SUPPORTING INFORMATION

For **Takuno and Gaut, "**Gene-body methylation is conserved between plant orthologs and is of evolutionary consequence"

SUPPORTING METHODS

Generating BS-seq data in *Brachypodium distachyon.* Briefly, approximately two micrograms of genomic DNA was sonicated to ~100 bp using the Covaris S2 System. Sonicated DNA was purified using Qiagen DNeasy mini-elute columns (Qiagen) and each sequencing library was constructed using the NEBNext DNA Sample Prep Reagent Set 1 (New England Biolabs, Ipswich, MA) according to the manufacturer's instructions but with methylated adapters used in place of the standard genomic DNA adapters. Ligation products were purified with AMPure XP beads (Beckman, Brea, CA). DNA was bisulfite treated using the MethylCode Kit (Invitrogen, Carlsbad, CA) following the manufacturer's guidelines and then PCR amplified using Pfu Cx Turbo (Agilent, Santa Clara). Libraries were sequenced using the Illumina HiSeq 2000 (Illumina) as per manufacturer's instructions. Image analysis and base calling were performed with the Illumina pipeline version RTA 2.8.0.

Methylation analysis and designation of genes as BM, IM or UM. As mentioned in the Main Text, we mapped BS-seq reads to reference genomes using BRAT software (1), allowing mismatches at only potentially methylated sites. We assessed methylation at individual cytosine sites by first assessing the average level of converted to non-converted reads across all cytosine sites. The averages were remarkably similar between tissues; for example over the entire genome, the average cytosine in the CG context had 54.3% non-converted reads, on average, in leaf tissue and 53.9% non-converted reads in floral tissue. These values were identical to those within coding regions.

We next constructed a histogram of the proportion of non-converted reads for each cytosine site against the number (frequency) of cytosine sites (Fig. S1). The distribution was notably bimodal, allowing clear designation of 'methylated' and 'non-methylated' sites. However, rather than rely on an arbitrary cut-off to designate as methylated, we employed the binomial test suggested by Lister et al. (2). This binomial test employs the estimated error rate (Table S1) to test for significant evidence of methylation at each site. We employed the test at a P-value of < 0.01 significance; the results correlated nicely with the bimodal distribution in Fig. S1. Given that this binomial approach has substantial precedence in the methylation literature (e.g., 3-5), we used it to define individual cytosines as methylated or non-methylated.

Given identification of cytosine residues as methylated or non-methylated, we then classified genes into one of three methylation classes: BM (body-methylated), IM (intermediately-methylated) and UM (under-methylated). To classify a gene, we first denoted a P-value for the CG, CHG and CHH contexts as P_{CG}, P_{CHG} and P_{CHH}, respectively (6). For example, the probability of the null hypothesis of CG methylation at the level of the genomic average was given by

$$P_{CG} = \sum_{i=m_{cg}}^{n_{cg}} \binom{n_{cg}}{i} p_{cg}^{i} (1 - p_{cg})^{n_{cg}-i},$$

where p_{cg} is the proportion of methylated CG sites in the genome, n_{cg} is the number of cytosine residues at CG sites with enough coverage of BS-seq, and m_{cg} is the number of methylated cytosine residues. Using the same approach, P_{CHG} and P_{CHH} were calculated for CHG sites and CHH sites, respectively. The lower these statistics, the more densely methylated a local region relative to the genomic average.

P_{CG}, P_{CHG} and P_{CHH} were calculated for each annotated region from translation start to the stop codon. We only considered genes with sufficient CG information (n_{cg} $\geq$20) and genes for which $\geq$ 40% and 60% of cytosine residues were covered by at least 2 reads for rice and *B. distachyon*, respectively (Fig. S6). Within *bona fide* genes, body methylation is enhanced at only CG sites (2, 7), so we discarded genes with high CHG and/or CHH methylation (i.e., with $P_{CHG} < 0.05$ and/or $P_{CHH} < 0.05$). This step resulted in the exclusion of 1,967 of 23,051 annotated genes in *B. distachyon* with sufficient methylation information and 1,501 of 26,505 annotated genes in rice. We then classified the remaining genes into three categories: BM ($P_{CG} < 0.05$), IM ($0.05 \leq P_{CG} \leq 0.95$) and UM ($P_{CG} > 0.95$).

Identifying orthologs and calculating evolutionary rates. We calculated substitution rates between *A. thaliana-A. lyrata* ortholog pairs and between *O. sativa-B. distachyon* ortholog pairs. The list of 18,330 orthologs for *A. thaliana-A. lyrata* pair was provided by J. A. Fawcett (The Graduate University for Advanced Studies, Hayama, Japan) (8), with the BM genes in *A. thaliana* previously defined (6). For *O. sativa-B. distachyon*, we inferred orthologous relationships following (8) with slight modifications. First, BlastP analyses [version 2.2.17 (9)] were performed with $E < 10^{-5}$. Homologous gene pairs whose alignments covered $\geq$ 50 % of both genes for both directions were retained. Second, synonymous divergence (denoted by K_S) was calculated for all gene pairs using the Nei and Gojobori method (10) after alignment by ClustalW version 1.83 (11). The major peak of K_S distribution was 0.35 to 0.40; we discarded gene pairs with $K_S > 0.7$ to enhance our further queries to *bona fide* orthologs as opposed to paralogs.

Homologous pairs were retained as queries for i-ADHoRe version 3.0 (12) to assess collinearity. i-ADHoRe was applied with gap size set to 10 genes, the minimum number of homolog anchors set to 10 genes, and the *P*-value cutoff set to 0.001. Duplicated genes may lead to ambiguous inference of orthologs. To avoid misassignment of tandemly duplicated genes, we excluded tandem duplications with $K_S < 0.7$ from the i-ADHoRe query. We ultimately detected 9,531 orthologs between rice and *B. distachyon* under these homology and collinearity parameters. These 9,531 orthologs were further pared to 7,826 orthologs based on two criteria: sufficient levels of methylation data (Fig. S6) and exclusion of genes with $P_{CHG} < 0.05$ and/or $P_{CHH} < 0.05$.

For the remaining orthologs, we calculated K_A and K_S using the Nei and Gojobori method after alignment with ClustalW, limiting our analyses to ortholog alignments that included $\geq$100 bp of synonymous change sites in alignment sequences. We also calculated CG [O/E] (13) from the ortholog data.

Analysis of maize BS-seq data and identification of maize orthologs. To assess gbM in maize, we employed three sets of data: *i*) the B73 maize genome sequence RefGen version 2 (14), *ii*) the filtered gene set for annotations (version 5bFGS), retrieved from MaizeSequence.org, and *iii*) BS-seq data generated from the outer layer of mature maize ears prior to fertilization in the reference inbred line B73 (15).

The BS Seeker software program was used to map the BS-seq short reads to the maize B73 reference genome (16), using default settings. BS Seeker was used for mapping because the BRAT software (1) mapped few reads uniquely within the maize genome, probably reflecting the polyploid history of the genome and the abundance of repetitive DNA(14, 17, 18). We note, however, that mapping with BRAT and BS Seeker resulted in < 0.058% differences in methylation calls within *B. distachyon* and also performed well with *A. thaliana* data (6) relative to the original BS-seq analyses (2). Thus, the limitations of mapping by BRAT applied only to the maize genome, which is substantially more repetitive than *A. thaliana*, rice or *B. distachyon* (19). Once reads were mapped, we applied the binomial method (2) at $p < 0.01$ to classify each cytosine residue with sufficient data as methylated or un-methylated.

We then assessed levels of cytosine methylation within the filtered gene set. Because BS-seq coverage was low for the maize data, we screened genes such that ≥20% of cytosines were covered by at least 2 reads and that ≥10 cytosines at CG sites were covered by at least 2 reads (Fig. S6).

Finally, the proportion of methylated cytosines within CG dinucleotides was plotted for orthologs to our set of 7,826 rice-*B. distachyon* orthology pairs. Maize orthologs to these pairs were retrieved from the quartet database of (20). Note that maize has two subgenomes, such that there were the possibility of two orthologs for maize relative to any of our rice and *B. distachyon* orthology pairs. We reported methylation correlations to orthologs from maize subgenome 1, simply because we identified more orthologs from this subgenome ($n = 914$) than from subgenome 2 ($n = 864$). However, pairwise correlations between rice/*B. distachyon* and subgenome1/subgenome2 were all highly significantly positive ($P < 10^{-5}$ by permutation test). Note also that the low coverage of maize BS-seq data results in substantial noise, undoubtedly weakening the observed correlations between maize orthologs relative to rice/*B. distachyon* orthologs.

Analysis of the distance from genes to methylated transposons. To investigate whether BM genes are typically within a chromatin environment defined by transposable element (TE) methylation, we identified TEs in *B. distachyon* using RepeatMasker (www.repeatmasker.org), with default parameters set to the *B. distachyon* genome. Sequence hits were further screened to retain hits > 100 bp in length that encompassed at least 70% of the reference TEs. By this method, 8,680 TEs were identified in the *B. distachyon* genome (version 1.0.) Given these TEs, we first quantified their level of DNA methylation, based on CG and CHG methylation, focusing on TEs for which 60% of CG and CHG sites had ≥2X coverage. Following a previous study (21), TEs with ≥10% methylation were denoted as methylated. Ultimately, the vast majority of assayable TEs were deemed to be methylated by this method – i.e., 5,861 methylated TEs and 179 unmethylated TEs. Given these designations, the distance between UM or BM genes to the nearest methylated TE was tallied. If the closest TE to a gene did not

have sufficient coverage of BS-seq to designate its methylation status – i.e., it was one of the ~2640 such TEs – then the gene was removed from the analysis.

Simulations of the mutation process at equilibrium. Methylated cytosines deaminate spontaneously, leading to high mutation rates and the preferential replacement of cytosine (C) with thymine (T). This mutation pressure reduces the observed number of CG dinucleotides relative to those expected based on the G+C content of a gene. A measure of this deviation, CG [O/E], has been used as a proxy to assess methylation content.

The question remains as to how genes can remain methylated over long evolutionary periods in the face of the mutation pressure that removed the very sites - CG dinucleotides – that are methylated. To study this process, we simulated DNA sequence evolution under a simplified model of nucleotide substitution.

We specified a simplified model of nucleotide substitution, following observations in reference (22) (Fig. S7). We assumed that the mutation rate of two types of transitions (i.e., C → T or G →A) is α times higher than the mutation rate of other nucleotide changes, μ, where μ is the rate of mutation to a specific nucleotide per site per generation (i.e., the rate of total mutations per site is 3μ). Once a substitution model is specified, the expected G+C content can be calculated by equation [3] in (23). Let μ_1 and μ_2 be the rate of mutation from G or C to A or T and the rate of mutation from A or T to G or C, respectively. The expected G+C content at equilibrium is calculated by $\mu_2/(\mu_1+\mu_2)$. In our model (Fig. S6), $\mu_1 = (2+2\alpha)\mu$ and $\mu_2 = 4\mu$. Thus, the expected G+C content under this model is calculated by $2/(3+\alpha)$. The genomic average of G+C content for coding regions was about 40% in *A. thaliana,* rice and *B. distachyon* (including both exons and introns). And therefore, we use $\alpha = 2$ because $2/(3+2) = 0.4$.

We further incorporated the effect of methylation at CG sites into the model, such that the C → T and G→A mutation rates at CG sites are β times higher than $\alpha\mu$. We also assumed all mutations are neutral, so that the mutation rate is equal to the substitution rate.

Given this model of nucleotide substitution, we simulated sequence of length L = 1000 bp with a given μ, α (=2μ) and a wide range of β values = {1,2,4,6,8,10,12,14,16,18,20}. We employed an interval of $m = 100/\mu$ generations as a time standard. To reach equilibrium, the simulation was run initially $2{,}000m$ generations. After this initial run, the CG [O/E] of the sequence was calculated every m generations. The simulation ran continuously, up to 10^5 observations, from which the average CG [O/E] was calculated. We do not specify μ itself because G+C content and CG [O/E] are independent from μ, which was verified by simulation.

The results are shown in Fig. S7. With $\alpha = 2$ and $\beta = 5$, the CG [O/E] at equilibrium ~ 40%, which is consistent with *A. thaliana*, *B. distachyon* and rice coding regions of BM genes (both exons and introns) (Figs. 3 & S5). The main interpretation is that when the deamination rate is at a reasonable level of $\beta = 5$ (24), then the level of CG[O/E] found in some plant species can be maintained indefinitely within a DNA sequence. This maintenance is not because the CG sites are themselves maintained indefinitely but rather because there is enough contravening mutation to create enough CG dinucleotides to maintain methylation.

SUPPLEMENTAL TABLES and FIGURES

Supplementary Table 1. Coverage of BS-seq in *B. distachyon* leaf and flower bud.

Tissue	Biological replicate	Number of uniquely mapped short reads	Coverage	Error rate [a]
Leaf	1	49,640,480	17.8X	1.06%
	2	44,466,382	15.9X	1.33%
	3	48,287,167	17.3X	1.12%
Flower bud	1	50,250,104	18.0X	0.94%
	2	47,322,329	17.0X	0.89%
	3	43,761,321	15.7X	0.98%

[a] Error rate is estimated from the proportion of non-converted reads that maps to the chloroplast genome.

Supplementary Table 2. Correlation between methylation level and the density of genes and TEs.

Tissue	Biological replicate [a]	Sequence context	Correlation to TE density [b]	Correlation to Gene density [b]
Leaf	1	CG	0.511	−0.692
		CHG	0.539	−0.788
		CHH	−0.303	0.303
Flower bud	1	CG	0.513	−0.696
		CHG	0.541	−0.789
		CHH	−0.207	0.171

[a] We obtained qualitatively identical results using other replicate samples.

[b] Correlation between methylation level and the density of genes or TEs in non-overlapped 100-kb sliding windows.

Supplementary Table 3. DNA methylation-related genes in *A. thaliana* [a] and *B. distachyon.*

Gene name	*A. thaliana* Gene ID	Activity	Sequence context	*B. distachyon* best-hit ID
MET1	At5g49160	DNA methyltranferase	CG	Bradi1g55290
CMT3	At1g69770	DNA methyltranferase	CHG, CHH	Bradi3g21450
DRM1	At5g15380	DNA methyltranferase	CHH	Bradi4g05680
DRM2	At5g14620	DNA methyltranferase	CHH	Bradi4g05680
HOG1	At4g13940	S adenosyl homocysteine hydrolase	CG, CHG, CHH	Bradi4g19460
DDM1	At5g66750	Chromatin remodelling ATPase	CG, CHG, CHH	Bradi1g10360
DRD1	At2g16390	Chromatin remodelling ATPase	CHH	Bradi3g19890

[a] The genes are excerpted from Table 1 in reference (25).

[b] Genes identified in *B. distachyon* relative to the *A. thaliana* query, with BlastP $E<10^{-100}$ (9) (version 2.2.17)

Supplementary Table 4. Methylation differentiation between leaf and immature floral tissues across the entire genome.

Sequence context	Avg. difference between tissues (%)	Avg. difference among replicates within tissues (%)
CG	1.10	1.04
CHG	1.33	1.31
CHH	0.43	0.39

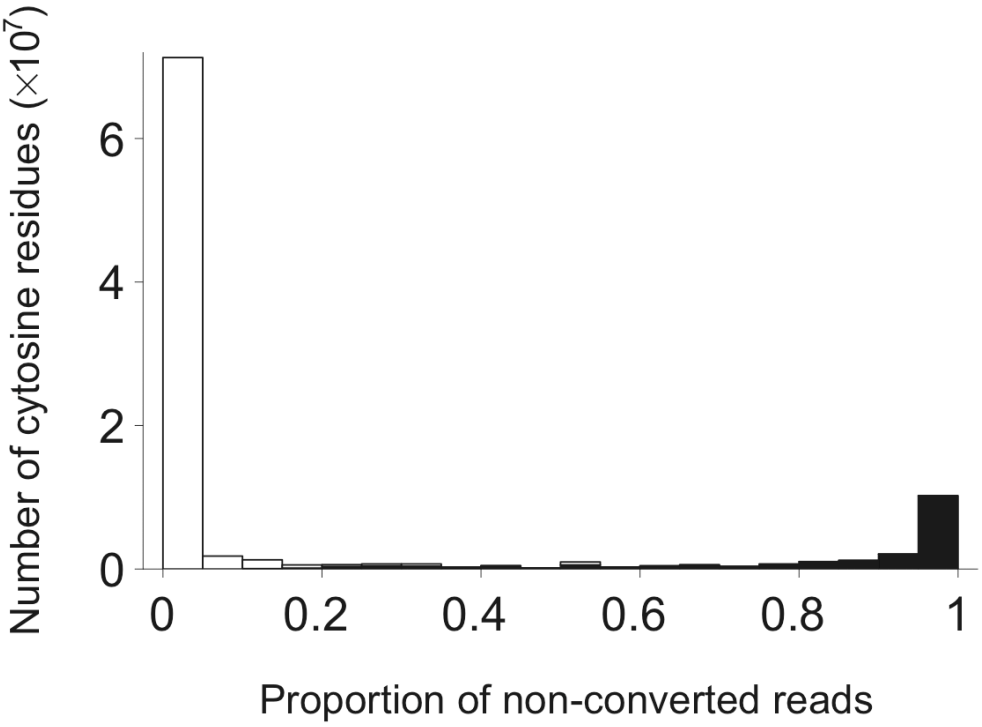

Fig. S1. The number of cytosines in *B. distachyon* (*y*-axis) vs. the proportion of non-converted reads at each site (*x*-axis). Blackened bars represent significantly methylated cytosine residues by the binomial test suggested by Lister et al. (2), which relies on the estimated error rates. In this study, the test was applied at the $p < 0.01$ level. The graph also shows that the vast majority of sites are non-ambiguous with respect to methylation status.

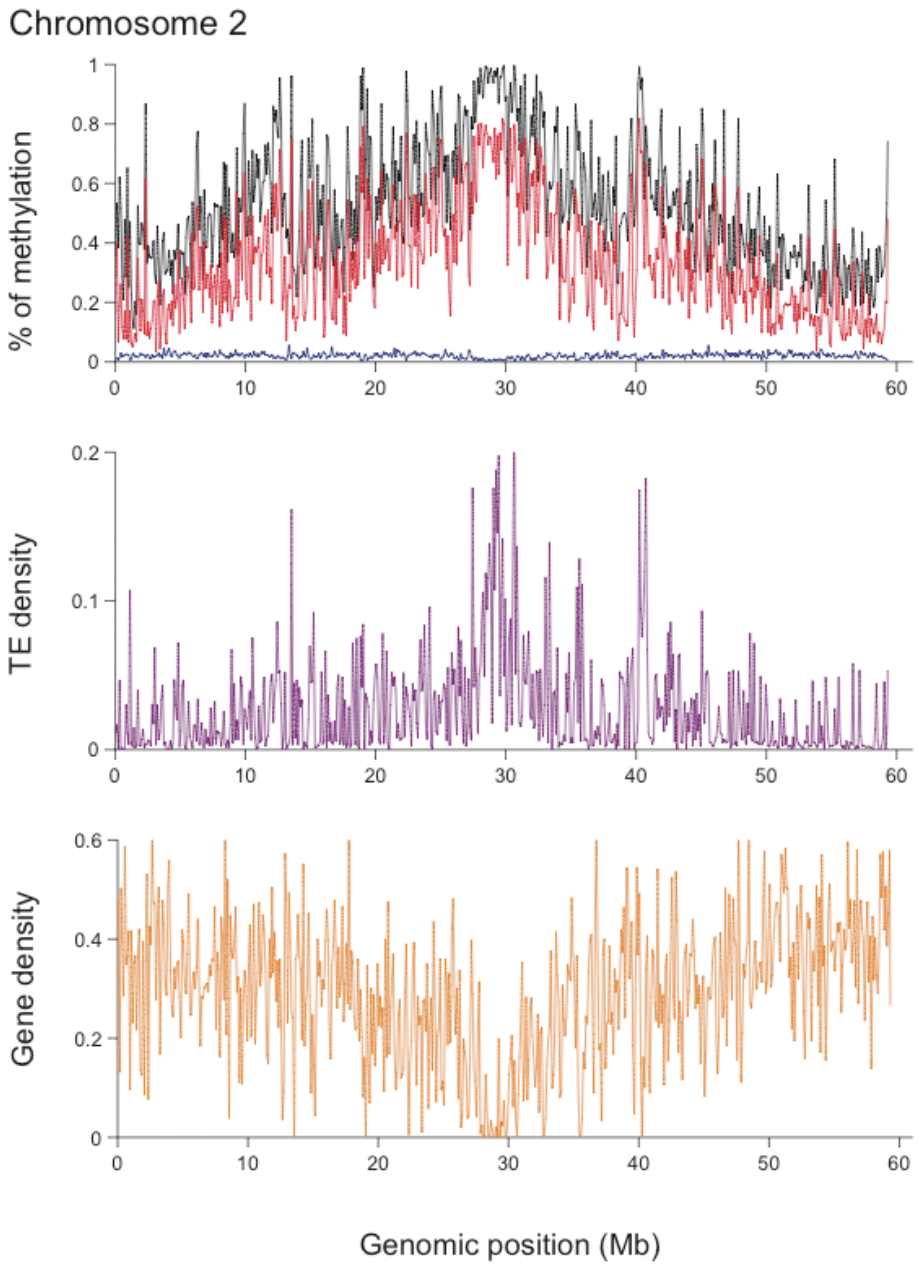

Fig. S2. Distribution of DNA methylation in chromosome 2-4 of *B. distachyon*. The first row shows methylation level at CG (red), CHG (black) and CHH (blue). The second row and third row show the density of TEs and genes, respectively. The size of non-overlapped windows is 100 kb.

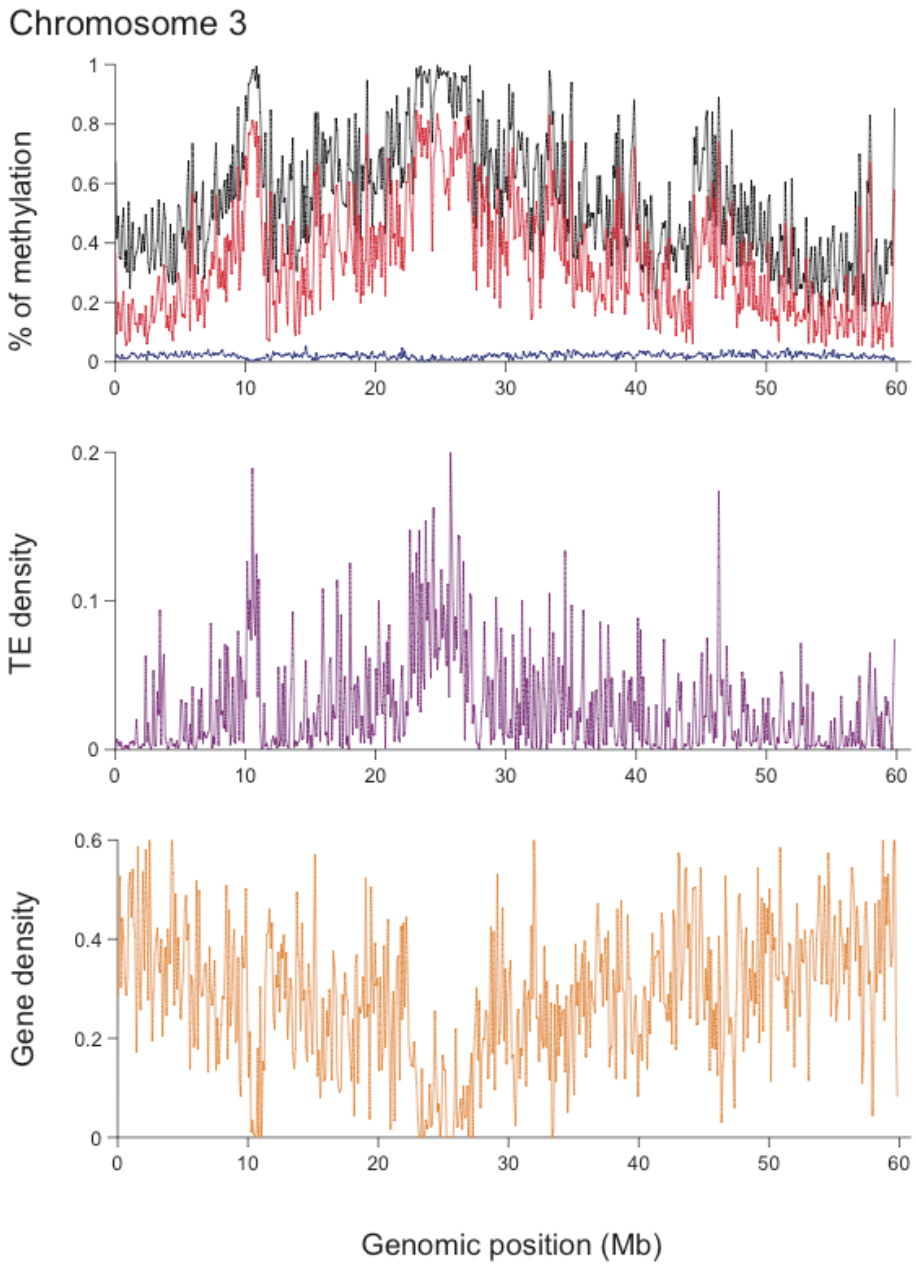

Fig. S2. (continued)

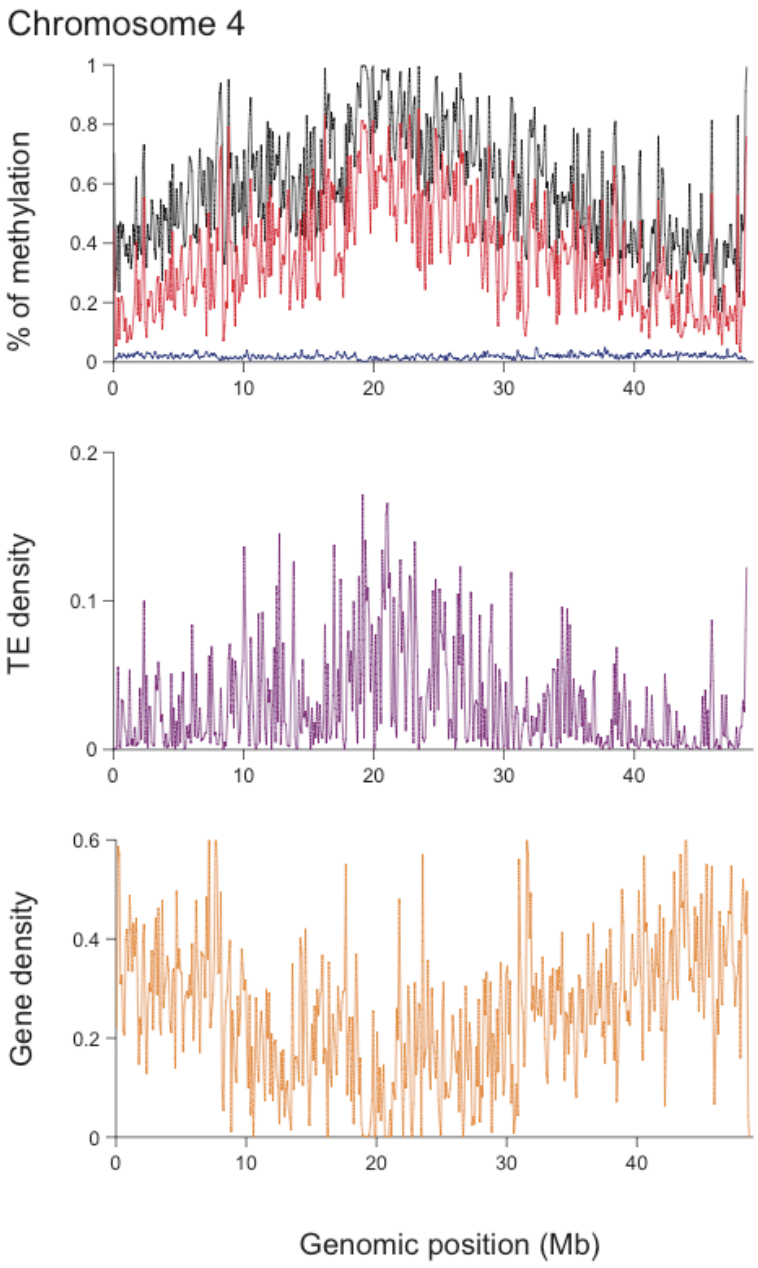

Fig. S2. (continued)

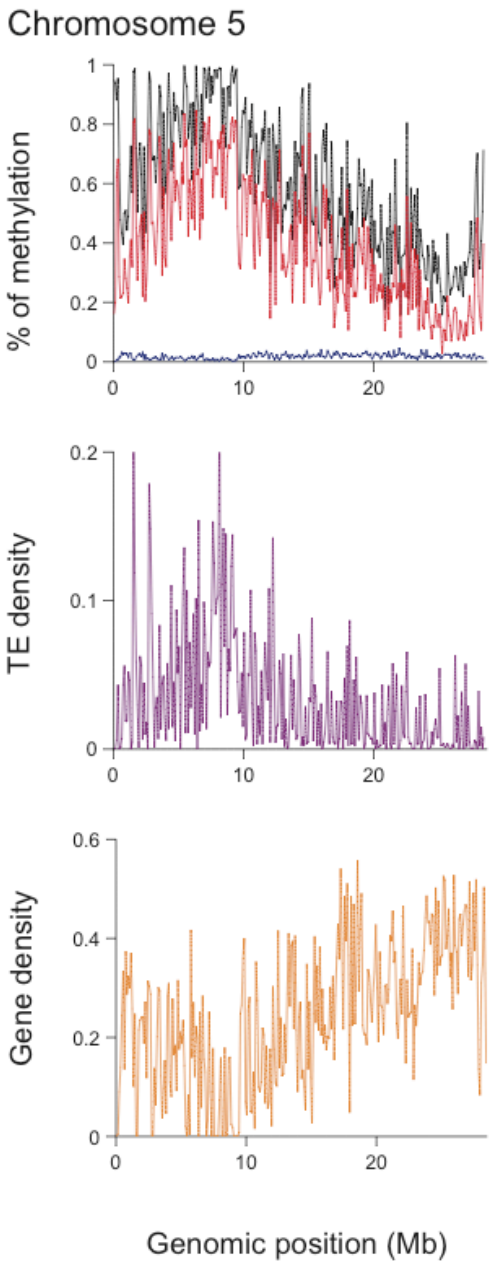

Fig. S2. (continued)

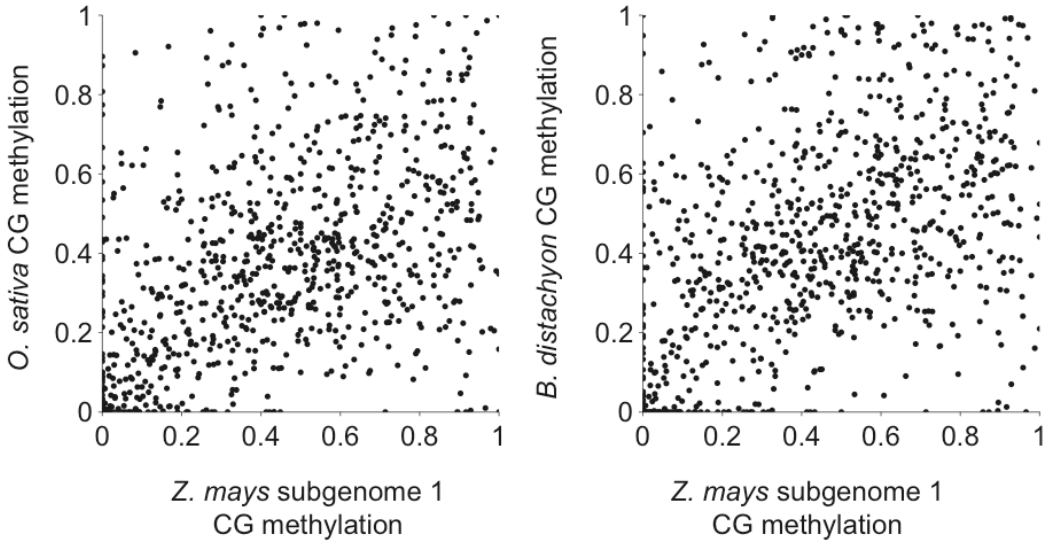

Fig. S3. Correlations in CG methylation across orthologs of maize and orthologs of either rice or *B. distachyon,* based on *n*=914 orthologs. Maize has two subgenomes (20). The correlations for maize vs. rice was $r = 0.510$ and for maize vs. *B. distachyon* was $r = 0.541$ ($P < 10^{-5}$ for both, based on permutation tests). We report analyses to orthlogs from maize subgenome 1, but procured qualitatively similar results from subgenome 2.

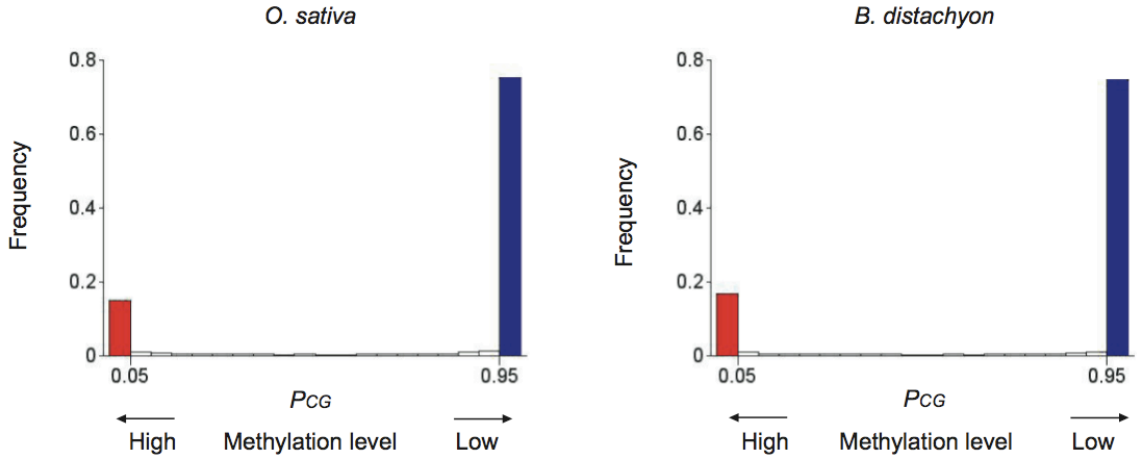

Fig. S4. The frequency distribution of P_{CG}, a significance test for body-methylation. Red and blue bars represent body- (BM) and under-methylated (UM) genes.

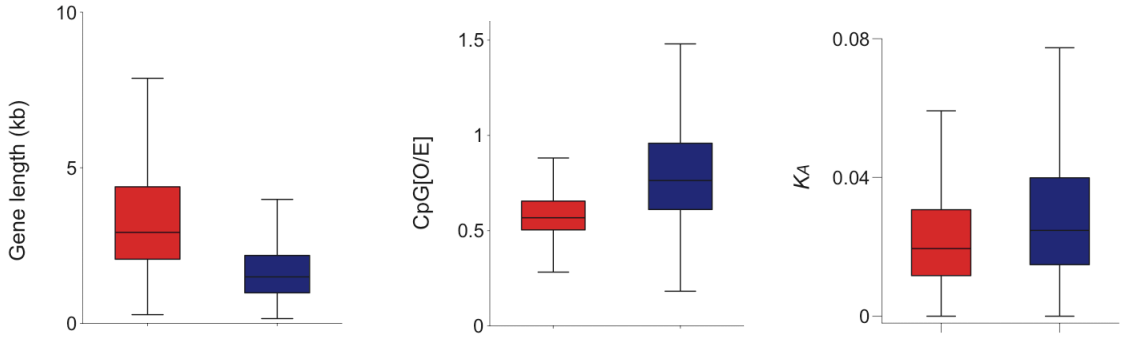

Fig. S5. Evolutionary analysis of BM (red) versus UM (blue) genes in *A. lyrata* orthologs of *A. thaliana* Col-0. Box plots show that BM genes are longer (left graph), have lower CG [O/E] ratios (middle graph), and diverge more slowly on average as measured by nonsynonymous divergence (K_A) (right graph).

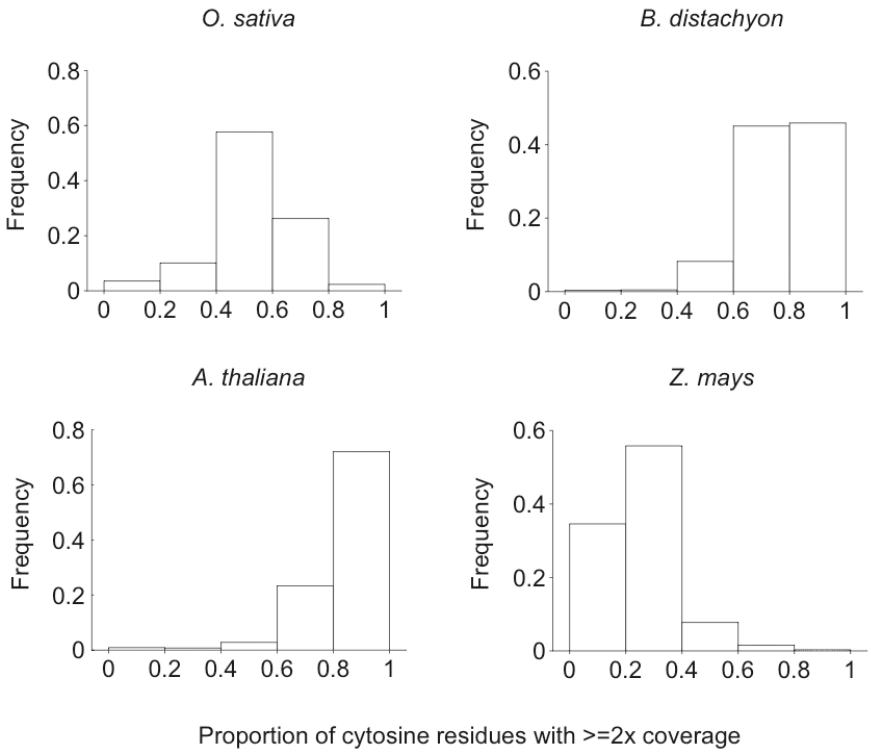

Fig. S6. The proportion of cytosine residues with ≥2 coverage of bisulfite short reads in rice, *B. distachyon, A. thaliana* and maize (*Z. mays*). For all graphs, the *x*-axes represents the proportion of cytosine residues that has ≥2-fold coverage. The *y*-axes represents frequencies of genes with given cytosine coverage. Based on these distributions, for analyses we included genes with ≥40% cytosines with ≥2-fold coverage for rice and with ≥60% cytosines ≥2-fold for *B. distachyon*. Using similar arguments, we used genes with ≥60% ≥2-fold coverage for the *A. thaliana* MA lines (for which an example of one of the 8 MA lines is shown) and genes with ≥20% ≥2-fold coverage for the maize data, which had lower coverage.

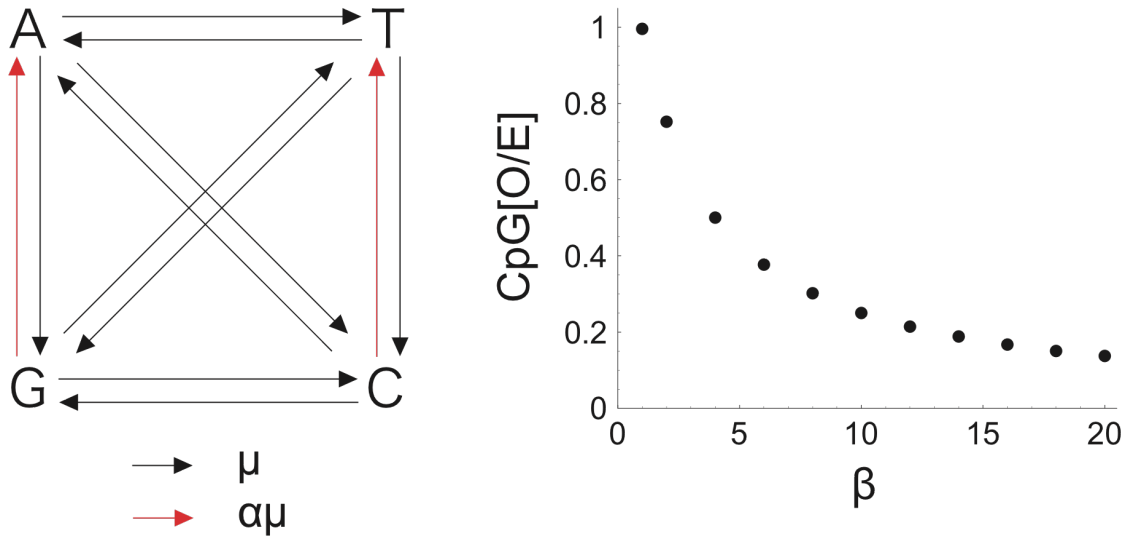

Fig. S7. A schematic diagram of the nucleotide substitution model (left) and simulation results (right). The nucleotide substitution model includes a separate parameter (α) for two types of transition (C->T and G->A) mutations, as well as a separate mutation parameter (β) at CG dinucleotide sites. The results in the graph on the right show that when transitional mutation rates at CG dinucleotides are ~5x higher than elsewhere in the sequences, then the equilibrium CG[O/E] level is about ~40%, which is similar to the observed level of CG[O/E] in rice, *B. distachyon* and *A. thaliana* BM genes.

SUPPORTING REFERENCES to SI Appendix

1. Harris EY, Ponts N, Levchuk A, Roch KL, Lonardi S (2010) BRAT: bisulfite-treated reads analysis tool. *Bioinformatics* 26:572–573.
2. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133:523–536.
3. Calarco JP, Borges F, Donoghue MT, Van Ex F, Jullien PE, Lopes T, Gardner R, Berger F, Feijo JA, Becker JD *et al.* (2012) Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell* 151:194–205.
4. Greaves IK, Groszmann M, Ying H, Taylor JM, Peacock WJ, Dennis ES (2012) Trans chromosomal methylation in *Arabidopsis* hybrids. *Proc Natl Acad Sci U S A* 109:3570–3575.
5. Li X, Zhu J, Hu F, Ge S, Ye M, Xiang H, Zhang G, Zheng X, Zhang H, Zhang S *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.
6. Takuno S, Gaut BS (2012) Body-methylated genes in *Arabidopsis thaliana* are functionally important and evolve slowly. *Mol Biol Evol* 29:219–227.
7. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452:215–219.
8. Fawcett JA, Rouze P, Van de Peer Y (2012) Higher intron loss rate in *Arabidopsis thaliana* than *A. lyrata* is consistent with stronger selection for a smaller genome. *Mol Biol Evol* 29:849–859.
9. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
10. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418–426.
11. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–480.
12. Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, Vandepoele K (2012) i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res* 40:e11.
13. Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* 8:1499–1504.
14. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115.
15. Ghent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, Dawe RK (2013) CHH

islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Research* in press
16. Chen PY, Cokus SJ, Pellegrini M (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* 11:203.
17. Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc Natl Acad Sci U S A* 94:6809–6814.
18. Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J (2004) On the tetraploid origin of the maize genome. *Comp Funct Genomics* 5:281–284.
19. Tenaillon MI, Hollister JD, Gaut BS (2010) A triptych of the evolution of plant transposable elements. *Trends Plant Sci* 15:471-478.
20. Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A* 108:4069–4074.
21. Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* 19:1419-1428.
22. Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327:92–94.
23. Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A* 48:582–592.
24. Gaut B, Yang L, Takuno S, Eguiarte LE (2011) The Patterns and Causes of Variation in Plant Nucleotide Substitution Rates. *Annu Rev Ecol Evol S* 42:245–266.
25. Furner IJ, Matzke M (2011) Methylation and demethylation of the *Arabidopsis* genome. *Current opinion in plant biology* 14:137–141.