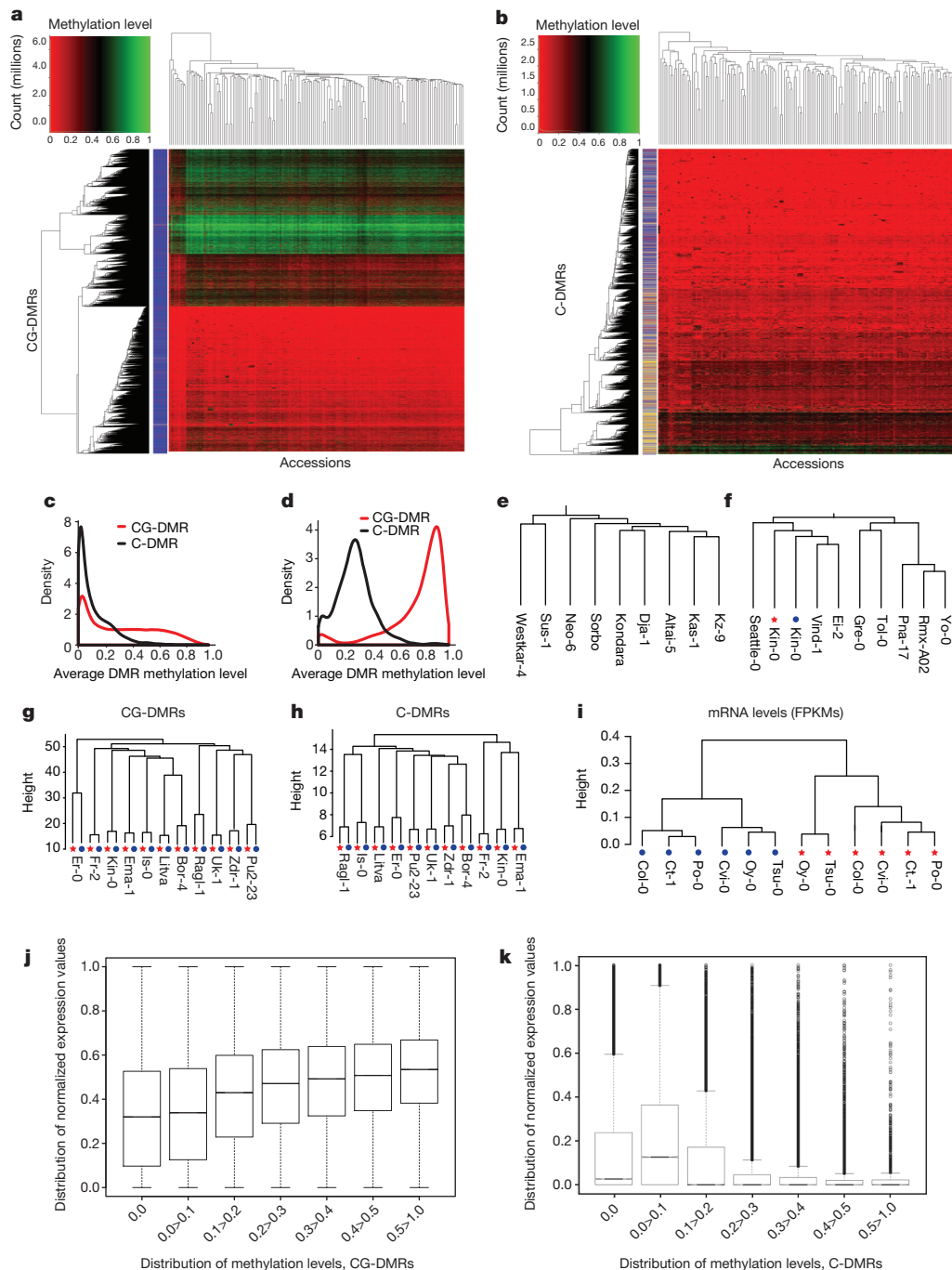


Figure 2 | Population-wide analyses of DMRs. **a, b**, Heatmaps representing the methylation levels across CG-DMRs (**a**) and C-DMRs (**b**) with coverage in all 152 accessions. The coloured row labels on the left of the heatmap indicate what feature a DMR falls in (blue, gene; gold, transposon; red, gene with a transposon inserted in an intron; grey, no feature). Rows indicate genomic locus of DMR and columns indicate accessions. **c, d**, The density (y axis; arbitrary units) and average methylation levels (x axis) of CG-DMRs and C-DMRs in genes (**c**) and transposons (**d**). **e, f**, Asian (**e**) and North American

(**f**) methylome profiles reflected geographical distribution. **g–i**, Dendrograms from hierarchical clustering of CG-DMRs (**g**), C-DMRs (**h**) and mRNA levels (**i**) from accessions that had samples from two different tissues (y axis, height in arbitrary units). Red stars and blue circles indicate leaf and mixed stage inflorescence samples, respectively. **j**, Boxplot representation of transcriptional variation reveals a positive association with higher levels of methylation levels of CG-DMRs. **k**, Increasing methylation levels of C-DMRs are negatively associated with gene expression.

rare variants similar to the results observed for C-DMRs and further indicate that the high epimutation rate for SMPs results in greater numbers of rare alleles. Therefore, even though the spontaneous epimutation rate is at least four orders of magnitude greater than SNPs, the reversible nature of certain SMPs governs their accumulation within populations^{2,3,5}.

Our analysis of gene families that contained the highest number of major effect mutations (see Supplementary Information: NBS-LRR, for defence response; F-box, for protein degradation; and MADS-box

transcription factor, for development) is consistent with previous studies^{24,25}, and these gene families also contained the highest frequency of C-DMRs (Fig. 3a). Furthermore, gene ontology analysis for genes overlapping with C-DMRs identified terms enriched in protein degradation and immune response functions, indicating that these three gene families are equally prone to hypervariable genetic and epigenetic states (that is, they are hypermutable; Supplementary Table 10). Although the frequency of major effect mutations and C-DMRs was similar for these hypermutable families, the remaining

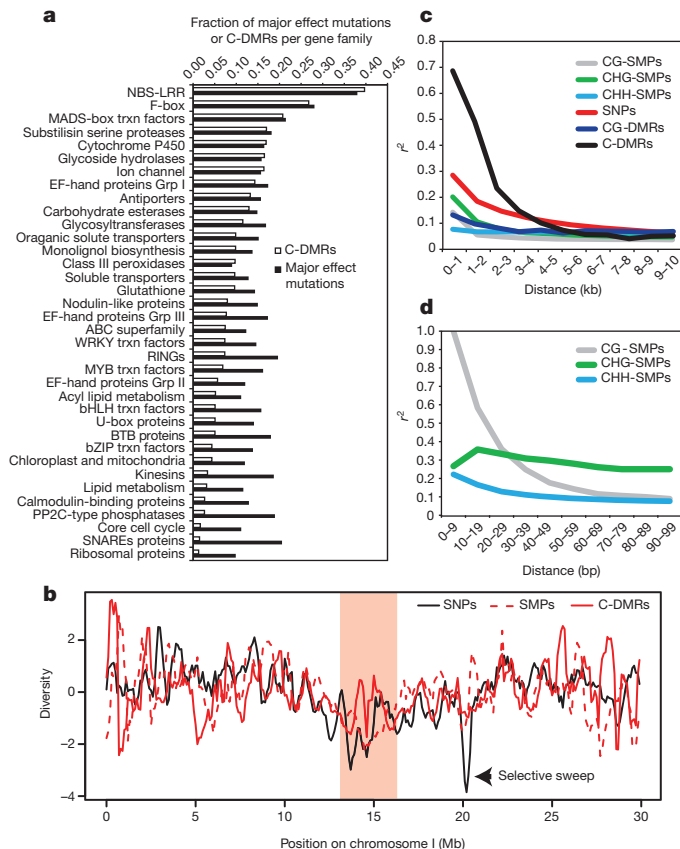


Figure 3 | Population-wide analyses of natural genetic and epigenetic variation. **a**, A comparison of genetic and epigenetic variation across gene families that have the highest incidence of C-DMRs. The y axis indicates the gene family and the x axis represents the relative fraction of major effect mutations or fraction of C-DMRs. Fractions were used to adjust for the size of the gene families. *txn*, transcription. **b**, A plot of SNP, SMP and C-DMR diversity across chromosome I. The shaded pink region indicates the position of the pericentromere. **c**, **d**, Linkage disequilibrium/positional association decay plots for genetic and epigenetic variants.

gene families tested revealed no such co-occurrence of genetic and methylation variation as the frequency of C-DMRs approached zero, whereas the frequency of major effect mutations reached a background rate (Fig. 3a). Apart from the hypermutable families discussed above, there is little relationship between major effect mutations and frequency of C-DMRs. Furthermore, there is no correlation between methylation level and mutation rate in genes containing C-DMRs (Supplementary Table 11). Therefore, the majority of genes targeted by RdDM are functional, and silencing by this pathway may limit their expression to specific stages of development (similar to observations made for transposons²⁶) and/or limit their expression until released from silencing by bacterial infection²⁷, possibly explaining the high frequency of C-DMRs in members of the NBS-LRR family.

To determine the extent to which variation in both DNA methylation and genotype are linked, diversity estimates were calculated for SNPs, all forms of SMPs and C-DMRs (Fig. 3b and Supplementary Fig. 10). A known selective sweep on chromosome I (ref. 26) was identified (Fig. 3b). However, no corresponding depletion was observed for either CG-SMPs or C-DMRs. At this resolution, no correlation between genotype and epigenotype was detected (Supplementary Table 12). Therefore, to understand the relationship and possible dependence of methylation variants on genotype, a higher-resolution positional association and linkage disequilibrium decay analysis was performed using SNPs, CG-SMPs, CHG-SMPs, CHH-SMPs, CG-DMRs and C-DMRs (Fig. 3c, d). Similar to past reports for SNPs, linkage disequilibrium decays within 10 kilobases

(kb), reaching 50% of its initial value at ~2 kb (refs 25, 28; Fig. 3c). This value is similar to the rate of decay for the association amongst C-DMRs (~10 kb), which reaches 50% of its initial value at ~1 kb (Fig. 3c). Surprisingly, the rate of decay for association amongst methylation variants such as CG-SMPs and CG-DMRs occurs rapidly, within 100 bp, which is especially true for genes when compared to transposons (Fig. 3d and Supplementary Figs 11 and 12). Collectively, these data indicate that SMPs and CG-DMRs are truly epigenetic in nature, as they occur largely independent of genetic variation. In contrast, although spontaneous C-DMR formation can occur independent of genetic variation^{2,3}, the linkage disequilibrium and association decay analysis revealed that the presence of C-DMRs may be due, in part, to local genetic variants.

Association-mapping methylation variants

Although there are many mechanisms that can give rise to DNA methylation variation^{2,3,15,29}, the extent to which each plays a role in the formation of the observable methylation variation is unknown. We noted that some sites of known transposition events possessed C-DMRs and posited that these structural variants could be responsible for these differences (Supplementary Fig. 13). To experimentally determine the proportion of C-DMRs with a local structural variant, regions surrounding 92 C-DMRs were PCR amplified and sequenced. Most of these C-DMRs failed to overlap with structural variants; however, structural variations were detected at ~17% (16/92) of the C-DMRs assayed (Fig. 4a, Supplementary Table 13). To better inspect any direct relationship between genetic variants and C-DMRs and to identify potential methylation quantitative trait loci (mQTL)³⁰, we used a genome-wide association technique, EMMAX, as this methodology was successfully used in another similarly sized *Arabidopsis* population^{28,31} (Supplementary Information). Furthermore, we employed two different methodologies to control for false discoveries and found them highly concordant (Supplementary Information). To minimize the number of false positives, we used SNPs that were significant in both methodologies. Application of EMMAX to the 152 accessions

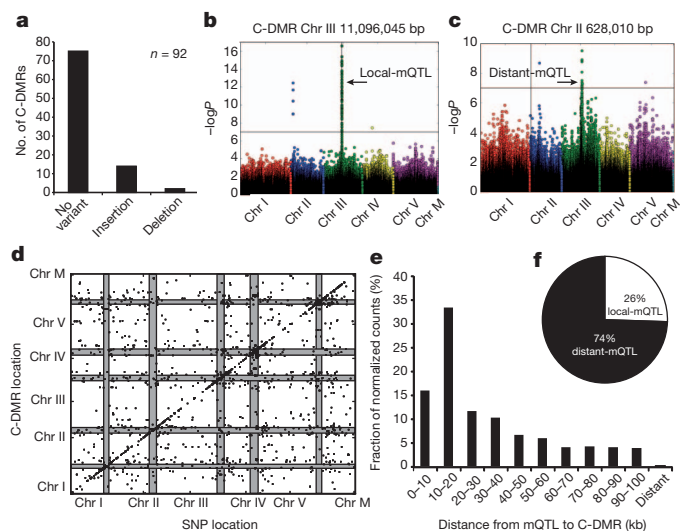


Figure 4 | Association of natural genetic variants and methylation variants. **a**, A summary of the type and number of variants (non-SNPs and small indels) discovered at 92 C-DMRs. **b**, **c**, Manhattan plots with examples of local (**b**) and distant (**c**) mQTL. Chr M denotes the mitochondrial genome. **d**, Distribution of significant mQTL and the C-DMRs with which they associate. Each point represents a significant association between a C-DMR and a block of SNPs. The x axis denotes the genomic location of the SNP block, and the y axis indicates the position of the C-DMR. The pericentromeres on each chromosome are shown as grey bars. **e**, The distribution of distances of mQTL from their C-DMRs normalized for the base space covered by each range of distances. **f**, The ratio of distant mQTL to local mQTL.

with SNP and C-DMR data uncovered C-DMRs that associated with local (Fig. 4b) and distant genetic variants (Fig. 4c) and identified the well-characterized *PAI* epialleles (Supplementary Fig. 14)¹⁵. In total, 2,739 significant mQTL were associated with 1,045 of the 3,023 tested C-DMRs (~35%; Supplementary Figs 15–24).

Of the tested C-DMRs, 377 (~12%) overlap with a genomic locus with which they associate, which is a similar proportion to the number of experimentally determined local variants. We grouped significant mQTL into blocks and plotted the position of these blocks and the corresponding C-DMR in Fig. 4d (Supplementary Information). An enrichment of local mQTL is visible in particular at the pericentromeric regions (Fig. 4d). When corrected for the genome space in which local events can occur, local mQTL account for a larger fraction of the overall results, although the raw number of distant mQTL exceeds the number of local mQTL (Fig. 4f). Furthermore, 61.3% of the local mQTL occur within 30 kb of the C-DMR (Fig. 4e). These association-mapping results also indicated that there were more than twice as many mQTL as C-DMRs. To address whether or not many of the C-DMRs are being controlled in a polygenic manner, we applied the tool MLM³² to the 1,045 C-DMRs with at least one mQTL. Roughly half of the significant C-DMRs reported as polygenic by EMMAX were also reported as polygenic by MLM (Supplementary Fig. 25). Given these results, there are polygenic C-DMRs, although it remains to be determined what types of mechanisms lead to the methylation variation of these C-DMRs. Lastly, applying EMMAX to CG-DMRs resulted in a much lower detection rate of mQTL (Supplementary Table 14 and Supplementary Fig. 26). Together, the above data demonstrate that a considerable fraction of C-DMRs and to a much lesser extent CG-DMRs exist as a result of genetic variation.

All C-DMRs randomly selected for Fig. 4a are rare in the population and had been filtered out before association mapping. Consequently, to determine potential causal variants that are associated with the methylation variants, we PCR amplified 96 C-DMRs associated with a local mQTL. Of these tested loci, 86 successfully amplified and revealed 16 structural variants (Supplementary Table 15), which are similar to the results from the randomly selected C-DMRs (16/92 versus 16/86). As an alternative to structural variation, distant mQTL may result from SNPs, as reported for the *VIM1* variant in the Bor-4 accession³³. Analysis of components with known involvements in DNA methylation within these distant mQTL regions (Supplementary Table 16) revealed *VIM3* and *AGO2* as possible causal loci. Potential causal variants for the remaining local and distant mQTL are likely to involve a combination of either SNPs or structural variations that will undoubtedly be uncovered with future whole-genome assemblies.

RdDM targets are activated in pollen

The mQTL that we have identified revealed that there is an association between some genetic variants and DNA methylation variants, especially for C-DMRs. It is well established that other genetic features, such as repeats, are important for guiding RdDM to target loci. For example, the intergenic sub-telomeric repeats 3' to the *MEDEA* locus and the repeated SINE elements and tandem repeats around the transcription start site of *FWA* are key regulatory sequences for controlling gene expression of these loci^{34,35}. Although these loci are under transcriptional control by genetic elements, these specific elements are present and invariably methylated in every accession examined. Therefore, to understand the role of regions of the epigenome that are less prone to natural epigenetic variation, we searched for loci that contained methylated alleles (methylation level $\geq 10\%$) in >90% of the accessions and identified 283 genes and 255 transposons. The expression of these loci was specifically activated in pollen (Fig. 5a and b). A previous study demonstrated that transposons are activated in the pollen vegetative nucleus, providing a substrate to generate mobile small RNAs, which can be transmitted to the sperm cells (germ line)²⁶.

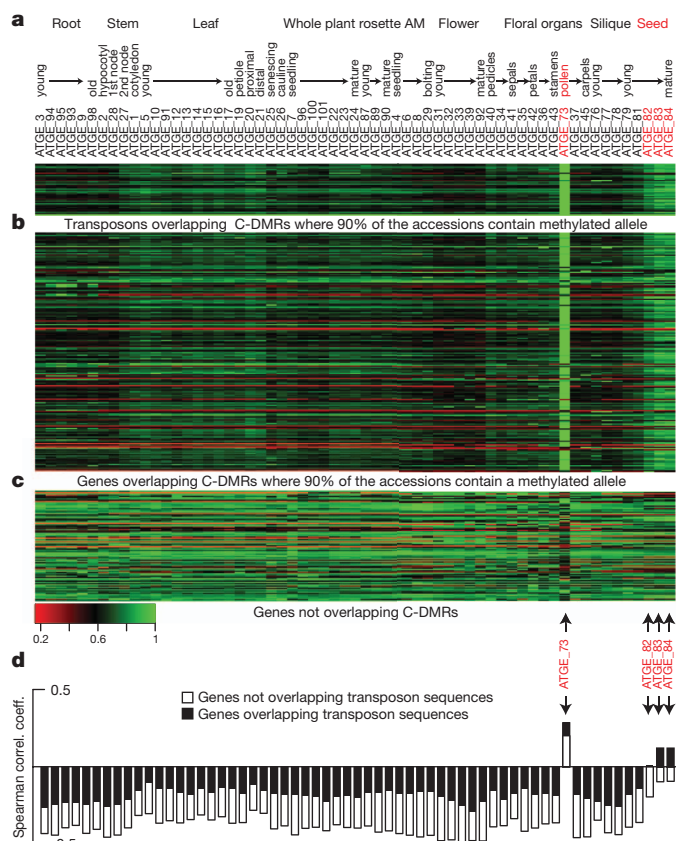


Figure 5 | Epigenetic reprogramming of genes targeted by the RdDM pathway. **a–c**, A heatmap representation of mRNA expression levels throughout a developmental time course²⁷ for transposons (**a**) and genes (**b**) that overlap with C-DMRs where >90% of the alleles are methylated across the population, and genes not overlapping with C-DMRs (**c**). Each row represents a locus with mRNA expression values. Each column represents a different developmental stage. AM, apical meristem. mRNA expression values on colour scale range from low (red) to medium (black) to high (green) and are represented as the average expression level across replicates normalized by the maximum average expression level for each probe. **d**, Correlation test between the nonCG methylation levels and microarray gene expression values for genes targeted by RdDM in Col-0. Open rectangles are genes that do not overlap transposon sequences; filled black rectangles represent genes overlapping transposon sequences.

This mechanism is not restricted to transposons, as we found that protein-coding genes that are under control of RdDM and invariably methylated across this population are also activated in pollen (Fig. 5b). This activation is not a general feature of pollen, as a control set of genes that are not targeted by RdDM are not activated in pollen (Fig. 5c). A closer examination of these invariably methylated genes with gene ontology revealed a significant enrichment for two major categories, cell wall biology and translation (Supplementary Table 17), both related to major functions of pollen development.

Although these invariably methylated loci are under similar epigenetic control as transposons (Fig. 5a, b), it is likely that all RdDM-targeted loci are under the control of this mechanism regardless of their variability within this population. In fact, Col-0 genes targeted by RdDM and their corresponding expression levels are positively correlated (Spearman correlation; $P = 5.81 \times 10^{-27}$) in pollen and seed development (Fig. 5d), whereas all 55 other tissues tested revealed either a negative correlation or no correlation (Fig. 5d, Supplementary Table 18). Furthermore, categories of genes with positive correlations are stronger for loci that overlap transposon sequences (Fig. 5d). These data indicate that these loci have probably come under control of sequences that are evolutionarily silenced, which acts to restrict their

expression to these specific stages of development (Fig. 5d: see Supplementary Information for an expanded discussion of this topic).

Concluding remarks

Natural epigenomic variation is widespread within *A. thaliana*, and the population-based epigenomics presented here has uncovered features of the DNA methylome that are not linked to underlying genetic variation, such as all forms of SMPs and CG-DMRs. However, C-DMRs have positional association decay patterns similar to linkage disequilibrium decay patterns for SNPs and in some cases are associated with genetic variants, but the majority of C-DMRs were not tested by association mapping due to low allele frequencies and could result from rare sequence variants. Our combined analyses of genetic and methylation variation did not uncover a correlation between major effect mutations and genes silenced by RdDM, suggesting that this pathway may target these genes for another purpose. This purpose could be to restrict expression from vegetative tissues similarly to transposons. Another possible purpose of being targeted by RdDM could be to coordinate expression specifically in pollen and in seed to ensure proper gametophytic and embryonic development. Animals also use small RNA-directed DNA methylation and heterochromatin formation mechanisms to maintain the epigenome of the germ line through the use of Piwi-interacting RNAs³⁶. In both plants and animals these small RNAs are derived from the genome of companion cells, which are terminal in nature and can afford widespread reactivation of transposon and repeat sequences as they are not passed on to the next generation. Our study provides evidence that RdDM-targeted genes may have co-opted this transposon silencing mechanism to maintain their silenced state in vegetative tissues and transgenerationally, as well as to ensure proper expression important for pollen, seed and germ line development.

METHODS SUMMARY

MethylC-seq library construction. Genomic DNA was sonicated to ~100 bp using the Covaris S2 system and then purified using Qiagen DNeasy MinElute columns (Qiagen). Each sequencing library was constructed similarly to genomic DNA libraries except the ligation was performed with methylated adapters provided by Illumina. Ligation products were purified with AMPure XP beads (Beckman). Ligated DNA was bisulphite treated using the MethylCode Kit (Invitrogen) and then PCR amplified using Pfu Cx Turbo (Agilent).

Full Methods and any associated references are available in the online version of the paper.

Received 5 July 2012; accepted 30 January 2013.

Published online 6 March 2013.

1. Calarco, J. P. *et al.* Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell* **151**, 194–205 (2012).
2. Becker, C. *et al.* Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**, 245–249 (2011).
3. Schmitz, R. J. *et al.* Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334**, 369–373 (2011).
4. Genereux, D. P., Miner, B. E., Bergstrom, C. T. & Laird, C. D. A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns. *Proc. Natl Acad. Sci. USA* **102**, 5802–5807 (2005).
5. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
6. Cubas, P., Vincent, C. & Coen, E. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401**, 157–161 (1999).
7. Manning, K. *et al.* A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nature Genet.* **38**, 948–952 (2006).
8. Johannes, F. *et al.* Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet.* **5**, e1000530 (2009).
9. Mirouze, M. *et al.* Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature* **461**, 427–430 (2009).
10. Reinders, J. *et al.* Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes Dev.* **23**, 939–950 (2009).
11. Teixeira, F. K. *et al.* A role for RNAi in the selective correction of DNA methylation defects. *Science* **323**, 1600–1604 (2009).
12. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA* **107**, 8689–8694 (2010).


13. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).
14. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Rev. Genet.* **11**, 204–220 (2010).
15. Bender, J. & Fink, G. R. Epigenetic control of an endogenous gene family is revealed by a novel blue fluorescent mutant of *Arabidopsis*. *Cell* **83**, 725–734 (1995).
16. Woo, H. R. & Richards, E. J. Natural variation in DNA methylation in ribosomal RNA genes of *Arabidopsis thaliana*. *BMC Plant Biol.* **8**, 92 (2008).
17. Martin, A. *et al.* A transposon-induced epigenetic change leads to sex determination in melon. *Nature* **461**, 1135–1138 (2009).
18. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
19. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genet.* **39**, 61–69 (2006).
20. 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).
21. Ibarra, C. A. *et al.* Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science* **337**, 1360–1364 (2012).
22. Gehring, M., Bubb, K. L. & Henikoff, S. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* **324**, 1447–1451 (2009).
23. Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033 (2008).
24. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genet.* **43**, 956–963 (2011).
25. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
26. Slotkin, R. K. *et al.* Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136**, 461–472 (2009).
27. Down, R. H. *et al.* Widespread dynamic DNA methylation in response to biotic stress. *Proc. Natl Acad. Sci. USA* **109**, E2183–E2191 (2012).
28. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
29. Ahmed, I., Sarazin, A., Bowler, C., Colot, V. & Quesneville, H. Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in *Arabidopsis*. *Nucleic Acids Res.* **39**, 6919–6931 (2011).
30. Johannes, F., Colot, V. & Jansen, R. C. Epigenome dynamics: a quantitative genetics perspective. *Nature Rev. Genet.* **9**, 883–890 (2008).
31. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genet.* **42**, 348–354 (2010).
32. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genet.* **44**, 825–830 (2012).
33. Woo, H. R., Pontes, O., Pikaard, C. S. & Richards, E. J. VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization. *Genes Dev.* **21**, 267–277 (2007).
34. Kinoshita, T. *et al.* One-way control of FWA imprinting in *Arabidopsis* endosperm by DNA methylation. *Science* **303**, 521–523 (2004).
35. Xiao, W. *et al.* Imprinting of the *MEA* Polycomb gene is controlled by antagonism between MET1 methyltransferase and DME glycosylase. *Dev. Cell* **5**, 891–901 (2003).
36. Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**, 320–324 (2006).
37. Schmid, M. *et al.* A gene expression map of *Arabidopsis thaliana* development. *Nature Genet.* **37**, 501–506 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Luo, R. Down and N. Krogan for critical reading of this manuscript; B. Coullahan for technical assistance with SOLiD RNA-seq; O. Loudet for seeds; and S. Ossowski, K. Schneeberger and D. Weigel for assistance in establishing a variant calling pipeline. R.J.S. was supported by NIH postdoctoral fellowships F32HG004830 and K99GM100000. M.D.S. was supported by an NSF IGERT training grant (DGE-0504645). O.L. and N.J.S. were supported by NIH/NCRR grant number UL1 RR025774. This work was supported by the NSF (MCB-0929402 and MCB-1122246), the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation (grant GBMF3034 to J.R.E.). J.R.E. is a HHMI-GBMF Investigator.

Author Contributions R.J.S., M.D.S. and J.R.E. conceived and designed the study. R.J.S., M.A.U., J.R.N., A.A. and R.B.M. performed experiments. R.J.S., M.D.S., M.P., O.L., H.C. and N.J.S. performed data analysis. R.J.S., M.D.S. and J.R.E. wrote the paper.

Author Information Genome sequence data can be downloaded from NCBI SRA (SRA012474). Processed datasets can be downloaded from NCBI GEO under accession GSE43857 for MethylC-seq data and accession GSE43858 for RNA-seq data and can be viewed at http://neomorph.salk.edu/1001_epigenomes.html and <http://signal.salk.edu/atg1001/index.php>. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and the online version of the paper is freely available to all readers. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.R.E. (ecker@salk.edu).

 This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported licence. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

METHODS

Plant material. Leaf and mixed stage inflorescence tissue were flash frozen in liquid nitrogen, and then the tissue was ground to a fine powder with a mortar and pestle. Leaf tissue was used for genomic and RNA-seq, and the tissues used for each MethylC-seq experiment are listed in Supplementary Table 1. DNA was isolated using a Plant DNeasy kit (Qiagen) following the manufacturer's recommendations. RNA was isolated using the Plant RNeasy kit (Qiagen) following the manufacturer's instructions.

Genomic DNA sequencing library construction. Approximately two micrograms of genomic DNA was sonicated to ~250 bp using the Covaris S2 system using the following parameters: cycle number, 2; duty cycle, 10%; intensity, 4; cycles per burst, 200; and time, 40 s. Sonicated DNA was purified with a PCR Purification MinElute column according to the manufacturer's instructions (Qiagen). Purified DNA was end repaired at room temperature for 45 min using the End-It Repair kit (Epicentre) and purified with a MinElute column (Qiagen). Purified samples were then A-tailed with dATP and Klenow 3'-5' exo minus (New England Biolabs) for 30 min at 37 °C and then purified with a MinElute column (Qiagen). Purified DNA was then used for an overnight ligation to TruSeq barcoded adapters (Illumina) with T4 DNA ligase at 16 °C (New England Biolabs). Ligated fragments were purified twice using Ampure XP purification beads (Beckman) at 1.3× ratio of beads to sample and then PCR amplified for 15 cycles using Phusion High Fidelity DNA Polymerase (New England Biolabs).

MethylC-seq library construction. Approximately one to three micrograms of genomic DNA was sonicated to ~100 bp using the Covaris S2 system using the following parameters: cycle number, 6; duty cycle, 20%; intensity, 5; cycles per burst, 200; and time, 60 s. Sonicated DNA was purified using DNeasy MinElute columns (Qiagen). Each sequencing library was constructed similarly to genomic DNA libraries except the ligation was performed with methylated adapters provided by Illumina. Ligation products were purified with AMPure XP beads (Beckman) at a ratio of 1.8 of beads to sample. Up to 450 ng of ligated DNA was bisulphite treated using the MethylCode kit (Invitrogen) following the manufacturer's guidelines and then PCR amplified using Pfu Cx Turbo (Agilent) using the following PCR conditions (2 min at 95 °C, 4 cycles of 15 s at 98 °C, 30 s at 60 °C, 4 min at 72 °C and 10 min at 72 °C).

RNA-seq library construction. RNA-seq libraries were prepared according to described methods³⁸ except for data collected for Fig. 2i. These libraries were prepared using a TruSeq RNA Sample Kit v2 (Illumina).

Sequencing. Paired-end genomic DNA and single-end MethylC-seq libraries were sequenced using the Illumina GAIIx as per manufacturer's instructions. Sequencing of genomic DNA and MethylC-seq libraries was performed up to 101 and 85 cycles, respectively. Image analysis and base calling were performed with the standard Illumina pipeline. Sequencing of RNA-seq libraries was performed on the SOLiD4 platform (Life Technologies) for 50 bp according to the manufacturer's instructions.

Variant identification. The SHORE package was used to call variants for all of our accessions²³. The following is a list of each submodule and arguments that we ran for the strains: shore import '-v' Fastq; '-e' Shore; '-a' genomic; '-x' forward reads; '-y' reverse reads; '-o' output directory; '-n' 200, shore mapflowcell '-l' TAIR10 Reference; '-f' output directory; '-v' bwa; '-n' 5%; '-g' 3; '-c' 7; '-b' 500000, shore correct4pe '-l' input directory; '-x' 250; '-e' 1001, shore merge '-P' input directory; '-d' output directory, shore consensus '-n' accession_name; '-f' TAIR10 Reference; '-o' output directory; '-i' input directory; '-g' 4; '-q' 7; '-a'; *Arabidopsis* default scoring matrix '-b' 0.51; '-v'; '-r'. Any variant with a quality score of 25 or above was deemed significant. These variants were then substituted into the TAIR10 reference genome to create sample specific references (also referred to as SNP-substituted references) for the mapping of other data sets. In the case of the MethylC-seq mapping, we were able to map, on average, an additional 943,182 reads and allowed us to call an additional 225,894 methylated cytosines (Supplementary Table 19).

MethylC-seq sequencing analysis. Fastq files were aligned to SNP-substituted reference genomes for each accession using Bowtie³⁹, and custom algorithms were used for identification of mC sites as described previously⁴⁰.

RNA-seq data analysis. Bioscope version 1.3 was used to align .csfasta and .qual files to SNP-substituted reference genomes for each accession using default parameters; this allows up to 10 locations per sequenced read. Cufflinks version 1.1 was used to quantify gene expression values using the following parameters: '-F' 0; '-b'; '-N'; '-library-type' fr-secondstrand; '-G' TAIR10.gtf.

Identification of SMPs. We identified SMPs by looking for sites that either were called methylated by our pipeline, or were covered by at least five reads, which we defined as an unmethylated site. Any other site was listed as missing. A SMP was defined as any site with an accession that had a methylation state different between at least two accessions but contained the same sequence as the Col-0 reference genome.

Dendrogram construction. Throughout this work, we present various clustering results of SMPs, SNPs and DMRs. In the cases where these dendrograms are presented with a heatmap, we used the R function heatmap.2 in the gplots package with the default clustering parameters to produce the figure. The dendrograms that lack heatmaps were produced by first generating a distance matrix with R's dist function and passing this matrix to the hclust function, both with their default parameters.

Clustering comparison. To compare the results of the clustering of SMPs and SNPs, we generated distance matrices using R's dist function with the methylation statuses of SMPs as well as the alleles of the SNPs and then compared the Spearman correlation coefficients between the SNP distance matrix and each of the SMP distance matrices (Supplementary Table 5).

Identification of DMRs. All classes of DMRs were identified as previously reported³. CG-DMRs and C-DMRs are not mutually exclusive because C-DMRs are a subset of CG-DMRs. Consequently, for any CG-DMR analyses the subset of C-DMRs were removed.

Definition of methylation levels. Throughout this work, we refer to the level of methylation of genomic regions. To compute this level for a given region, we summed the number of sequenced C bases across all cytosines that were called statistically significantly methylated by our pipeline and divided that sum by the number of sequenced bases covering all cytosines in the given region.

Relationship between DNA methylation and mutation. In an attempt to look at the relationship between mutation and DNA methylation, we calculated the weighted average of DNA methylation and mutation rates across all genes. Genes were defined as entries in the TAIR10 reference GFF file having the word "gene" in the feature column. Methylation levels were calculated as described above, and SNP effects were determined using the snpEff tool (<http://www.ncbi.nlm.nih.gov/pubmed/22728672>) and its *A. thaliana* TAIR10 reference file. We computed two mutation rates, the overall mutation rate and the major effect mutation rate, which we obtained by calculating the fraction of mutations in that gene out of the total number of mutations that were observed in that gene across all accessions. Major effect mutations were defined as mutations that introduced or removed a start or stop codon. The methylation level and mutation rates for each locus were normalized to the maximum value observed at that locus for each measurement type. This normalization yielded measurements on a scale from 0 to 1. We performed a correlation test on these measurements to try and detect a relationship between methylation level and either of the mutation types. As we had no reason to suspect a linear relationship between these variables, we chose to use a Kendall statistic to evaluate the correlation. We detected small but statistically significant relationships between all three of our measurements. Although these results are statistically significant given the small magnitude of the correlation coefficients, we believe that these relationships are at least difficult to interpret but probably not biologically meaningful (Supplementary Table 11).

Enrichment of DMRs in genes and transposons. To determine if CG- and C-DMRs were enriched or depleted in genes or transposons, we performed a binomial test based on the proportions of these features throughout the genome. The results of these tests can be found in Supplementary Table 7.

Linkage disequilibrium/positional association decay analysis. To determine the rate of decay for C-DMRs and CG-DMRs we computed a Pearson correlation coefficient between each pair of DMRs within 10 kb of one another. These coefficients were then separated into 1 kb or 200 bp bins based on the distances between the midpoints of the DMRs. We took the median correlation coefficient of each bin as the rate of decay at a particular distance. In the case of SMPs and SNPs, we used the software package PLINK to determine the association/ linkage disequilibrium between all pairs of sites with a minor allele frequency of 20% and that were within 10 kb of one another. In the case of DMRs, we computed the minor allele frequency by first scoring each accession's DMR as methylated (methylation level $\geq 10\%$) or unmethylated (methylation level $< 10\%$). These scores were binned as in the case of DMRs, and the median value of each bin was taken as the decay rate for a particular distance.

DMR saturation analysis. We estimated how close we are to saturating the discovery of DMRs by randomly subsetting our data and calling DMRs on those subsets (Supplementary Fig. 27). For each of the sample sizes, five random subsets were drawn from the samples and run using the same DMR calling pipeline previously outlined. Although the discovery of new CHH-DMRs seems to be saturated, DMRs in the other contexts remain to be found.

mQTL analysis. Given our small sample size, we made several efforts to control for the number of false positives we undoubtedly found. To this end, we only tested DMRs that had at least 75% (114 samples) of their observations present and at least 10% of their observations over a 10% methylation level (that is, what we defined as a methylated allele). Additionally, we only tested phenotypes that had genomic inflation factors (GIFs) between 0.985 and 1.015. To obtain these GIFs, we calculated the 50th percentile of each tested C-DMR's distribution of *P*-values

as well as the 50th percentile of the distribution of P -values generated by randomly permuting the phenotypes of 20 randomly chosen C-DMRs 10 times (200 permutations in total). These filtering steps left us with 3,023 C-DMRs and 1,877 CG-DMRs to test. We then randomly sampled 1% of the P -values tested and input them to the R package Q-Value⁴¹. The P -value corresponding to a 1% false discovery rate was then used as a cut-off to determine the significance of each association test (we refer to this methodology as the 'Q-value method'). The results for significant SNPs are detailed in Supplementary Table 20. As further validation to ensure that this methodology was working, we compared it to the randomization method outlined elsewhere⁴² (we refer to the following methodology as the 'randomization method'). To this end, we randomized the labels in our genotype matrix (that is, so every sample now had genotypes from a different, randomly chosen sample) and ran EMMAX on the DMRs that had passed our quality control thresholds. Specifically, we ran those DMRs that had at least 10% of their DMRs in the 'methylated' state, at least 75% of their observations present, and a GIF between 0.985 and 1.015. For each DMR tested, we attempted to find the largest P -value that kept the false discovery rate (FDR) under 1%. In this case, we defined the FDR of a given P -value cut-off as the fraction of significant (that is, below the P -value cut-off in question) hits found in the randomized set out of the total number of significant hits found in the randomized and non-randomized sets. The results for significant SNPs are detailed in Supplementary Table 21. We found that the methodology employing Q-value discovered fewer mQTL than the randomization method (Supplementary Table 22), but both methods found a similar proportion of cis and trans mQTL (Supplementary Fig. 28). Furthermore, the Q-value results are nearly a perfect subset of the randomization results (~93% overlap). Consequently, to be conservative, we used the SNPs that overlapped in both methodologies for the analysis in the paper. We grouped these significant SNPs into blocks with the following method.

If a significant SNP lies within 10 kb of another significant SNP combine these two SNPs into a block (that is, the block's start and end are now the positions of these two SNPs). Using this block as a starting point, look for other significant SNPs that are within 10 kb of either end of the block. If such SNPs exist, add them to the existing block, update the block ends with the new SNP, and look for significant SNPs within 10 kb of these new block ends. Repeat this procedure until no significant SNPs can be found within 10 kb of the block ends. These blocks are what we refer to as mQTL throughout the paper. To prioritize candidate loci for follow up studies, we have listed all genes (that is, protein-coding genes defined in the file here (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff)) that fall within the mQTL blocks defined by these significant SNPs, the number of significant SNPs that directly overlap these genes, and whether or not they have been implicated in DNA methylation processes (Supplementary Tables 16 and 23). To better address the validity of mQTL that associated with more than one mQTL, we ran the 1,045 C-DMRs with at least one significant mQTL through the MLM software provided in ref. 32. When evaluating results from this program, we chose the model that minimized the EBIC criterion reported. We used the same P -value cut-off given by the Q-value method above to determine which results were significant and collapsed them in the same fashion as mentioned above. We have included the individual results for the significant SNPs in Supplementary Table 24.

Expression of genes containing DMRs. The lists of C-DMRs and CG-DMRs were used to find the overlap between them and a list of protein coding genes (that is, genes with the 'protein-coding gene' descriptor in the TAIR10 reference annotation file found here (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff)). We then compared the methylation level of these DMRs with the expression levels of the genes they overlapped. We created boxplots of the expression levels for various methylation levels (for example, the expression values for all genes with a DMR that had a methylation level greater than 0.2 but less than 0.3). All the expression values of a locus were divided by the maximum observed value at that locus, so the expression values plotted are the fraction of the maximum expression level observed at a

given locus. It is interesting to note that genes with no C methylation are expressed at a lower level than those that have a methylation level between (but excluding) 0 and 0.1. This dip is due to genes that have no gene body (that is, CG methylation) as has been shown in ref. 20 and is also apparent in these loci (Supplementary Fig. 8). Consequently, we plotted these data again excluding those sites without gene body methylation (that is, 0 now represents loci with no CHG or CHH methylation) and saw the median expression rise to match the median expression level at the 0 to 0.1 level (Supplementary Fig. 9). To make the differences in the medians clearer, we have plotted the median values for the boxplots in Fig. 2k and 2j along with the bootstrap confidence intervals in Supplementary Figs 29 and 30.

Developmental gene expression profiling. Microarray analysis was previously performed for a broad range of developmental stages throughout the plant life cycle³⁷. These data were downloaded from http://www.weigelworld.org/resources/microarray/AtGenExpress/AtGE_dev_gcRNA.txt.zip/at_download/file. These lists of loci that are targeted by the RdDM pathway were matched against probe IDs and the resulting information was extracted. Triplicate data for each developmental time point was averaged and then row normalized according to the developmental time point that displayed the highest expression level and then plotted as a heatmap.

Analysis of local sequence variants at C-DMRs overlapping genes. Primer sets were designed and used for PCR amplification of 92 methylated C-DMRs and for amplification of 86 C-DMRs with local mQTL. Individual PCR products were purified with a PCR purification column (Qiagen) and then sequenced with Sanger sequencing technology. All primer sets can be found in Supplementary Tables 13 and 15.

SMP conservation. To get a global look at the diversity of methylation across each chromosome, we binned cytosine positions into 10 kb windows. To examine the conservation of methylation state at cytosines throughout the genome, we computed a score for each site. Any cytosine that had less than five reads covering it was excluded. We used the following formula to estimate the amount of conservation at each site that was missing data from no more than 50 samples: $(\text{count}(\text{methylated accessions}) - \text{count}(\text{unmethylated accessions})) / (\text{count}(\text{methylated accessions}) + \text{count}(\text{unmethylated accessions}))$. This score reaches its maximum value of 1 when all accessions are methylated and a minimum of -1 when all accessions are unmethylated. We computed this score for each site within a bin (Fig. 1a and Supplementary Fig. 5) and then averaged those statistics together. The distributions of these scores are plotted across features in Fig. 1b, d and e.

Genome-wide running correlation of SMP, SNP and C-DMR diversity measures. To evaluate how the correlation between the diversity measures calculated for SMPs, SNPs and C-DMRs changed across the genome, we calculated diversity measures in the same way as in Fig. 3b, but in 100 kb windows offset by 20 kb instead of 500 kb windows offset by 100 kb. We changed the window size and offset in order to generate more points with which to perform correlation tests. First, we calculated the percentiles of all the diversity measures. Next, we performed a Kendall tau correlation test on these percentiles for all windows that started within 500 kb (upstream or downstream) of a genomic coordinate (listed as the Window Center in Supplementary Table 25). The coefficients from these tests as well as their P -values are listed in Supplementary Table 25.

38. Li, L. *et al.* Linking photoreceptor excitation to changes in plant architecture. *Genes Dev.* **26**, 785–790 (2012).
39. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
40. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
41. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. A* **64**, 479–498 (2002).
42. Breitling, R. *et al.* Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* **4**, e1000232 (2008).