OPEN ACCESS Freely available online

PLOS | GENETICS

Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization

Danelle K. Seymour[9], **Daniel Koenig**[9], **Jörg Hagmann, Claude Becker, Detlef Weigel***

Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany

Abstract

DNA methylation is an ancient molecular modification found in most eukaryotes. In plants, DNA methylation is not only critical for transcriptionally silencing transposons, but can also affect phenotype by altering expression of protein coding genes. The extent of its contribution to phenotypic diversity over evolutionary time is, however, unclear, because of limited stability of epialleles that are not linked to DNA mutations. To dissect the relative contribution of DNA methylation to transposon surveillance and host gene regulation, we leveraged information from three species in the Brassicaceae that vary in genome architecture, *Capsella rubella*, *Arabidopsis lyrata*, and *Arabidopsis thaliana*. We found that the lineage-specific expansion and contraction of transposon and repeat sequences is the main driver of interspecific differences in DNA methylation. The most heavily methylated portions of the genome are thus not conserved at the sequence level. Outside of repeat-associated methylation, there is a surprising degree of conservation in methylation at single nucleotides located in gene bodies. Finally, dynamic DNA methylation is affected more by tissue type than by environmental differences in all species, but these responses are not conserved. The majority of DNA methylation variation between species resides in hypervariable genomic regions, and thus, in the context of macroevolution, is of limited phenotypic consequence.

Citation: Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D (2014) Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. PLoS Genet 10(11): e1004785. doi:10.1371/journal.pgen.1004785

Editor: Brandon S. Gaut, University of California Irvine, United States of America

Received July 18, 2014; **Accepted** September 26, 2014; **Published** November 13, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Raw reads are deposited at the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession number PRJEB6701.

Funding: Supported by the Deutsche Forschungsgemeinschaft (DFG) (SPP1529 Adaptomics), a Human Frontier Science Program (HFSP) Long-Term Fellowship (DK), and the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: weigel@weigelworld.org

[9] These authors contributed equally to this work.

Introduction

Cytosine methylation is a heritable epigenetic modification found in the genomes of organisms spanning the eukaryotic phylogeny [1,2,3,4]. It occurs in three nucleotide contexts, CG, CHG, or CHH (where H is any nucleotide except G) [5], and is enriched in the repeat rich heterochromatic regions of genomes, in nucleosome linkers, and at CG sites in the exon sequences of genes (gene body methylation) [4,6,7,8,9,10,11]. Repeat-localized DNA methylation plays a role in transposon silencing [12,13], but the direct relationship between transcription of protein coding genes and DNA methylation remains unclear. In contrast to repeat methylation, gene body methylation is associated with moderately transcribed sequences [6,7,14,15,16], and has been proposed to stabilize gene expression levels by excluding H2A.Z [17]. Nevertheless, DNA methylation can vary between tissues and environments [18,19,20], and in a handful of cases changes in methylation state contribute to heritable phenotypic variation, although the majority have been linked to structural differences near the affected genes [21,22,23,24,25,26,27]. These observations suggest that DNA methylation may regulate developmental processes and that it could potentiate phenotypic variation during evolution.

Unlike mutational processes acting on DNA sequences, our understanding of the factors contributing to meiotically stable variation in DNA methylation is in its infancy [28]. The different molecular mechanisms governing DNA methylation constitute one factor impacting stability and subsequent inheritance at symmetric and asymmetric sites. In the plant *Arabidopsis thaliana*, initiation and maintenance of methylation at CG and CHG sites is divided primarily between DNA METHYLTRANSFERASE 1 (MET1) and CHROMOMETHYLASE3 (CMT3) [29,30,31]. During DNA replication these two enzymes copy symmetrically methylated cytosines onto the newly synthesized DNA strand using the parental strand as a template [32,33]. Unlike symmetric cytosine methylation, CHH methylation cannot be replicated from the template strand [34]. Instead, methylation at newly synthesized CHH sites is established after cell division by the RdDM RNA-directed DNA methylation pathway through the concerted action of small RNAs (sRNAs) produced from the methylated locus and the de novo DNA methyltransferases DRM1/DRM2 (DOMAINS REARRANGED METHYLTRANSFERASE1/2) [34,35,36,37]. In addition, RdDM-independent asymmetric DNA methylation relies on DDM1 (DECREASE IN DNA METHYLATION1) and CMT2 [38].

Author Summary

DNA methylation is an epigenetic mark that has received a great deal of attention in plants because it can be stably transmitted across generations. However, the rate of DNA methylation change, or epimutation, is greater than that of DNA mutation. In addition, different from DNA sequence, DNA methylation can vary within an individual in response to developmental or environmental cues. Whether altered characters can be passed on to the next generation via directed modifications in DNA methylation is a question of great interest. We have compared how DNA methylation changes between species, tissues, and environments using three closely related crucifers as examples. We found that DNA methylation is different between roots and shoots and changes with temperatures, but that such changes are not conserved across species. Moreover, most of the methylated sites are not conserved between species. This suggests that DNA methylation may respond to immediate fluctuations in the environment, but this response is not retained over long evolutionary periods. Thus, in contrast to transcriptional responses, conserved epigenetic responses at the level of DNA methylation are not widespread. Instead, the patterns of DNA methylation are largely determined by the evolution of genome structure, and responsive loci are likely short-lived accidents of this process.

The extent to which DNA methylation varies at individual sites across generations, or the epimutation rate, has only recently been characterized in isogenic plant lines [39,40]. Repeat-associated methylation was remarkably stable over 30 generations, but some variability arose outside of repeats in euchromatic sequence [39,40]. Changes in DNA methylation accumulated non-linearly, indicating that a subset of methylated sites is particularly prone to spontaneous changes in methylation and, as a result, the absolute DNA methylation differences quickly reach saturation [39,40]. Variation of methylation across generations has been linked to the transgenerational cycling of transposon and repeats between methylated and unmethylated states in the germline [41].

Armed with the knowledge of within-species epimutation rate, the degree of epigenome stability over short evolutionary periods, within a single species, for example, can be addressed [18]. Using *A. thaliana*, intraspecific variation in methylation was surveyed in 140 geographically diverse accessions [18]. Most single site and RdDM-derived regional epimutations were rare, occurring in only a few of the 140 accessions [18]. The lack of intermediate frequency epimutations in these categories is consistent with the view that the vast majority of new methylation variants within a species may only exist for brief periods during evolution. Not too surprisingly, a significant subset of both rare and intermediate frequency RdDM-derived regional epimutations were associated with previously unknown structural variants [18]. Expansion and contraction of repeat-associated sequences leads to intraspecific structural variation; therefore, as a result of RdDM silencing, such structural variants should be linked to methylation variation.

Over longer evolutionary periods, broad similarities in DNA methylation are observed across a variety of genomic features. Large-scale patterns of methylation are shared across flowering plants, including extensive methylation of heterochromatic transposon and repeat-associated sequences [6,7,8,9,10,11] likely due to conservation of the RdDM machinery in plants. Over shorter divergence times, similar levels of gene body methylation have been observed at orthologous genes within the grasses [11,42]. Similarly, in vertebrates, where most of the CG sites in the genome are methylated, absence of methylation at so-called CpG islands is usually found in all species examined [43]. Regardless of organism, the degree of DNA methylation conservation depends on both the evolutionary time scale under consideration and on the genomic feature of interest.

Here we compare at single base resolution DNA methylation in three closely related Brassicaceae - *Capsella rubella*, *Arabidopsis lyrata*, and *Arabidopsis thaliana*. These three species, which diverged about 10 to 20 million years ago [44], vary in genome size and architecture [45,46,47]. Both *C. rubella* and *A. lyrata* have a Brassicaceae typical set of eight chromosomes, while *A. thaliana* has only five chromosomes [48,49]. Both the *A. lyrata* and *C. rubella* genomes are about 50% larger than that of *A. thaliana*, but for very different reasons. Expansion of centromeric, heterochromatic regions has enlarged the *C. rubella* genome, but predominantly euchromatic regions have expanded in *A. lyrata*, driven by insertions of transposable elements (TEs) adjacent to genic sequences [46,47]. Reflecting these differences in genome architecture, the reference genome assemblies represent about 85% of the entire genome in *A. lyrata*, about 75% in *A. thaliana*, and about 60% in *C. rubella* (Table S1) [46,47,50,51,52,53,54]. We show that the difference in genome structure is a major factor influencing the evolution of DNA methylation in these species. Furthermore, while overall DNA methylation is similar between species at many sites, dynamic DNA methylation responses between environments and tissues are rarely conserved. Using a comparative framework we were able to disentangle the contribution of genomic, environmental, and developmental factors to DNA methylation variation between species.

Results

Genome-wide distribution of DNA methylation

Using a factorial design, we subjected seedlings of the inbred reference strains, *A. thaliana* Col-0, *A. lyrata* MN47, and *C. rubella* MTE, to either a control or 23-hour cold treatment and separately harvested root and shoot tissues. This design provides the opportunity to determine conservation of DNA methylation as well as dynamic changes between and within species. In addition to extracting DNA for bisulfite-sequencing in duplicate, we also extracted RNA in triplicate for RNA-seq.

Bisulfite-treated samples were sequenced to an average of 20× strand-specific coverage (Table S2). With this coverage, over 97.5% of the cytosines in the non-repetitive portion of the reference genome of each species could be interrogated (99.5% for *C. rubella*, 97.5% for *A. lyrata*, and 98.7% for *A. thaliana*). With a minimum coverage of three, we confidently estimated methylation rates at two thirds to three quarter of cytosines (62% for *C. rubella*, 65% for *A. lyrata*, and 75% for *A. thaliana*). Sites with significant methylation levels were identified using a binomial test [39]. False positive rates, determined from incomplete conversion of exogenous unmethylated phage lambda DNA, were very low (Table S3).

Global patterns of DNA methylation in *A. lyrata* and *C. rubella* are similar to those reported before for *A. thaliana*, with highest levels in regions near the centromeres, which are populated by TEs and repeats, but contain few genes [6,14,15] (Fig. 1). There is little correlation between DNA methylation density and gene expression at the 500 kb scale (Fig. 1). Centromeric regions are plagued with TEs, and as expected, methylation is found preferentially at sites annotated as residing in TEs (Fig. 2A). Methylation at CHG and CHH sites, which account for over half of methylated sites in all three species, occurs almost exclusively in TEs (Fig. 2A).

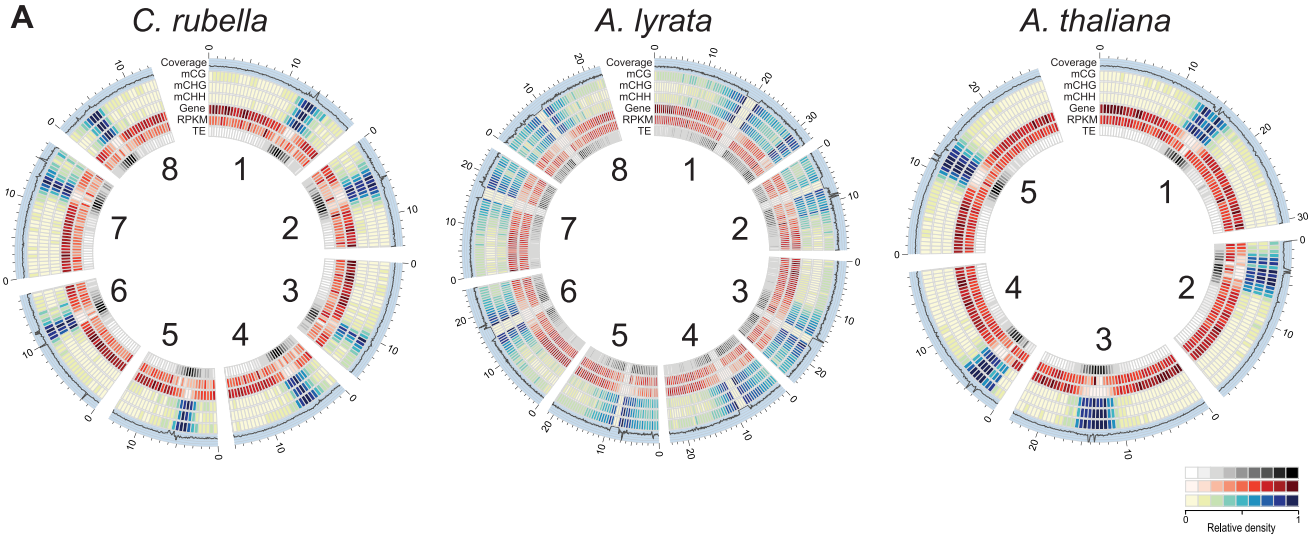

Figure 1. Genomic distribution of DNA methylation. A) Circos plots [74] of *C. rubella*, *A. lyrata*, and *A. thaliana*. Chromosome number is indicated on the inner circle. Data is plotted for 500 kb windows, except for sequencing coverage (100 kb). Gene expression (RPKM) was calculated using the sum of the expression counts from all samples within a species.
doi:10.1371/journal.pgen.1004785.g001

Methylation patterns in the three species reflect their genome architecture. While we mapped a similar number of methylated cytosines in *A. thaliana* and *C. rubella*, consistent with the almost equal size of euchromatic sequences in both species, we identified almost three times as many methylated cytosines in *A. lyrata*, even though its reference genome assembly is only 50 to 75% longer than that of the other two species. The larger number of methylated cytosines in *A. lyrata* has led to an elevation in the methylation rate at a number of genomic features (Fig. 2B). This increase has only occurred at CHG and CHH sites, hallmarks of RdDM at TEs, and is especially evident in introns, correlating with the invasion of introns by TEs in this species (Fig. 2B, C). Almost one third of intronic bases in *A. lyrata* overlap with a TE or repeat, compared to fewer than 10% in the other two species (Fig. 2C), with the expansion found for all TE classes (Fig. 2D). Intron-inserted TEs are frequently found in non-expressed genes (Fig. S1) and are associated with increased methylation in flanking intronic and exonic sequences (Fig. S2), potentially due to pseudogenization or incomplete annotation of repeats. However, when a TE is inserted into the intron of an expressed gene, elevation of CHG and CHH methylation of exon sequences is not evident (Fig. S2, S3). Despite TE expansion in *A. lyrata*, the level of *A. lyrata* gene body methylation is comparable to that of *C. rubella*, which has few TEs in its introns (Fig. 2E). However, species-specific differences in methylation patterns are evident in flanking UTR and intergenic sequence (Fig. 2E). In these regions *A. lyrata* is the most highly methylated in all contexts (Fig. 2E). Depending on context, *C. rubella* displays methylation levels either similar to *A. thaliana* or intermediate between the two other species (Fig. 2E).

Arabidopsis thaliana lost three centromeres relative to *A. lyrata* and *C. rubella*, and this loss has been estimated to account for about 10% of the genome size reduction in *A. thaliana* [46]. Using orthologous genes, it is possible to reconstruct the gene, repeat, and methylation density using the ancestral chromosome positions (Fig. 3). As expected, repeat density and cytosine methylation next to these degraded centromeres is reduced in *A. thaliana*, while gene density is higher (Fig. 3). Particularly notable is the decrease in CG gene body methylation (Fig. 3). Although gene body methylation is positively correlated with gene expression in several species [6,7,14,15,16], gene expression is not noticeably different in these regions between the three species (Fig. 3). Thus, the elimination of centromeres has had a measurable impact on repeat and methylation distribution in *A. thaliana*, but did not strongly affect the expression of ancestrally pericentromeric genes.

Methylated regions are not conserved across species

Methylation of plant genomes is driven to a large extent by TEs, which are silenced via either the sRNA-mediated RdDM pathway [36] or the RdDM-independent pathway which relies on DDM1 [38]. Using a Hidden Markov Model algorithm, we identified methylated regions (MR) in each genome, which have a median length of 300 to 530 bp and cover between 26 and 73 Mb (Table S4). MRs are preferentially found in heterochromatic sequence next to centromeres, as they are enriched for TEs (Fig. S4, Fig. 4A). Since TEs are rapidly turned over, we expected MRs to be only poorly conserved. To test this assumption, we identified nearly 60 Mb of sequences with a 1:1:1 relationship in whole-genome alignments (Table S5) [47]. Less than 1% of the MR space is contained in the alignable portion of the genomes (Fig. 4B). In the rare cases where an MR spans alignable sequences, such sequences are almost always methylated in only one of the three species (Fig. 4C). We conclude that DNA methylation targets primarily the variable portion of the genome, which is subject to species-specific expansion and contraction of TEs.

To determine whether specific orthologs tend to be associated with methylation in all species, even in the absence of MR sequence conservation, we analyzed orthologs that contained a MR overlapping or within 1 kb of their coding region. Again, we found that the presence of MRs is rarely conserved (Fig. 4D, Table S6), although MR sharing is seen more often than expected by chance (Fig. 4D, Tables S7, S8). This could, however, be simply due to genes near centromeres being more often associated with MRs because they are in an MR-rich genome environment.

Conservation of CG gene body methylation

In contrast to RdDM of TEs and other repeats, the function of CG gene body methylation is still enigmatic, although it correlates

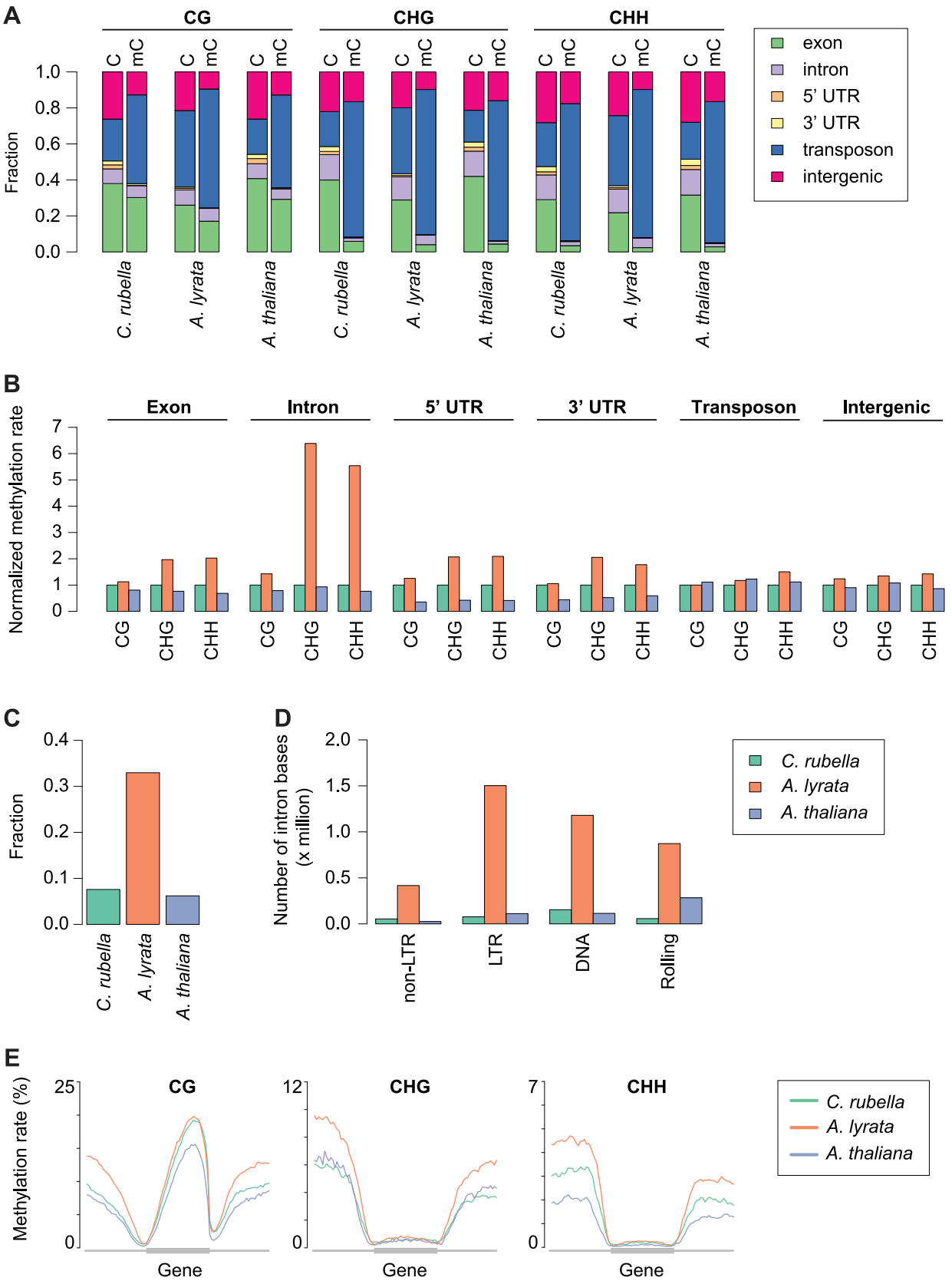

Figure 2. Impact of repeat expansion on DNA methylation at genomic features. A) Feature annotation of all cytosines and methylated cytosines. Annotations are shown for all three contexts. B) Genome average of methylation rates for each genomic feature. Methylation rates are normalized to the outgroup species *C. rubella*. C) Fraction of intron bases annotated as transposable element or other repeat sequence. D) Total number of intron bases (millions) that are annotated as a particular transposable element class. E) Methylation rate distribution across gene bodies of

orthologous genes and flanking sequences (1.5 kb up - and downstream). Orthologs that lacked methylation in both their gene body and flanking sequences were excluded. Distributions are plotted by context.
doi:10.1371/journal.pgen.1004785.g002

positively with gene expression and negatively with mean normalized expression variance, or the coefficient of variation, across tissues and treatments (Fig. S5) [6,7,14,15,16,17]. CG gene body methylation is found in the majority of genes (Table S9), and its rate is highly correlated between orthologs, while CG methylation up- and downstream of genes is much less correlated (Fig. 5).

CHG and CHH methylation in gene bodies is often indicative of transcriptionally inactive pseudogenes, paralogs, or transposons wrongly annotated as protein coding genes [14,15,55]. Between 10 and 20% of genes exhibit CHG or CHH methylation, most of which were not expressed in our samples (Table S9). Genes with CHG or CHH methylation are underrepresented in the orthologous gene set, where their fraction drops to less than half of their fraction among all genes, supporting the assertion that CHG and CHH methylation point to a tendency toward pseudogenization (Table S9). Moreover, CHG and CHH methylation are generally not conserved, suggesting that these marks arise in a lineage-specific fashion.

Site-specific gains and losses of methylation in euchromatic sequence

We used the cross-species alignments to identify 15.1 million conserved CG, CHG and CHH sites, which are located particularly in exons (Fig. 6A, Table S5). Although only a small portion, 2%, had significant methylation, most were shared between at least two species, with *A. thaliana* having the fewest methylated sites, reflecting the general decrease in global DNA methylation in this species (Fig. 6B–D, Table S10). Sites

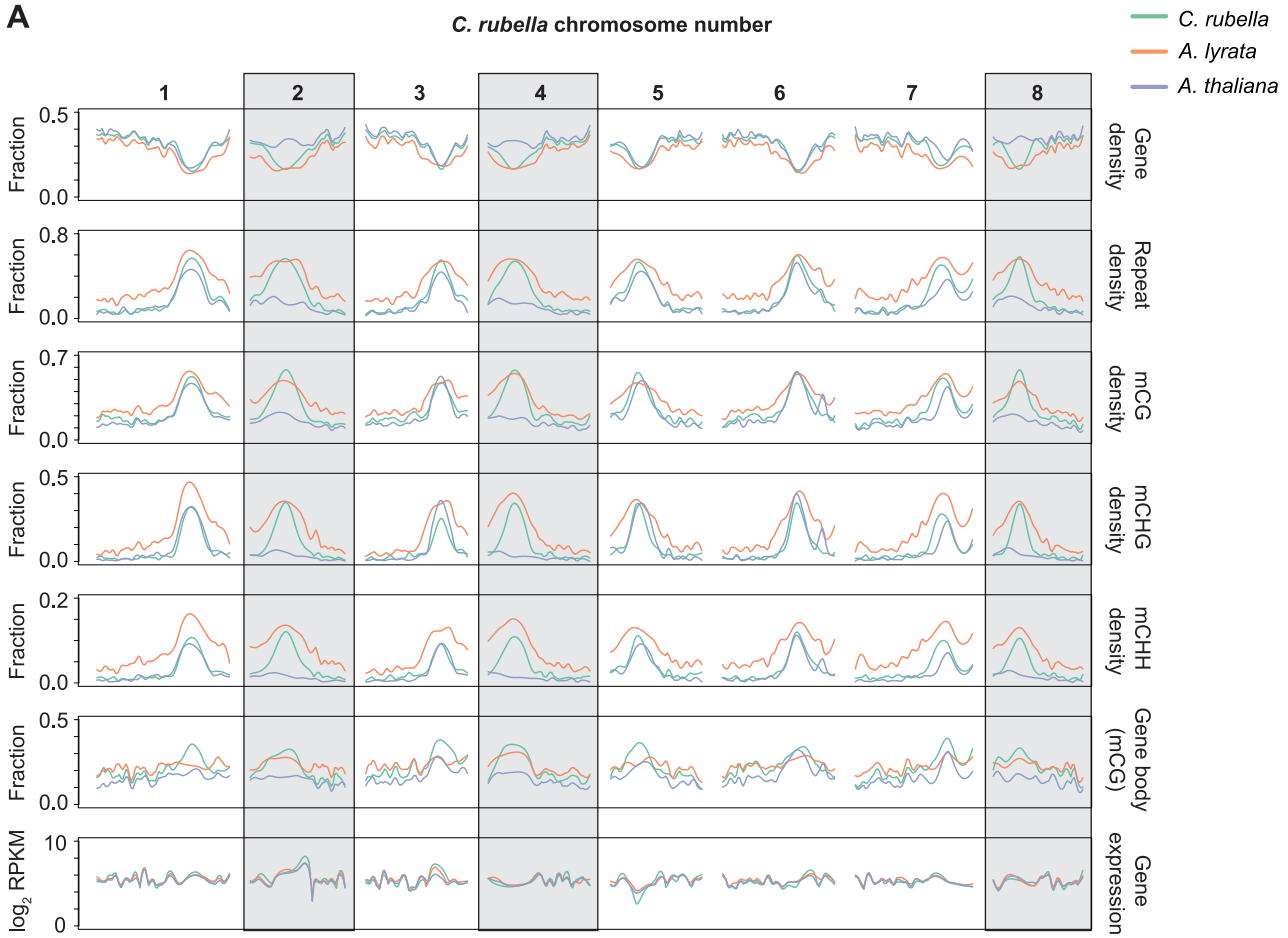

Figure 3. Centromere loss impacts DNA methylation in *A. thaliana*. A) Orthologous genes, anchored on the *C. rubella* genome, were used to calculate several statistics to investigate the impact of centromere loss on DNA methylation in *A. thaliana*. *Capsella rubella* centromeres 2, 4, and 8 (grey boxes) were lost during chromosomal fusion events that occurred on the branch leading to *A. thaliana*. Gene density, repeat density, and methylation densities were calculated for a 20 Kb window centered on the midpoint of each orthologous gene (10 kb up- and 10 kb downstream). Gene density and repeat density were calculated as fractions of each 20 kb window annotated as either a gene (ATG to STOP) or a repeat. Methylation densities were calculated as fractions of cytosines methylated in each context. Gene body methylation and gene expression (RPKM) were calculated for each ortholog. Gene body methylation was calculated as the fraction of methylated CG sites in a gene (ATG to STOP). Gene expression data from all samples within a species were used to calculate the RPKM values. For each statistic, local linear regression was performed to smooth the data in 250 kb bins. Smoothing parameter was relative to chromosome length.
doi:10.1371/journal.pgen.1004785.g003

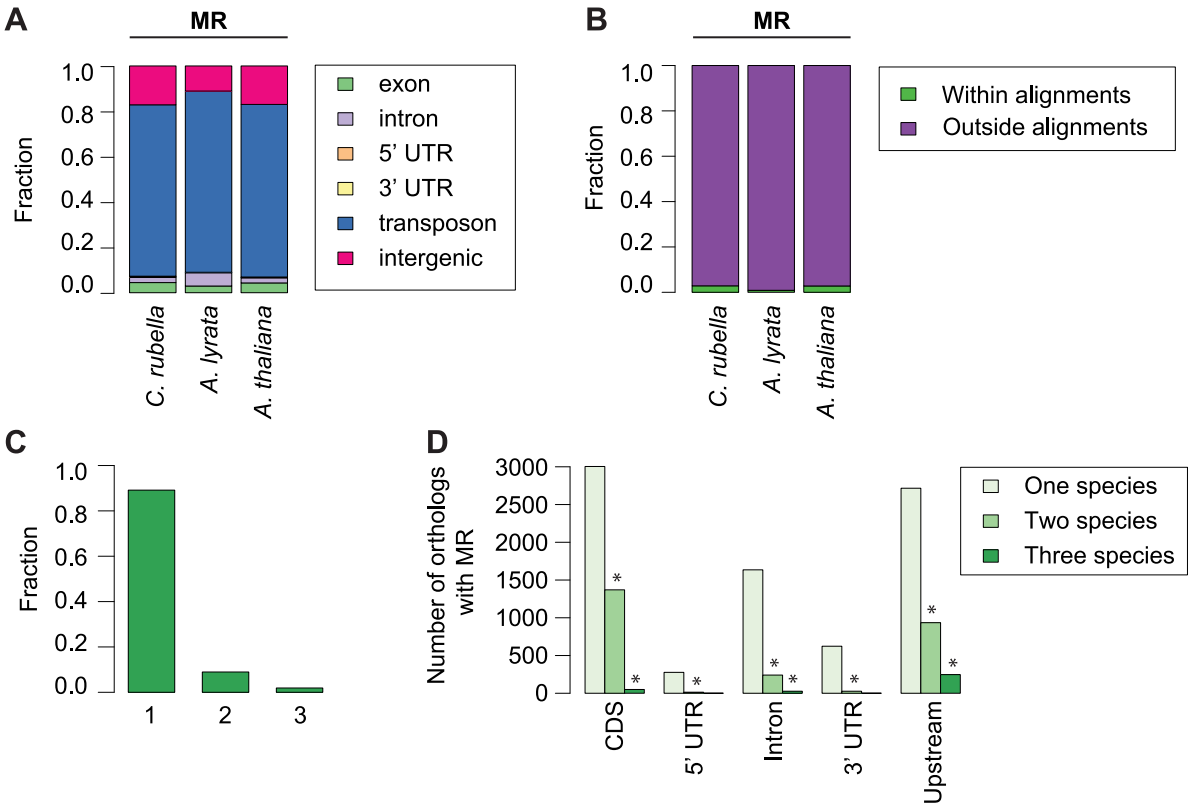

Figure 4. Conservation of methylated regions (MR). A) Annotation of all bases in MRs. B) Fraction of bases in MRs that occur either within or outside of the three-way whole genome alignments. C) Fraction of MR bases found within three-way whole genome alignments that occur in one, two, or three species. D) Conservation of MRs in the absence of sequence alignments. The total number of orthologous genes overlapping an MR in one, two, or three species is given, with location of MR overlap separated by genomic feature. Upstream region was defined as 1 kb before the start codon. Asterisk indicates two or three-way sharing of MRs that exceeds permutation values.
doi:10.1371/journal.pgen.1004785.g004

methylated in multiple species are further enriched in exons, with very few of these conserved sites being CHG or CHH sites (Fig. 6B,C, Fig. S6).

Sites that differ in methylation between species can be used to study gain and loss of methylation. We consider sites that are methylated only in a single species as lineage-specific gains, and absence of methylation in only one species as lineage-specific losses. We found that the number of gains and losses reflect the differences in genome architecture between the three species (Fig. 6 B,D). The many methylation losses in *A. thaliana* appear to be the result of genome shrinkage, and this species has also the fewest gains. In contrast, *A. lyrata* has the most gains, likely reflecting recent TE expansion (Fig. 6 B,D). The density of variable sites across the genome (in 10 kb windows) illustrates that gains and losses are not randomly distributed (Fig. 6D). Species-specific gains, which occur in all three sequence contexts, are concentrated in a subset of windows that are strongly enriched for TEs (Fig. 6D,E), but are also frequently found in exons (Fig. S6). That methylation gains are particularly likely in first and last exons suggests that methylation spreading from nearby TEs makes an important contribution to newly methylated sites, regardless of TE class (Fig. 6F, S7) [56,57,58].

Lineage-specific losses are more evenly distributed, without any signature of TE association. In addition, sites that are conserved in not only two, but all three species occur across a similar spectrum of genomic features (Fig. S6). Together these results indicate that unlike gains, losses occur in a random fashion, with the proviso that there is an overall global loss of methylation in *A. thaliana* (Fig. 6D). Though centromere elimination contributes to the different methylation pattern in *A. thaliana*, this explains only a minority of these losses (Fig. S8). It appears more likely that they are caused by the global reduction in TE content. We also attempted to understand what factors might contribute to conservation of DNA methylation over time. Sites found in more than one species are enriched in exons of conserved length and are more frequent in the center of exons (Fig. S9, S10).

Methylation variation within individuals

Because several studies have shown that DNA methylation can change between tissues and in response to external stimuli [19,20], we wanted to address whether these responses are conserved. Principal component analysis on the four types of samples, control shoots, cold-treated shoots, control roots and cold-treated roots, for all three species according to global RNA-seq measurements revealed that tissue is the most important factor, with over 7,000 genes being differentially expressed between roots and shoots (Fig. 7A, S11). Tissue-specific differences in gene expression are the largest source of expression variance in this data set (Fig. 7A). In contrast, species is the most important factor for differences in DNA methylation and explains 80% of the variance in our data (Fig. 7B, Fig. S12). Moreover, PC2 places *A. lyrata* closest to *C. rubella* instead of its congener *A. thaliana*, reflecting the methylation losses in *A. thaliana* (Fig. 7B).

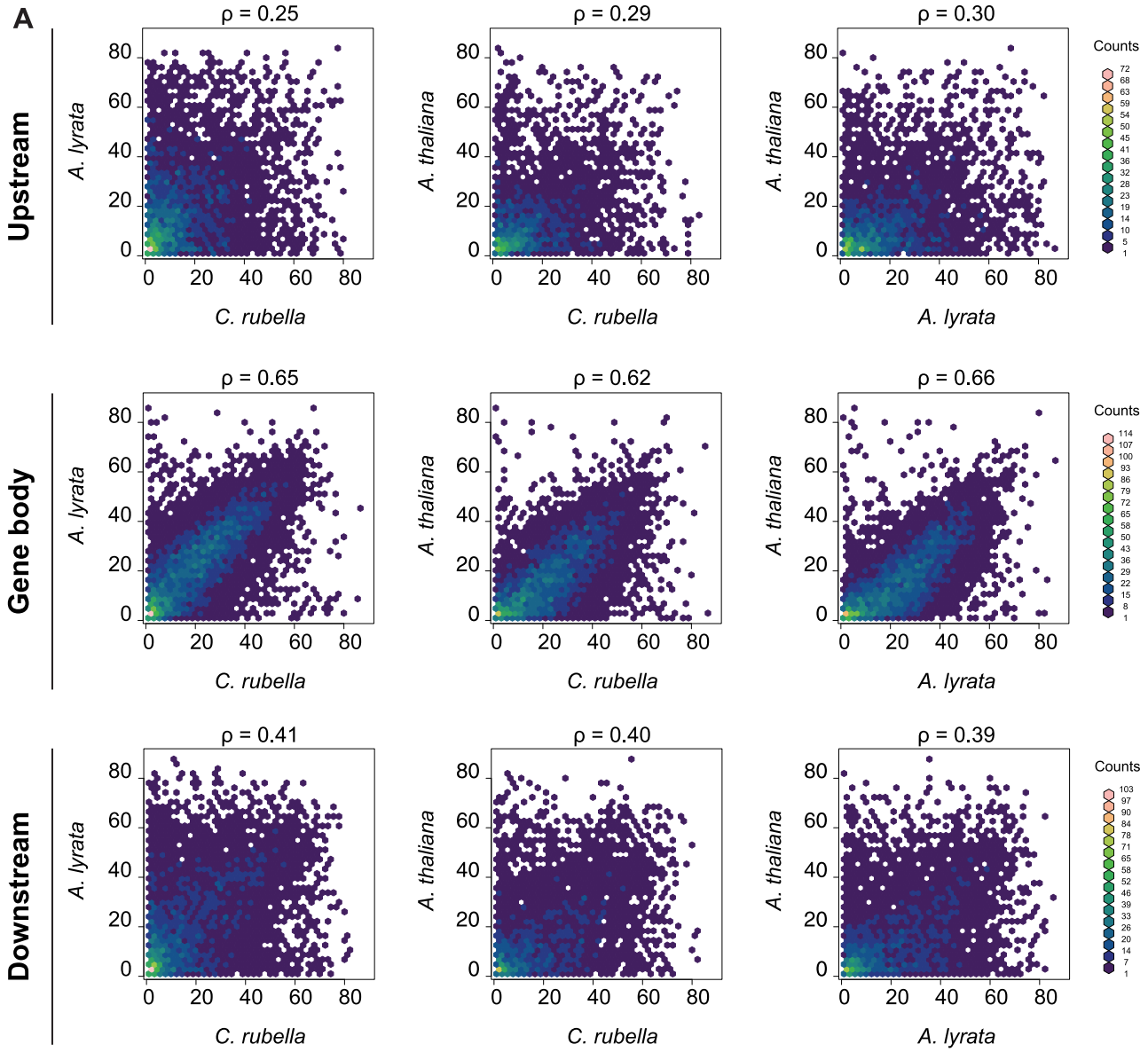

Figure 5. Methylation rates at orthologs. A) Pairwise comparison of the average methylation rates at orthologs. Average methylation rate was calculated as the average of all CG sites in the feature, including non-methylated CG sites. Pairwise comparisons are shown for upstream regions (1.5 kb), gene bodies, and downstream regions (1.5 kb). Spearman rank correlation coefficient (ρ) is included for each comparison.
doi:10.1371/journal.pgen.1004785.g005

To evaluate the degree to which within-species DNA methylation changes are conserved, we first estimated significant differential methylation at site and region levels. Four biologically appropriate comparisons were performed for each species to minimize multiple testing problems. Two tests identified differentially methylated positions (DMPs) between roots and shoots, and two tests identified DMPs between cold and control conditions regardless of tissue type. In each species, ten times as many DMPs were found between tissues than between treatments (Figure 8A, Table S11). Similar to DMPs, 20 to 50 times as many differentially methylated regions (DMRs) were detected between tissues than between treatments (Fig. 8B, Table S12).

Importantly, DMPs and DMRs do not necessarily coincide (Fig. S4, S13). DMPs in all contexts are rarely found within DMRs, indicating that significant regional changes in methylation are not just the extension of single base differences (Fig. 8C). CHG and CHH DMPs reside mainly within MRs (Fig. 8C); since these are almost exclusively found in the non-alignable portions of the genome, including TEs (Fig. 4A, Fig. 8D), the positions of DMPs and DMRs are typically not conserved between species (Fig. 8E). In the rare case that DMPs or DMRs can be found in the portion of a species' genome that can be aligned with the genomes of the other two species (Fig. 8E), they are only variable in a single species (Fig. 8F). Methylation variation at both the site and region level is therefore not conserved across species.

In the absence of sequence conservation at DMRs, we looked for conservation of their presence at orthologous genes. When only considering orthologs, fewer than 700 genes coincide with a DMR (405 in *C. rubella*, 652 in *A. lyrata*, and 221 in *A. thaliana*) (Table S13). Orthologs only rarely shared the presence of an overlapping or adjacent DMR, similar to what we see for MRs. Despite the rarity of such cases, they occur more often than expected by

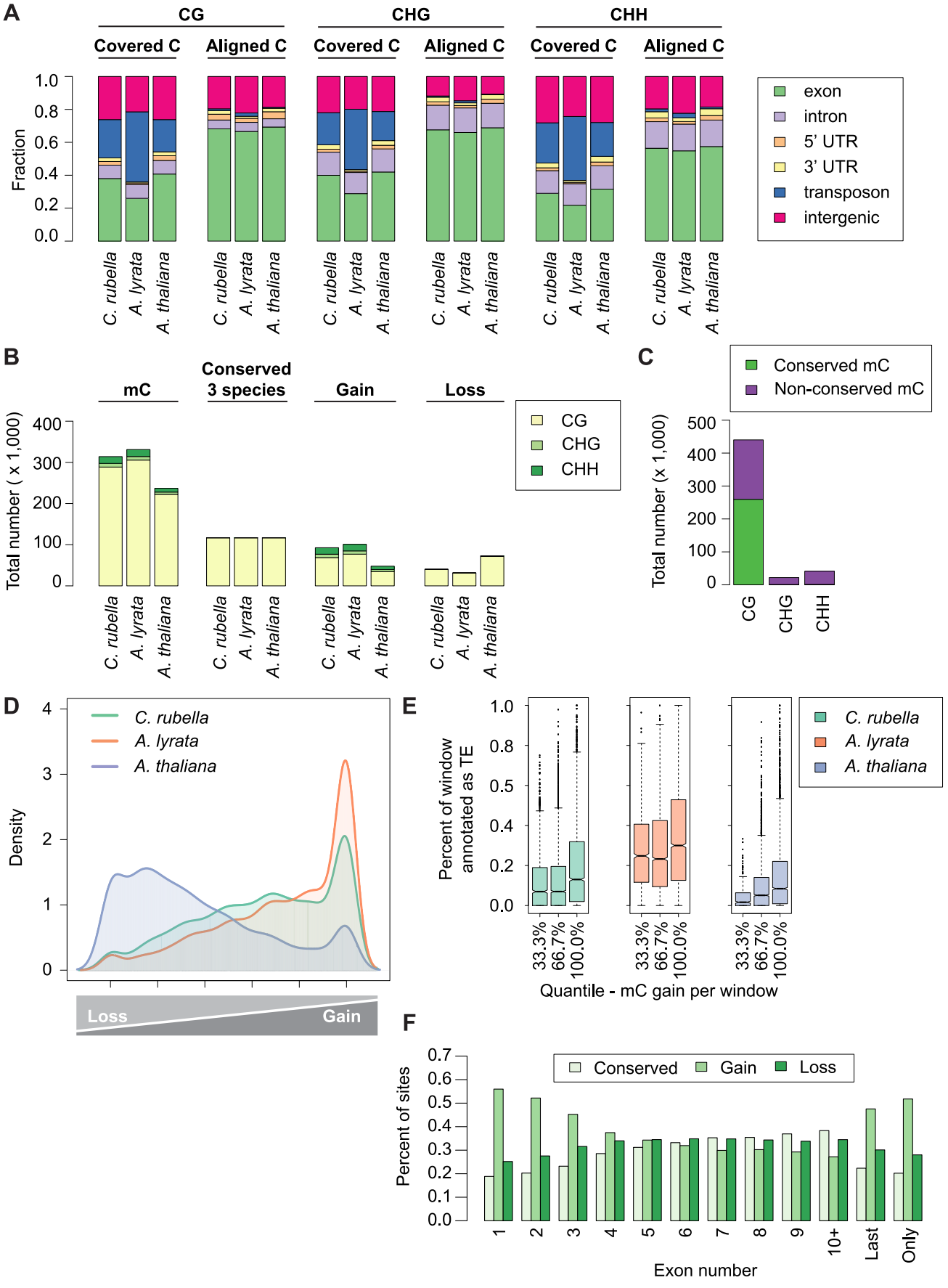
A
CG
CHG
CHH
Covered C
Aligned C
Fraction
exon
intron
5' UTR
3' UTR
transposon
intergenic
C. rubella
A. lyrata
A. thaliana
B
mC
Conserved 3 species
Gain
Loss
Total number (x 1,000)
CG
CHG
CHH
C
Conserved mC
Non-conserved mC
Total number (x 1,000)
D
Density
Loss
Gain
E
Percent of window annotated as TE
33.3%
66.7%
100.0%
Quantile - mC gain per window
F
Conserved
Gain
Loss
Percent of sites
Exon number
1
2
3
4
5
6
7
8
9
10+
Last
Only

Figure 6. Site-level comparison of methylation. A) Annotation of all cytosines within a species (covered C) compared to the annotation of cytosines found in the three-way whole genome alignments (aligned C). B) Total number of mC by context for aligned site classes. Site classes are as follows: mC - methylated sites within a species. Conserved (3 species) - sites that are methylated in all three species. Gain - sites that are methylated in a single species. Loss - sites that have lost methylation in a single species. C) Total number of conserved mC and non-conserved mC by context. D) Density plot describing the distribution of variable sites in the genome (10 kb windows). For each window the following statistic was calculated: species-specific methylation gains/sum of species-specific methylation gains and losses. E) Windows with a high density of gains have more transposons and repetitive sequences. Density of transposons plotted against density of methylation gains (10 kb window). F) Methylation gains are enriched at the beginning and end of genes. Fraction of mC in each site class is plotted by exon position in a gene.
doi:10.1371/journal.pgen.1004785.g006

chance for a subset of genomic features and species comparisons (Fig. S14, Table S14, Table S15). Lack of sequence conservation together with minimal overlap of DMR presence at orthologs supports the transitory nature of methylation variation during genome evolution.

We also asked whether differential methylation in or near coding sequences is correlated with changes in gene expression. DMP and DMR overlap with genes was analyzed separately for those that overlapped with exons, introns, 5′ UTRs, 3′ UTRs and 1 kb upstream regions (Table S13, S16). DMPs occur in many genes in all three species, and most of them are expressed in our samples (9,631 in *C. rubella*, 12,216 in *A. lyrata*, and 6,345 in *A. thaliana*), but there is no evidence for correlation between DMPs and gene expression. This holds true for tissue as well as treatment DMPs (average Spearman rank correlation coefficient tissue = −0.04, treatment = 0.02, Table S17). Only a small number of DMRs overlap with expressed genes (529 in *C. rubella*, 801 in *A. lyrata*, and 284 in *A. thaliana*). Again, there is no correlation with gene expression (average Spearman rank correlation coefficient for CG DMRs = −0.16, CHG DMRs = −0.06, CHH DMRs = 0.00, Table S18).

Although DMPs and DMRs are not conserved across species, there is consistently more variability between root and shoot samples at a number of genomic features. Importantly, the methylation profile across transposons is quite different between tissues. Transposons are consistently more highly methylated in all sequence contexts in shoots (Fig. 9A). A similar trend is apparent for CHG and CHH sites in intergenic regions in *A. lyrata*, reflecting that TEs are closer to genes in this species (Fig. 9B) [46].

Discussion

DNA methylation is an ancient epigenetic modification that appears in the genomes of organisms throughout the eukaryotic phylogeny [1,2,3]. This mark is associated with a number of cellular processes including transposon silencing and host gene regulation, but the cause-and-effect relationship between gene expression and DNA methylation remains unclear [6,7,12,13,14,15,16]. From an evolutionary standpoint, it is useful to consider methylated cytosines from two differing perspectives, either as a non-canonical nucleotide or as a molecular phenotype akin to transcription, and each perspective has important implications for the interpretation of its evolutionary dynamics.

Dynamics of DNA methylation as a molecular phenotype

As a molecular phenotype, many characteristics of DNA methylation are conserved between the species we examined. DNA methylation is generally associated with the repeat-dense sequences found in the centromeres, with CG methylation being in addition present at high levels in exonic sequences

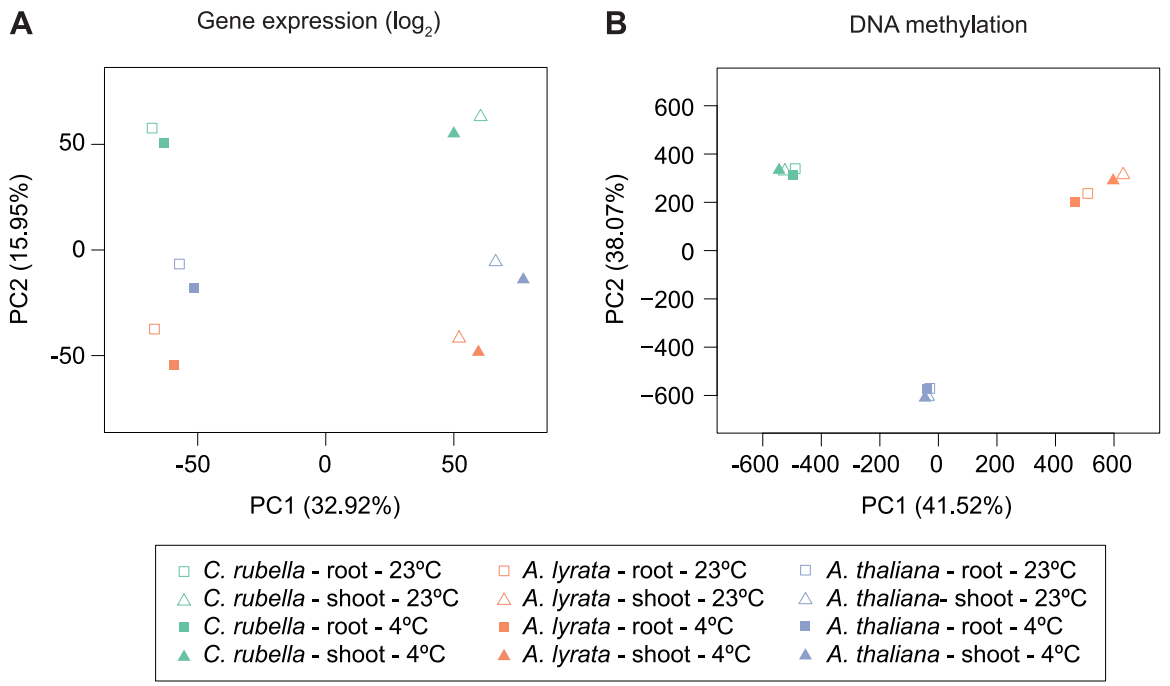

Figure 7. Species gene expression and mC relationships. A) Principal component analysis on fitted gene expression values (log_2) and B) mC rates at aligned methylated positions. All contexts are considered (see Fig. 6B,C and Table S10 for further description of mC sites).
doi:10.1371/journal.pgen.1004785.g007

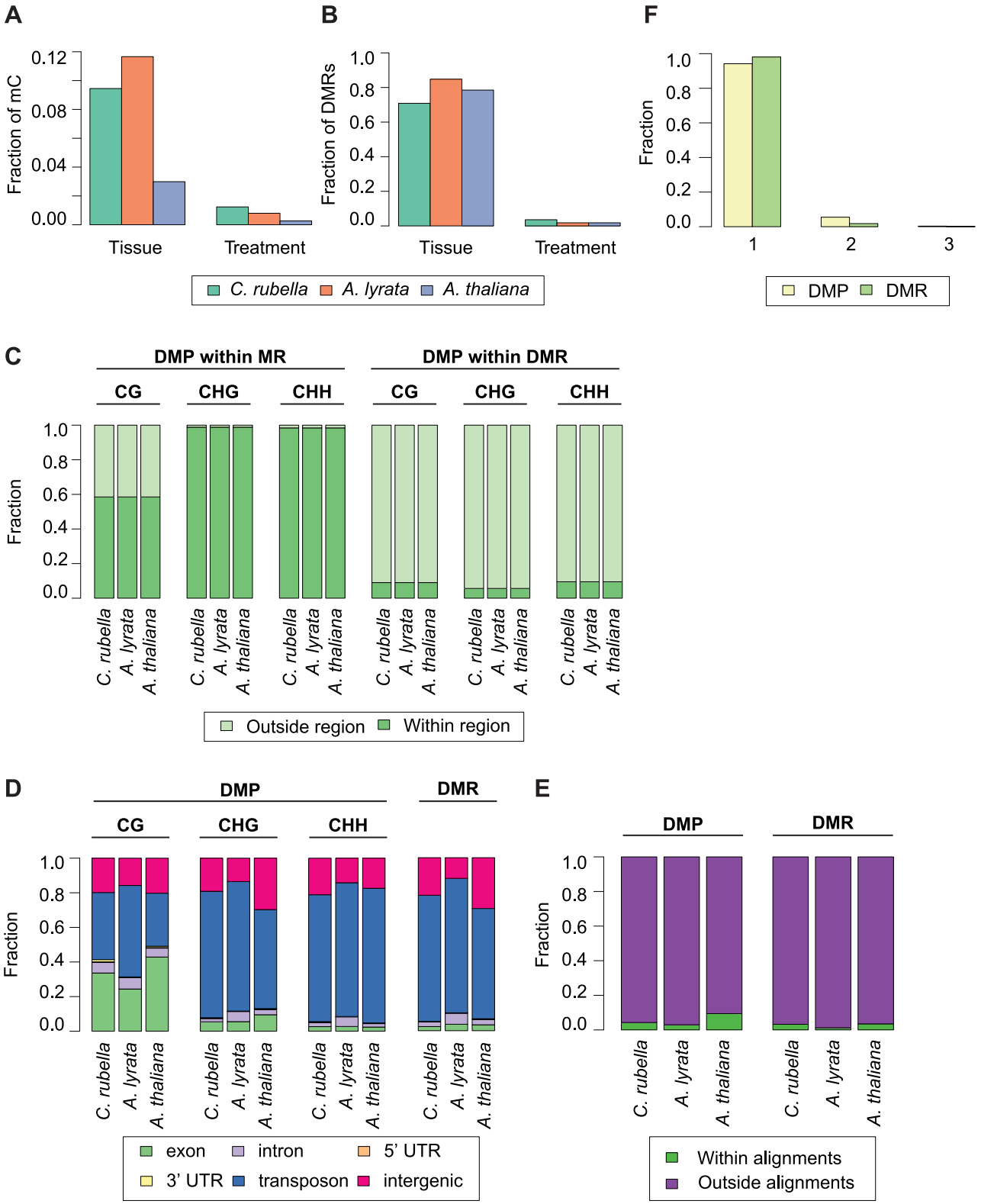

Figure 8. Intraspecific variation in DNA methylation. A) Fraction of mC that are variable between either tissue (root and shoot) or treatment (23°C and 4°C) comparisons. B) Fraction of DMRs that are variable between either tissue (root and shoot) or treatment (23°C and 4°C) comparisons. C) Fraction of DMPs in each context that reside either within a MR or DMR. D) Feature annotation of DMPs by context and DMR bases. E) Fraction of DMPs and DMR bases found within three-way whole genome alignments. F) Fraction of DMPs and DMR bases found within three-way whole genome alignments that occur in one, two, or three species.
doi:10.1371/journal.pgen.1004785.g008

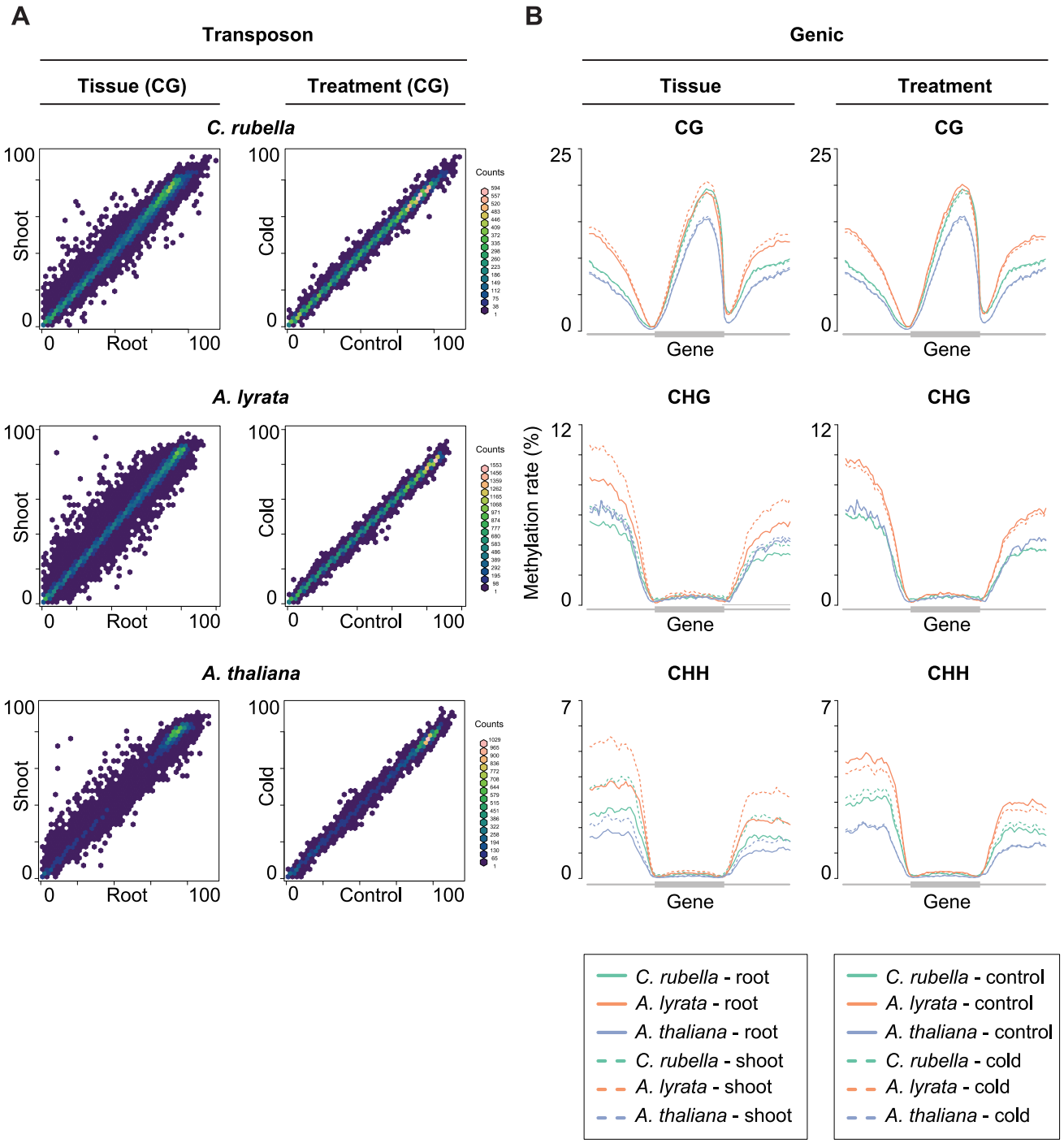

Figure 9. Intraspecific variation of transposon and gene body methylation. A) Comparison of the average methylation rates at annotated transposons and repeats between tissues (root and shoot) and treatments (23°C and 4°C). Average methylation rate is calculated as the average of methylation rates at all cytosines in the feature, including non-methylated cytosines. B) Methylation rate distribution across gene bodies of orthologous genes and flanking sequences (1.5 kb up- and 1.5 kb downstream). Orthologs that lacked methylation in both their gene body and flanking sequences are excluded. Distributions are plotted by context.
doi:10.1371/journal.pgen.1004785.g009

[6,7,8,9,10,11,14,15,16]. Furthermore, gene body methylation levels are conserved in orthologous genes indicating that DNA methylation rate may be subject to purifying selection, a finding consistent with previous wider evolutionary comparisons [42]. The close relationship of the species used in our experiments allows us to make inferences at base pair resolution. Given the substantial rate of epimutation in non-repetitive sequences [39,40], we were surprised to discover that a large fraction of sites is methylated in more than one species. These sites were predominantly found in gene bodies, providing additional evidence for selective constraint. While gene body methylation is poorly understood, there is some evidence that it is correlated with nucleosome positioning in exons

[14,59]. If nucleosome position is conserved, it could potentially explain long-term conservation of DNA methylation at some sites.

An additional proposed feature of DNA methylation as a molecular phenotype is the ability to respond to external stimuli or internal developmental cues. In theory, such variation could control changes in gene expression. We found evidence for DNA methylation variation in all three species across both tissue type and environment. The changes in DNA methylation were in all three species much greater between tissues, and consistently resulted in lower methylation levels in the root [19]. Differences between the root and shoot tissues also explain a majority of the expression variation in the transcriptional data, but these changes are not directional. We found no evidence that changes in DNA methylation across tissues is associated with changes in gene expression. In fact, a large proportion of methylation changes were found in repetitive sequences. This pattern may result from the increased stringency of transposon silencing in the shoot, which includes the plant germline [60].

While transcriptional responses are highly conserved across all three species, we found no evidence for conservation of DNA methylation response at the sequence level. MRs and DMRs are predominantly found in the rapidly evolving repeat-rich regions of the genome and rarely reside in or near the same orthologous gene in more than one species. In many of the classical epimutants, epigenetic regulation of nearby transposon insertions can impact neighboring genes and cause phenotypic variation [21,24,25,26]. This additional regulation is in some cases beneficial; for example, for genes specifically expressed in the pollen [41,61]. The data presented here demonstrates that these events are both rare and likely lineage-specific. It is possible that the reported cases of differential methylation as a regulator of transcription are short-term innovations that are eventually replaced by genetically encoded regulation.

DNA methylation from an epimutational perspective

The mode of inheritance of symmetrically methylated cytosines motivates the interpretation of DNA methylation as a molecular modification that increases the complexity of the genetic code. While mutational processes affecting DNA sequence are well described, epimutational processes are poorly understood. DNA mutations rarely revert and occur in a largely random fashion throughout the genome [62]. In contrast, recent studies have shown that the transgenerational stability of DNA methylation is very context dependent [39,40]. Over short evolutionary times, epimutations are more likely to occur in euchromatic sequences and are biased away from heavily methylated repetitive sequences [39,40].

Over the longer evolutionary times examined here, we find that changes in genome content and structure are the major contributors to DNA methylation variation. While the majority of single site and regional methylation is found in repetitive sequences that are unlikely under evolutionary constraint, the remaining observed patterns in euchromatic sequence reflect lineage-specific evolution of transposons. This is particularly obvious in *A. lyrata*, which has experienced a recent invasion of transposable elements into euchromatic sequences [46] and subsequent elevation in the methylation rate of euchromatic features, particularly introns.

Large-scale structural changes that have perturbed the genome-wide DNA methylation landscape have also occurred in *A. thaliana* [48,49]. Loss of three repeat-rich centromeres in *A. thaliana* caused a decrease in DNA methylation in sequences flanking the ancestral centromeres. The impact of lineage-specific transposon evolution and subsequent methylation is similarly evident in genic sequences. Approximately 40% of methylation in conserved exon sequence is species-specific. These sites are non-uniformly distributed near the 5′ or 3′ edges of genes, likely due to spreading from adjacent transposons [56,57,58]. These observations support the hypothesis that surveillance of transposons is the primary contributor to the genomic distribution of DNA methylation in plants. Since transposon content and genome structure vary extensively even over short evolutionary time periods, DNA methylation appears to be similarly variable. This is supported by the poor resolution of species relationships in a principal component analysis of DNA methylation and a nearly ten-fold increase in divergence between *A. lyrata* and *A. thaliana* when comparing DNA methylation as opposed to nucleotide sequence [46]. Together, these results indicate that DNA methylation as a non-canonical nucleotide is very rarely conserved over intermediate evolutionary times scales.

Despite the fact that we can estimate the epimutation rate of methylated cytosines and other parameters related to nucleotide mutations, it is misleading to equate DNA methylation changes to nucleotide substitutions. Our results indicate that the rapid evolution of repeat sequences is the major contributor to the equally rapid changes in the genomic distribution of DNA methylation. In this respect, it is more reasonable to regard DNA methylation primarily as a molecular phenotype resulting from the underlying genetic sequences. Although a few "pure" epialleles have been identified in nature, the majority of natural epimutations are linked to nearby transposon insertions or other genetic changes [21,24,25,26]. Fast evolution of repeat-sequences can, however, provide opportunities for lineage-specific cooption of DNA methylation for regulation of endogenous genes in response to various stimuli.

Materials and Methods

Experimental design

Seeds from the reference strain for each species (*A. thaliana* Col-0, *A. lyrata* MN47, *C. rubella* MTE) were sterilized with a 15 minute treatment of 30% bleach and 0.1% Triton X-100. Sterilized seeds were plated onto 0.5× MS 0.7% agar plates with 1% sucrose. Each plate represented a single replicate consisting of 20 seedlings. In total, 7 replicates were sown and randomized into a 3×2×2 factorial design. The three factors in this experiment were species, tissue, and cold treatment. After sowing, plates were stratified in the dark at 4°C for 8 days, before being shifted to 23°C short-day conditions (8 hr light:16 hr dark). Plates were oriented vertically. After 6 days in 23°C, half of the plates were exposed to 4°C short-day conditions for 23 hours. At the end of the cold treatment, both control (23°C) and treated (4°C) samples were harvested. Root and shoot tissues were harvested independently. Plants were cut just above and below the root-shoot junction to separate the tissues and avoid cross contamination of tissue types. To minimize daily collection times, replicates were blocked by day.

RNA extraction and RNA-seq library preparation

Total RNA was isolated from three replicates of each factor combination using the Qiagen RNAeasy Plant Mini Kit (catalog # 74904). An on-column DNase digestion was included (catalog # 79254). Total RNA integrity was confirmed on the Agilent BioAnalyzer. Illumina TruSeq RNA libraries were constructed using 3 μg of total RNA. Samples were randomized before library construction. The manufacturer's protocol was followed with one exception - 12 PCR cycles were used instead of the recommended 15. Libraries were quantified on an Agilent BioAnalyzer (DNA 1000 chip). Samples were normalized to 10 nM library molecules

and then pooled for sequencing. Three pools were constructed, each consisting of 12 random samples. Each pool was sequenced across three lanes of an Illumina GAII flowcell.

DNA extraction and bisulfite library preparation

DNA was extracted from two replicates of each factor combination using the Qiagen DNeasy Plant Mini Kit (catalog # 69104). DNA was quantified using the Qubit BR assay (Life Technologies, catalog # Q32853). Bisulfite libraries were confstructed using modifications to the Illumina TruSeq DNA kit and published bisulfite library protocols [15,39]. Depending on the sample, starting material ranged from 200 ng to 1 μg. Changes to the manufacturer's protocol will be noted here. After shearing of genomic DNA with a Covaris S220 instrument, sheared lambda DNA was spiked into each sample (1:0.001 sample:lambda ratio) as a control., for accurate estimation of failure to bisulfite convert non-methylated cytosines. Samples were randomized before library construction. During the ligation step, the amount of adapter was adjusted based on the amount of starting material in each sample. For 1 μg of input DNA, 2.5 μl of adapter were used. Adapter input was scaled linearly for samples with less starting DNA. For the second AMPure bead clean up after the ligation step, the ratio of sample to beads was adjusted to 1:0.74. A final elution volume of 42.5 μl was used for this step. After ligation, 40 μl of eluate was transferred to a new tube for subsequent bisulfite treatment.

The Qiagen Epitect Plus Kit (catalog # 59124) was used for bisulfite treatment. The manufacturer's protocol for 'low concentrated and fragmented samples' was followed, using 85 μl of bisulfite mix for conversion. Clean up of the bisulfite reaction included ethanol as a final wash step. The sample was eluted in 17 μl. After bisulfite treatment samples were amplified using Pfu Cx HotStart Polymerase from Agilent (catalog # 600410) instead of the supplied PCR mix. Reaction conditions are all follows: 32.9 μl of water, 5 μl of 10× Pfu Cx Buffer, 5 μl of 2 mM dNTP, 1.6 μl of Illumina PCR Primer Cocktail, 0.5 μl of Cx Polymerase (2.5 U/μl), 5 μl of bisulfite-treated DNA eluate. Three PCR reactions were pooled for each bisulfite-treated sample. The following cycling conditions were used: 98°C - 30 seconds; 18 cycles of 98°C - 10 seconds, 65°C - 30 seconds, 72°C - 30 seconds; 72°C - 5 minutes. An AMPure bead clean up was used to purify the final PCR product (1:1 sample to bead ratio). Samples were eluted in 32.5 μl of Illumina supplied Resuspension buffer. 30 μl of the final eluate was transferred to a new plate for subsequent quantification and sequencing. Libraries were quantified using the Agilent BioAnalyzer (DNA 1000 chip). Libraries were diluted to 10 nM and then pooled. Samples were pooled based on genome size - and each pool consists of 2 random samples from each species. Four pools were constructed and each was sequenced across three lanes of the Illumina HiSeq 2000.

Bisulfite sequencing

We sequenced bisulfite-converted libraries with 2×101 base pair paired-end reads on an Illumina HiSeq 2000 instrument with conventional *A. thaliana* DNA genomic libraries in control lanes. Each sample contained 0.1% lambda DNA as an unmethylated control. We pooled six different samples in each lane. The Illumina RTA software (version 1.13.48) performed image analysis and base calling.

Processing and alignment of bisulfite-treated reads

Reads were filtered and trimmed as previously described [39]. Subsequently, trimmed reads were mapped against the corresponding reference genomes (Crubella_183, Alyrata_107, Athaliana_167 (TAIR9) [46,47,50,51]. The lambda genome sequence was appended to each species genome sequence in order to estimate the false methylation rates of each sample. All reads were aligned using the mapping tool bismark v0.7.3 [63]. Applying the 'scoring matrix approach' of SHORE as previously described [39], we retrieved unique and non-duplicated read counts per position. Read and alignment statistics can be found in Table S2. All command line arguments are listed in Text S1. Raw reads are deposited at the European Nucleotide Archive under accession number PRJEB6701.

Determination of methylated sites

We used published methods [39], with a few exceptions. Here we retrieved incomplete bisulfite conversion rates, or false methylation rates (FMRs), from the alignments against the lambda genome rather than the chloroplast sequence. False methylation rates are found in Table S3. In addition, we combined the read counts of replicate samples after removing sites that were differentially methylated between replicates. The methylation rates for combined replicates were used for all subsequent analyses. The number of DMPs detected between replicates can be found in Table S19. In each species we required a methylation rate of at least 20% in one of the four tissue-treatment combinations in order for a site to be considered significantly methylated.

Identification of differentially methylated positions (DMPs)

To identify DMPs we followed published methods [39], but we required positions to have a methylation rate of at least 20% in one of the treatment combinations before performing Fisher's exact test. This increased statistical power by reducing the number of multiple testing corrections. Pairwise tests were not performed between all treatment combinations, instead only relevant comparisons were performed within each species (Root-23°C vs Shoot-23°C, Root-4°C vs Shoot-4°C, Root-23°C vs Root-4°C, Shoot-23°C vs Shoot-4°C).

Identification of methylated regions (MRs)

To detect contiguously methylated parts of the genome we modified a Hidden Markov Model (HMM) implementation [64]. Briefly, each cytosine can be in either an unmethylated or methylated state. The model trains methylation rate distributions for each state and sequence context (CG, CHG, CHH) independently using genome-wide data. In addition, transition probabilities between the states are trained. To make the original HMM implementation applicable to plant data, three different (beta binomial) distributions were estimated for each state (methylated and unmethylated) instead of just the single distribution used in mammals, which have almost only CG methylation [64]. To prevent identification of regions over uncovered bases, the genome was split at locations that lacked a covered cytosine position for 50 adjacent base pairs. On each of these segments, the most probable path through the methylation states was estimated after genome-wide parameter training. Transitions between states demarcated the methylated regions (MR). Replicates of each treatment combination were combined for this analysis. The combined read counts at cytosines were used to calculate methylation rates, train the HMM, and identify methylated regions. As a result, there is a single segmentation of the genome per treatment combination. Methylated regions were trimmed on both 5′ and 3′ ends by removing positions with a methylation rate

below 10%. Further details will be described in a manuscript by Hagmann, Becker et al. [65].

Identification of differentially methylated regions (DMRs)

Based on the MRs identified for each sample using the HMM algorithm described above, we selected regions of variable methylation state between samples to test for differential methylation. Due to the very large number of MRs, it was critical to reduce the number of tests performed to identify DMRs. By filtering MRs using the criteria outlined in a forthcoming manuscript by Hagmann, Becker et al. [65], we reduced the number of MRs four fold in each species. For each identified region, pairwise statistical tests were performed for the relevant comparisons listed above. The statistical test approximates the context-specific beta binomial distribution for the region of interest. Individual and joint distributions are approximated for two samples being compared. The statistical test compares the individual sample distributions to the joint distribution using a log-odds ratio. This ratio is compared against a chi-squared distribution to obtain confidence values. For each identified region, samples were assigned to groups by separating the samples with statistically significant methylation. To confirm groupings, we first combined read counts from treatment combinations in the same group. With the combined data, the same statistical test as described above was performed to test for differential methylation. Groups were confirmed in this way to identify and filter potentially erroneous DMRs. After false discovery rate (FDR) correction using Storey's method [66], regions with an FDR below 0.01 were defined as differentially methylated regions (DMRs). To resolve overlapping DMRs, we retained the non-overlapping regions containing the maximum number of samples with statistically significant differential methylation. Apart from the criterion used to resolve overlapping DMRs, the methods follow those that will be described in detail in a manuscript by Hagmann, Becker et al. [65].

Site-level conservation of methylation

We identified conserved sites using a published three-way whole genome alignment [47]. For CG sites, identical context was required while substitutions at the H positions were allowed in degenerate contexts as long as they did not mutate to G. Sites that transitioned contexts were not considered. Methylation rates for significantly methylated sites were then extracted from each species, tissue, and treatment combination for subsequent analysis.

Identification of 1:1:1 orthologous gene pairs

Three-way orthologs were identified using the reciprocal-best blastp hit approach as implemented in the multiParanoid pipline (inParanoid v. 4.1, blast v. 2.2.26) [67].

RNA sequencing

We sequenced each RNAseq library with 101 base pair single-end reads on the Illumina GAII instrument. We pooled twelve different samples in each lane. Each pool was sequenced over three lanes. The Illumina RTA software (version 1.13.48) performed image analysis and base calling.

Processing and alignment of RNAseq reads

Reads were trimmed using the shore import function in SHORE version 0.9.0 [68]. Command line arguments can be found in Text S1. This function simultaneously trims reads and separates samples by barcode. Since all samples were sequenced over three lanes, after lanes are de-multiplexed sample reads were combined. Due to variable annotation qualities between species, only sequences annotated as CDS annotations were used to map RNA-seq reads. The following representative gene model annotation versions were used for each species: Crubella_183, Alyrata_107, Athaliana_167 (TAIR10) [46,47,50,51]. Reads were aligned with one allowed mismatch to the appropriate annotation using bwa version 0.6.1 [69]. Read counts were obtained for each gene using a custom perl script. In summary, the script identified uniquely aligned read with a mapping quality score above 30 and stored the total read count for each target sequence. Read and alignment statistics can be found in Table S20. Raw reads are deposited at the European Nucleotide Archive under accession number PRJEB6701.

Differential expression analysis

Differentially expressed genes were identified using the R package edgeR (3.4.2) with minor modifications [70]. Using edgeR, we estimated the dispersion parameter for each gene using estimateGLMTagwiseDisp(). Next, we fit a negative binomial generalized linear model (GLM) using glmFit(). Significance testing for differential expression was performed using a custom GLM. Significance testing in edgeR was done via term-dropping of each factor level (likelihood ratio test), and as a result performed more statistical tests than necessary. To minimize multiple testing problems, we implemented a negative binomial GLM that tested for differential expression significance using an ANOVA [71]. Dispersion estimates from edgeR were provided to the modified GLM. Using this model, differential expression analysis was performed in two ways. First, expression analysis was performed within species. There were 12 samples consisting of three replicates and four unique treatment combinations. All representative gene models were considered. The following custom GLM model was used: expression~tissue*treatment. This included the main effects of tissue and treatment as well as their interaction. Secondly, we performed differential expression analysis between all species simultaneously. In this case, there are a total of 36 samples consisting of three replicates of each species, tissue, and treatment combination. Only 1:1:1 orthologous gene pairs were considered (14,395 in total). The following custom GLM model was used: expression~species*tissue*treatment. This includes the main effects of species, tissue, and treatment as well as all two and three-way interactions. Corrections for gene length were performed, but this did not impact the results and was subsequently ignored.

Repeat annotations

Transposon and repeat annotations for all three species were derived from the *Capsella rubella* genome paper [47,72,73].

Supporting Information

Figure S1 Effects of intron insertions of transposons on gene expression. Genes with and without TEs in their introns are compared. A gene is considered expressed if it had at least 3 RPKM in three of the twelve species-specific RNA-seq samples. (EPS)

Figure S2 Methylation rates of sequences flanking intron-inserted transposons. All cytosines in sequences flanking TEs in introns were extracted (+/−500 bp). Methylation rate for each annotated feature and context is calculated as the number of methylated cytosines over the total number of possible cytosines. Methylation rates are normalized to genome-wide methylation rates for each feature-context combination. Sites considered in our current analysis (intronic TE and +/−500 bp) were excluded from

the calculation of background methylation rates. This plot also accounts for expression of the gene containing the intronic TE.
(EPS)

Figure S3 Methylation rates at genomic features of expressed genes. Genome average of methylation rates for each genomic feature. Similar to figure 2B, except annotations are only considered for genes that are expressed. A gene is considered expressed if it received at least 3 RPKM in three of the twelve species-specific RNA-seq samples. Methylation rates are normalized to the outgroup species *C. rubella*.
(EPS)

Figure S4 Genomic distribution of MRs and DMRs. A) Circos plots [74] to demonstrate the genomic distribution of MRs and DMRs in *C. rubella*, *A. lyrata*, and *A. thaliana*. Chromosome number is indicated on the inner circle. Data is plotted for 500 kb windows.
(EPS)

Figure S5 Relationship between gene body methylation and gene expression. Gene body methylation rates are plotted against either gene expression ($\log_2$) deciles or coefficient of variation (CV) deciles. When comparing gene body methylation with gene expression the Spearman rank correlation coefficient in *C. rubella* = 0.21, *A. lyrata* = 0.23, and *A. thaliana* = 0.24. In contrast, when comparing gene body methylation with CV the Spearman rank correlation coefficient in *C. rubella* = −0.34, *A. lyrata* = −0.19, and *A. thaliana* = −0.33.
(EPS)

Figure S6 Annotation of methylated site classes in three-way alignments. Feature annotation is shown for each methylation context. Site classes are as follows: Aligned - all C in three-way alignments. mC - methylated sites within a species. Consv. (3 species) - sites that are methylated in all three species. Gain - sites that are methylated in a single species. Loss - sites that have lost methylation in a single species.
(EPS)

Figure S7 Transposon categories for aligned methylated site classes. The top 5% of windows (10 kb) for three-way conserved sites, gains, and losses were identified. As a control, an equal number of random genomic windows were chosen. Shown is the number of bases annotated as a transposon category for the top 5% of windows in each site class normalized to the control annotation.
(EPS)

Figure S8 Centromere loss is not associated with methylation loss at aligned cytosines. Fraction of species-specific losses in methylation is plotted for each ortholog residing within ancestral centromere boundaries. Orthologs were categorized based on genomic position, either in or outside of ancestral centromere boundaries. Centromere boundaries were defined in *C. rubella* using repeat density (Fig. 3, 0.3 threshold). Orthologs residing in maintained ancestral centromeres ("No Loss") were compared to orthologs residing in ancestral centromeres lost in *A. thaliana* ("Loss").
(EPS)

Figure S9 Conserved methylated sites associated with conservation of exon length. Fraction of site categories that reside in exons with conserved lengths across all three species or exons of variable lengths.
(EPS)

Figure S10 Distribution of cytosines across exons. The density of exon methylation at aligned cytosines is shown for conserved methylated sites as well as for lineage-specific gains and losses of methylation. On top is the density of non-methylated aligned cytosines. There is no bias in location within an exon for non-methylated sites.
(EPS)

Figure S11 Differential gene expression. For each model, within species (top) and between species (bottom), the number of differentially expressed genes (absolute and as a fraction of expressed genes) is shown for each main effect and all interactions ($p<0.05$).
(EPS)

Figure S12 Species mC relationship of replicates. A) Principal component analysis on mC rates at aligned methylated positions. All contexts are considered (see Fig. 6B,C and Table S10 for further description of mC sites). Unlike figure 7, this plot considers the mC rate of each replicate at all aligned methylated positions.
(EPS)

Figure S13 Genomic distribution of DMPs. A) Circos plots [74] to demonstrate the genomic distribution of DMPs in *C. rubella*, *A. lyrata*, and *A. thaliana*. Plots are separate for tissue specific DMPs (root and shoot) or treatment specific DMPs (23°C and 4°C). Chromosome number is indicated on the inner circle. Data is plotted for 500 kb windows.
(EPS)

Figure S14 Conservation of DMRs in the absence of sequence alignments. The total number of orthologous genes containing a DMR in one, two, or three species is shown. Location of DMR overlap is separated by genomic feature. Upstream region is considered 1 kb before the start codon. Asterisk indicates two or three-way sharing of DMRs that exceeds permutation values.
(EPS)

Table S1 References for genome size. References for the genome size (in pg and Mb) as well as the total size of the genome assembly are listed for each species. Genome size references are derived from the Kew Royal Botanic Gardens Plant DNA C-values database.
(XLSX)

Table S2 Bisulfite-sequencing coverage and alignment statistics. For each sample, the total number of sequenced reads (paired and single) is shown. Also, the total number of CG, CHG, and CHH sites covered is reported along with the average genome-wide coverage of each context.
(XLSX)

Table S3 False methylation rates by coverage bin. The incomplete bisulfite conversion rate, or false methylation rate (FMR), for each sample is shown by coverage bin. For each bin, FMR is calculated as the number of cytosines in lambda DNA that are not converted to U (T in the DNA sequence) after bisulfite treatment over the total number of converted (U/T) and unconverted (C) reads.
(XLSX)

Table S4 MR and DMR statistics by sample (A) and species (B). Mean and median length of region, total number of regions, and genomic bases covered by regions are shown. Sample statistics were calculated from the combination of biological replicates.
(XLSX)

Table S5 Genome alignment metrics. Number of bases covered in three-way whole genome alignments is shown. In addition, total number of bases in methylated site classes is shown.
(XLSX)

Table S6 MR presence at genomic features. The number of genes in each species annotation that contain a MR is shown. Upstream refers to 1 kb upstream of the start codon. The number of orthologous genes with an overlapping MR is also shown.
(XLSX)

Table S7 P values of pairwise MR overlap. To test the significance of MR co-occurrence at orthologous genes a hypergeometric test was used. Significance of each test is shown here.
(XLSX)

Table S8 Two and three-way species MR overlap. The number of orthologs that contain an MR in one, two, or three species is shown (A). Permutation analysis was performed to estimate the random occurrence of one, two, and three-way overlap (10,000 permutation tests). Maximum permutation values are shown in (B). Features where the data exceeds the maximum permutation value are indicated in (C).
(XLSX)

Table S9 Gene body methylation by context. The total numbers of genes with CG, CHG, or CHH gene body methylation are shown for all genes (A) and orthologous genes (B).
(XLSX)

Table S10 Three-way genome alignment site classes by context. Total numbers of CG, CHG, and CHH sites for each alignment site class are shown.
(XLSX)

Table S11 DMP statistics by comparison. Total numbers of DMPs in each tissue and treatment comparison are shown.
(XLSX)

Table S12 DMR statistics by comparison. Total numbers of DMRs in each tissue and treatment comparison are shown.
(XLSX)

Table S13 DMR presence at genomic features. The number of genes in each species' annotation that contain a DMR is shown. Upstream refers to 1 kb upstream of the start codon. The number of orthologous genes with an overlapping DMR is also shown.
(XLSX)

Table S14 P values of pairwise DMR overlap. To test the significance of DMR co-occurrence at orthologous genes a hypergeometric test was used. Significance of each test is shown here.
(XLSX)

Table S15 Two and three-way species DMR overlap. The number of orthologs that contain a DMR in one, two, or three species is shown (A). Permutation analysis was performed to estimate the random occurrence of one, two, and three-way overlap (10,000 permutation tests). Maximum permutation values are shown in (B). Features where the data exceeds the maximum permutation value are indicated in (C).
(XLSX)

Table S16 DMPs at genomic features. The number of genes in each species annotation that contain a DMP is shown. Upstream refers to 1 kb upstream of the start codon. The number of orthologous genes with an overlapping DMP is also shown.
(XLSX)

Table S17 DMP correlation with gene expression by feature. Spearman rank correlation coefficient was calculated between the direction of differential methylation and the appropriate $\log_2$ fold change. Correlation coefficients were calculated separately for tissue and treatment specific DMPs. An NA value indicates that there were too few genes in a given category. Expression values are from the intraspecific expression analysis.
(XLSX)

Table S18 DMR correlation with gene expression by feature. Spearman rank correlation coefficients when comparing the degree of differential methylation for each context (extracted from the HMM model) with the appropriate $\log_2$ fold change. All annotated genes overlapping a DMR were considered. Expression values are from the intraspecific expression analysis. Correlation coefficients were calculated only for tissue-specific DMRs as there are too few treatment-specific DMRs. Results are only shown for DMRs overlapping CDS, intron, and upstream sequences because too few expressed genes reside in the other categories (5′ and 3′ UTRs). Upstream refers to 1 kb upstream of the start codon.
(XLSX)

Table S19 Number of DMPs between replicates. For each species, tissue, treatment combination, differentially methylated positions were identified between biological replicates. The total number of DMPs for each comparison is listed. These positions were removed from all further analyses.
(XLSX)

Table S20 RNA-seq sequencing coverage and alignment statistics. For each sample, the total number of RNA sequencing reads is shown. Read counts are also shown for mapped reads, uniquely mapped reads, and the reads that passed a mapping quality threshold (30).
(XLSX)

Text S1 Command lines for alignments. Command lines and arguments for the processing of bisulfite reads and RNA-seq reads.
(TXT)

Acknowledgments

We thank Florian Maumus and Hadi Quesneville for early access to their transposon annotation of the three genomes.

Author Contributions

Conceived and designed the experiments: DKS DK CB DW. Performed the experiments: DKS DK. Analyzed the data: DKS DK JH CB. Contributed to the writing of the manuscript: DKS DK DW.

References

1. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 107: 8689–8694.
2. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328: 916–919.
3. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9: 465–476.
4. Huff JT, Zilberman D (2014) Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. Cell 156: 1286–1297.
5. Gruenbaum Y, Navehmany T, Cedar H, Razin A (1981) Sequence specificity of methylation in higher-plant DNA. Nature 292: 860–862.
6. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126: 1189–1201.

7. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet 39: 61–69.
8. Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, et al. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat Genet 23: 305–308.
9. Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, et al. (2013) The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. Genome Res 23: 1651–1662.
10. Li X, Zhu J, Hu F, Ge S, Ye M, et al. (2012) Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. BMC Genomics 13: 300.
11. Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc Natl Acad Sci U S A 110: 1797–1802.
12. Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8: 272–285.
13. Lisch D (2013) How important are transposons for plant evolution? Nat Rev Genet 14: 49–61.
14. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452: 215–219.
15. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523–536.
16. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315–322.
17. Coleman-Derr D, Zilberman D (2012) Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. PLoS Genet 8: e1002988.
18. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, et al. (2013) Patterns of population epigenomic diversity. Nature 495: 193–198.
19. Widman N, Feng S, Jacobsen SE, Pellegrini M (2014) Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation. Epigenetics 9: 236–242.
20. Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, et al. (2012) Widespread dynamic DNA methylation in response to biotic stress. Proc Natl Acad Sci U S A 109: E2183–2191.
21. Morgan HD, Sutherland HG, Martin DI, Whitelaw E (1999) Epigenetic inheritance at the agouti locus in the mouse. Nat Genet 23: 314–318.
22. Cubas P, Vincent C, Coen E (1999) An epigenetic mutation responsible for natural variation in floral symmetry. Nature 401: 157–161.
23. Manning K, Tor M, Poole M, Hong Y, Thompson AJ, et al. (2006) A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet 38: 948–952.
24. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, et al. (2009) A transposon-induced epigenetic change leads to sex determination in melon. Nature 461: 1135–1138.
25. Das OP, Messing J (1994) Variegated phenotype and developmental methylation changes of a maize allele originating from epimutation. Genetics 136: 1121–1141.
26. Liu J, He Y, Amasino R, Chen X (2004) siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. Genes Dev 18: 2873–2878.
27. Miura K, Agetsuma M, Kitano H, Yoshimura A, Matsuoka M, et al. (2009) A metastable DWARF1 epigenetic mutant affecting plant stature in rice. Proc Natl Acad Sci U S A 106: 11218–11223.
28. Heard E, Martienssen RA (2014) Transgenerational epigenetic inheritance: myths and mechanisms. Cell 157: 95–109.
29. Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, et al. (2001) Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. Science 292: 2077–2080.
30. Bestor T, Laudano A, Mattaliano R, Ingram V (1988) Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. J Mol Biol 203: 971–983.
31. Finnegan EJ, Dennis ES (1993) Isolation and identification by sequence homology of a putative cytosine methyltransferase from Arabidopsis thaliana. Nucleic Acids Res 21: 2383–2388.
32. Leonhardt H, Page AW, Weier HU, Bestor TH (1992) A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell 71: 865–873.
33. Chuang LSH, Ian HI, Koh TW, Ng HH, Xu GL, et al. (1997) Human DNA (cytosine-5) methyltransferase PCNA complex as a target for p21(WAF1). Science 277: 1996–2000.
34. Pelissier T, Thalmeir S, Kempe D, Sanger HL, Wassenegger M (1999) Heavy de novo methylation at symmetrical and non-symmetrical sites is a hallmark of RNA-directed DNA methylation. Nucleic Acids Res 27: 1625–1634.
35. Cao X, Jacobsen SE (2002) Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. Proc Natl Acad Sci U S A 99 Suppl 4: 16491–16498.
36. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11: 204–220.
37. Chan SW, Zilberman D, Xie Z, Johansen LK, Carrington JC, et al. (2004) RNA silencing genes control de novo DNA methylation. Science 303: 1336.
38. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, et al. (2013) The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell 153: 193–205.
39. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, et al. (2011) Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature 480: 245–249.
40. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, et al. (2011) Transgenerational epigenetic instability is a source of novel methylation variants. Science 334: 369–373.
41. Calarco JP, Borges F, Donoghue MT, Van Ex F, Jullien PE, et al. (2012) Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell 151: 194–205.
42. Takuno S, Gaut BS (2012) Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol Biol Evol 29: 219–227.
43. Long HK, Sims D, Heger A, Blackledge NP, Kutter C, et al. (2013) Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. eLife 2: e00348.
44. Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. Proc Natl Acad Sci U S A 107: 18724–18728.
45. Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, et al. (2005) Evolution of genome size in Brassicaceae. Ann Bot 95: 229–235.
46. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, et al. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet 43: 476–481.
47. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, et al. (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet 45: 831–835.
48. Yogeeswaran K, Frary A, York TL, Amenta A, Lesser AH, et al. (2005) Comparative genome analyses of Arabidopsis spp.: inferring chromosomal rearrangement events in the evolutionary history of A. thaliana. Genome Res 15: 505–515.
49. Lysak MA, Berr A, Pecinka A, Schmidt R, McBreen K, et al. (2006) Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. Proc Natl Acad Sci U S A 103: 5224–5229.
50. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.
51. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, et al. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36: D1009–1014.
52. Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ (2009) The dynamic ups and downs of genome size evolution in Brassicaceae. Mol Biol Evol 26: 85–98.
53. Bennett MD, Leitch IJ, Price HJ, Johnston JS (2003) Comparisons with Caenorhabditis (similar to 100 Mb) and Drosophila (similar to 175 Mb) using flow cytometry show genome size in Arabidopsis to be similar to 157 Mb and thus similar to 25% larger than the Arabidopsis genome initiative estimate of similar to 125 Mb. Ann Bot 91: 547–557.
54. Bennett MD, Leitch IJ (2011) Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. Ann Bot 107: 467–590.
55. Schmitz RJ, He Y, Valdes-Lopez O, Khan SM, Joshi T, et al. (2013) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res 23: 1663–1674.
56. Arnaud P, Goubely C, Pelissier T, Deragon JM (2000) SINE retroposons can be used in vivo as nucleation centers for de novo methylation. Mol Cell Biol 20: 3434–3441.
57. Sun FL, Haynes K, Simpson CL, Lee SD, Collins L, et al. (2004) cis-Acting determinants of heterochromatin formation on Drosophila melanogaster chromosome four. Mol Cell Biol 24: 8210–8220.
58. Saze H, Kakutani T (2007) Heritable epigenetic mutation of a transposon-flanked Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1. EMBO J 26: 3641–3652.
59. Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, et al. (2010) Relationship between nucleosome positioning and DNA methylation. Nature 466: 388–392.
60. Baubec T, Finke A, Scheid OM, Pecinka A (2014) Meristem-specific expression of epigenetic regulators safeguards transposon silencing in Arabidopsis. EMBO Rep 15:446–52.
61. Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, et al. (2009) Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell 136: 461–472.
62. Lynch M (2007) The Origins of Genome Architecture. Sunderland, MA: Sinauer Associates. 389 p.
63. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27: 1571–1572.
64. Molaro A, Hodges E, Fang F, Song Q, McCombie WR, et al. (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell 146: 1029–1041.
65. Hagmann J, Becker C, Muller J, Stegle O, Meyer RC, et al. (2014) Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. bioRxiv doi: 101101/009225.

66. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100: 9440–9445.
67. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314: 1041–1052.
68. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, et al. (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res 18: 2024–2033.
69. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
70. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140.
71. Koenig D, Jimenez-Gomez JM, Kimura S, Fulop D, Chitwood DH, et al. (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proc Natl Acad Sci U S A 110: E2655–2662.
72. Maumus F, Quesneville H (2014) Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter. PLoS ONE 9: e94101.
73. Maumus F, Quesneville H (2014) Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. Nat Commun 5: 4104.
74. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19: 1639–1645.