Divergence of Gene Body DNA Methylation and Evolution of Plant Duplicate Genes

Jun Wang, Nicholas C. Marowsky, Chuanzhu Fan*

Department of Biological Sciences, Wayne State University, Detroit, Michigan, United States of America

Abstract

It has been shown that gene body DNA methylation is associated with gene expression. However, whether and how divergence of gene body DNA methylation between duplicate genes can influence their divergence remains largely unexplored. Here, we aim to elucidate the potential role of gene body DNA methylation in the fate of duplicate genes. We identified paralogous gene pairs from Arabidopsis and rice (Oryza sativa ssp. japonica) genomes and reprocessed their single-base resolution methylome data. We show that methylation in paralogous genes nonlinearly correlates with several gene properties including exon number/gene length, expression level and mutation rate. Further, we demonstrated that divergence of methylation level and pattern in paralogs indeed positively correlate with their sequence and expression divergences. This result held even after controlling for other confounding factors known to influence the divergence of paralogs. We observed that methylation level divergence might be more relevant to the expression divergence of paralogs than methylation pattern divergence. Finally, we explored the mechanisms that might give rise to the divergence of gene body methylation in paralogs. We found that exon methylation divergence more closely correlates with expression divergence than intronic methylation divergence. We show that genomic environments (e.g., flanked by transposable elements and repetitive sequences) of paralogs generated by various duplication mechanisms are associated with the methylation divergence of paralogs. Overall, our results suggest that the changes in gene body DNA methylation could provide another avenue for duplicate genes to develop differential expression patterns and undergo different evolutionary fates in plant genomes.


Editor: Shin-Han Shiu, Michigan State University, United States of America

Received June 24, 2014; Accepted September 22, 2014; Published October 13, 2014

Copyright: © 2014 Wang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: CF was supported by start-up fund from Wayne State University. 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: cfan@wayne.edu

Introduction

Epigenetic modifications of DNA and histones can inheritably regulate the access to the genetic information encoded by DNA [1–4]. DNA methylation, defined as adding a methyl group to the cytosine base of DNA to form 5-methylcytosine, is one of such epigenetic modifications [5]. In plants, cytosine can be methylated in three sequence contexts, CG, CHG and CHH (H = A, C, or T) [6]. DNA methylation can be established de novo and maintained by various methyltransferases, e.g. DRM2, MET1, and CMT3 in plants [7–11]. It can also be removed either passively by demethylation enzymes, e.g. DME and ROS1 in plants [12], or actively by demethylation enzymes, e.g. DME and ROS1 in plants [13,14]. DNA methylation is involved in various important biological processes such as repressing the expression of TEs and repetitive elements, and participating in earlyembryogenesis, stem cell differentiation, X chromosome inactivation, genomic imprinting, neuronal and cancer development [15–22].

DNA methylation can occur in the promoter and gene body regions of genes, which influence gene expression differently. DNA methylation in promoter regions is usually negatively associated with gene expression [23,24]. Whereas, in gene body extreme low or high DNA methylation level is associated with lower gene expression, while modest DNA methylation levels is related to higher expression [25–28]. However, the underlying mechanisms of gene body methylation in regulating gene expression have not been well understood. The causal relationships between gene body methylation and gene expression have been found in preventing transcriptional initiation from alternative promoters within genes, ensuring the accuracy of alternative-splicing and hindering transcriptional elongation [26,29,30]. Although an alternative hypothesis claimed that gene body methylation could be the by-product of expression [31,32], the conservation of gene body methylation among orthologs in evolutionarily distant species and higher proportion of body-methylated genes with phenotypic effects supported that gene body methylation may play certain functional roles [33,34].

Gene duplication plays a critical role in the origination of functional novelties in organisms. How duplicate genes evolve and become fixed in a genome is one of the central questions in molecular evolution [35]. Duplicate genes can be fixed through genetic drift or subfunctionalization/neofunctionalization driven by natural selection [35–45]. Previous studies revealed that the expression and function divergence of duplicate genes can be achieved through various mechanisms, such as nucleotide substitution, cis-regulation, post-translational regulation, promoter
epigenetic marks and so on [37,40,46,47]. A recent study demonstrated that the DNA methylation in promoter may play a significant role for functional divergence of duplicated genes in human [48].

In previous studies of methylation in duplicate genes, Wildman et al (2009) was the first to investigate the methylation pattern conservation between duplicate genes in Arabidopsis [49]. Chang and Liao (2012) discovered that DNA methylation in upstream flanking regions of genes repressed and thus “rebalance” the overall expression dosage of paralogs [46]. Wang et al. (2013) found that ~20% of the lineage-specific new duplicate genes had methylation pattern significantly more divergent from their parental genes in gene body than the methylation conservation of all the paralogs in Arabidopsis [50]. Wang et al. (2013) examined the divergence of gene body methylation levels of duplicate genes generated from different mechanisms, and the relationship between gene body methylation patterns and synonymous substitutions of paralogs [51]. However, it remains to be explored whether and how DNA methylation, particularly in gene body, influences the fate of duplicate genes in plants [52–54].

Here, we aim to determine how divergence of DNA methylation level and pattern in gene body is associated with expression divergence and evolution of duplicate genes in plant genomes. We chose rice (O. sativa ssp. japonica) and Arabidopsis thaliana as the model systems, which have high abundance of gene duplication and modest DNA methylation level in the genomes (in CG context, the methylation levels of rice and Arabidopsis are 23.27% and 38.21%, respectively, see Table S1 and [27,55]). Our results show that gene body methylation divergence is associated with divergence of duplicate genes, suggesting that gene body methylation might play an important role in the expression divergence and evolutionary fate of duplicate genes in plant genomes.

Materials and Methods

Plant species selected and genome sequences, bisulfite-sequencing (BS-seq), expression and small RNA data sets

We selected two species, A. thaliana, and O. sativa ssp. japonica for our analyses. We downloaded their genome, CDS, protein sequences, and “.gff” files from Phytozome v8.0 http://www.phytozome.net with A. thaliana 167 (TAIR10) for A. thaliana, and http://rice.plantbiology.msu.edu/downloads_gad.shtml (MSU7) for O. sativa ssp. japonica. We downloaded the BS-seq raw data from the NCBI Short Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra/) with accession numbers SRA000284 (immature floral tissue) for Arabidopsis [55] and SRA012190 (panicle) for O. sativa ssp. japonica [27], and from the NCBI Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) with accession numbers GSM560562 (embryo), GSM560563 (endosperm), GSM560564 (seedling roots), and GSM560565 (seedling shoots) for rice [28,56]. We downloaded the RNA-seq raw data from SRA with accession numbers SRA000286 (immature floral tissue) for Arabidopsis [55], and digital gene expression (DGE) raw data with accession number GSE20871 (panicle) from the NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) for O. sativa ssp. japonica [27]. We downloaded Arabidopsis Affymetrix expression data of 56 microarray tissues/conditions that were not based on genetic mutants or overlapping tissues from http://www.ebi.ac.uk/arrayexpress/experiments with accession number E-AFMX-9 [57]. We downloaded the RNA-seq data of endogenous small RNAs in panicle of rice from GSE32973 (GSM816731) of GEO [58], and that of small RNAs in wild-type immature flower of Arabidopsis from GSM277608 of GEO [55], which are the same types of tissues, where the methylation data were measured for the two species.

Process methylation data, expression data and small RNA data

We re-analyzed the BS-seq data for Arabidopsis and O. sativa ssp. japonica with Bismark [59]. The intermediate steps included (1) running quality control and trimming the low quality bases with trim_galore (Version 0.2.5) (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), (2) mapping the reads with Bismark v0.7.7, (3) removing the duplicates generated by PCR (deduplicate_bismark_alignment_output.pl), (4) generating cytosine methylation reports with Bismark v0.7.7. The plant chloroplast genome has no methylation activity [60], so any methylation reads detected in chloroplast genome should be accounted for the error. To test whether a cytosine is methylated, we conducted binomial test for each cytosine site based on the number of methylated reads, non-methylated reads and the error rate estimated from the chloroplast genome. To correct for multiple comparison problem, we computed the false discovery rate (FDR) ‘p’ value for each binomial test ‘p’ value [61–63]. The ‘q’ value=0.05 was taken as the criterion of the methylated cytosine [27,55]. We only considered the cytosine mapped with ≥ 5 BS-reads.

We reprocessed the raw RNA-seq data of Arabidopsis and DGE data of O. sativa ssp. japonica. For RNA-seq data, we trimmed the low quality data with trim_galore (Version 0.2.5), mapped the RNA-seq reads to Arabidopsis genome with Bowtie-0.12.8 [64], removed duplications due to PCR with picard-tools-1.79 (http://picard.sourceforge.net/), and finally we used cufflinks (v2.0.2) [65] to calculate gene-level relative abundance of reads in Fragments Per Kilobase million mapped fragments (FPKM) format. For reprocessing DGE data, we chose the longest transcript for each gene in O. sativa ssp. japonica genome to build the reference sequences, then mapped DGE reads to the reference sequences with Bowtie-0.12.8, finally counted the number of DGE reads for each gene and normalized the number of DGE reads with the gene length.

We annotated Affymetrix microarray expression data to Arabidopsis genes with customized CDF file downloaded from http://brainarray.mbi.uc.edu/Brainarray/Database/CustomCDF/17.1.0/taig.asp. We processed the expression data with RNA package of R. The expression specificity index of each gene was computed according to the approach by Yanai et al [66]. Namely, we computed the tissue specificity index, τ.

\[
\tau = \frac{N}{\sum_{i=1}^{N} (1 - \chi_i)}
\]

where N is the number of tissues, and \( \chi_i \) is the expression component (EC) in the ith tissue normalized by the maximal EC value among all the tissues. For the detailed steps to compute \( \chi_i \) and \( \tau \), please refer to Yanai et al’s paper [66]. The \( \tau \) value of each gene ranges from 0 to 1. The higher the \( \tau \) value, the more specifically expressed the gene is [66].

We collected the processed counts of each small RNA in the genomes. We then mapped the small RNA sequences to the genomes of corresponding species with BWA (bwa-12–17–2013- git) [67]. Based on the above two information plus the coordinates of gene annotations, we estimated the number of small RNAs mapped in each paralogous gene.
Identification and classification of paralogous gene pairs in the two species

We used all the peptide sequences of the genome to Blat [68] against all the peptides for the two species, respectively. We discarded the self-hit pairs and only kept the gene pairs whose reciprocal best hits were each other (namely, we only kept the gene pair of A and B, when the best hit of A is B and the best hit of B is A in the genome) with ≥70% sequence coverage and ≥50% sequence identity at the protein level. The Blat identity score was calculated following the formula in http://genome.ucsc.edu/FAQ/FAQblat.html#. We classified the duplicate gene pairs according to four duplication mechanisms: dispersed duplication, retrotransposition, tandem duplication and segmental duplication. If a group (≥2) of genes, where any adjacent ones are within at most 10-gene distance, and their corresponding paralogs can also be clustered together with any adjacent ones within at most 10-gene distance, we defined them as generated by segmental duplication. If query and target genes of a paralogous pair were adjacent to each other, we defined them as generated by tandem duplication. If one duplicate gene had one exon, and the other copy had multiple exons, and intron losses in the CDS regions of the two genes were identified, we defined them as generated by RNA-based retrotransposition. The rest of paralogous gene pairs, both of which have single or multiple exons and are not belong to segmental, tandem duplication or retrotransposition, were defined as dispersed duplication. If a paralogous pair was classified into more than one duplication mechanism, we didn’t count this pair in the analyses of duplication mechanisms. We used Codeml of PAML4.7 [69] to compute Ka, Ks, and K for all the paralogs.

Defining the methylation level/pattern divergence

For methylation level comparison, we only considered the paralogous pairs, where both genes have ≥50% of the cytosines mapped by BS-seq. We conducted Fisher test using 2×2 contingency table to compare numbers of methylated cytosines and non-methylated cytosines (in CG, CHG, CHH context, respectively) in the gene bodies of two paralogous genes. We used FDR (the ‘q’ value of the Fisher test 'p') to correct for multiple test problems. With the ‘q’ value of the Fisher test 'p' value≤0.05, we classified the paralogous pairs as the ones with divergent methylation level, otherwise as the ones with conserved methylation level. And we used the Fisher test 'p' value as the proxy of methylation level conservation to perform other analyses.

To define the paralogs with conserved methylation level in the promoter regions, we first chose different cutoff length of the upstream transcription starting sites for two species respectively, and gene expression level in these regions. We identified that methylation level conservation to perform other analyses.

FDR (the ‘q’ value of the Fisher test 'p') to correct for multiple test problems. We defined the paralogs with non-methylated cytosines in the promoter regions at 5’ upstream regions of 500 bp in rice and 5000 bp in Arabidopsis. We used the Fisher test 'p' value as the proxy of methylation pattern conservation for all the paralogs. We conducted a binomial test with this average percentage, the number of cytosines with conserved status, and the number of cytosine with non-conserved status. We computed the FDR 'q' value associated with the 'p' value of binomial test. We defined the paralogs with pattern conservation percentage significantly (FDR 'q' ≤0.05) lower than the average value as the ones with non-conserved methylation patterns, and the rest as those with conserved methylation patterns.

We further divided paralogous gene pairs into conserved highly methylated, conserved lowly methylated, and conserved non-methylated pairs. If the methylation of both paralogous copies was zero, we classified the pair into the conserved non-methylated group. If the methylation levels of both copies of a pair were higher than the median methylation level of all paralogs with conserved methylation level/pattern, we classified this pair into the conserved highly methylated group. And if they were lower than the median, we classified the pair into the conserved lowly methylated group.

Population genetics analysis and mutation frequency estimation

We obtained A. thaliana single nucleotide polymorphism (SNP) data generated from a complete re-sequencing of 80 strains using next-generation sequencing technology [71], downloading from http://1001genomes.org/data/MP/L/MPICao2010/releases/current/. We then computed the \( \theta_w \) values for duplicate genes in Arabidopsis. \( \theta_w \) was defined as the number of segregation sites, which can be estimated using SNP data, divided by \( a_w \) where

\[
a_w = \frac{1}{\sum_{i=1}^{n-1} i}
\]

To estimate the G/C to A/T mutation frequency of A. thaliana genes, we compared A. thaliana to its A. lyrata orthologous sequences. The extracted genomic sequences of A. thaliana genes were Blat against A. lyrata genome. The genomic sequences of A. thaliana genes and their A. lyrata orthologous sequences were identified and then aligned. We collected the SNPs of the aligned regions of these A. thaliana genes. For each SNP, the derived allele was distinguished from ancestral using the A. lyrata orthologous sequence as outgroup. We only considered the G/C→A/T mutations in the sites of the A. thaliana gene. The mutation frequency of G/C→A/T was calculated from the quotient of the total number of G/C→A/T mutations and the total number of G/C bases in the aligned region of A. thaliana genes.

Analysis of TE distribution of paralogs

We downloaded the repeat library repeatmasker-libraries-20130422 from http://www.girinst.org/ and mapped the repeat sequences to the genome with RepeatMasker 4.0.1 [73]. Based on the mapping result, we divided paralog pairs into three categories: 1) both copies contain TEs/repeat sequences within 1 kb upstream or downstream region; 2) neither copy has TEs/repeat sequences within 1 kb upstream or downstream region; 3) only one of the two copies has TEs/repeat sequences within 1 kb upstream or downstream region. We conducted further analyses based on this classification. Additionally, we used different cutoff length of paralog flanking region to search for the nearby TEs and examined the reliability of our analyses.
Statistical analyses

Correlations were measured with Spearman’s rank correlation method. We used “ppcor” package in R to compute partial correlation [51,74,75]. ‘ppcor’ can calculate the pairwise partial correlations for each pair of variables while controlling a third or more other variables with three correlation methods (i.e. Pearson, Kendall, and Spearman). Each variable has mathematical expressions and variances to compute the partial correlation with the response variable [76]. It provides the ‘p’ value as well as statistic for each pair of variables.

The goodness of fit was measured with R-squared (coefficient of determination, $R^2$), and coefficient significance of linear regression with ‘lm’ function in R. $R^2$ is defined as the percentage of the response variable variation that can be explained by a statistical model, so $R^2 = \frac{\sum (y_i - f_i)^2}{\sum (y_i - \bar{y})^2}$. The binomial test of the significance of paralog pairs overlapped in the same categories of the two species was conducted using ‘binom.test’ function in R. The expected probability to overlap is $P = (\text{conserved speciesA\%}*\text{conserved speciesB\%} + \text{non-conserved speciesA\%}*\text{non-conserved speciesB\%})$. The number of paralog pairs with data available in both species A and species B is ‘N’, and the observed number of paralog pairs with same categories in the two species is ‘X’.

All the intermediate steps were carried out with Perl and R scripts.

Results

Plant duplicate genes identification and gene body DNA methylation measurement

We identified total 3459 and 2911 paralogous gene pairs in Arabidopsis and rice genomes, respectively. We first reprocessed the BS-seq data generated from the immature floral tissue of Arabidopsis and the panicle of rice. Our reprocessed results are consistent with previous reports (Table S1) [27,55]. We assessed the DNA methylation divergence between the two paralogous genes from two aspects: (1) methylation level, which estimates the overall methylation level difference in the entire gene body; (2) methylation pattern, which estimates the effect of methylation change in specific cytosine sites. The methylation level was defined as the percentage of methylated cytosines over the total cytosines mapped by BS-seq, for each of the cytosine contexts respectively. Methylation pattern was measured by the percentage of cytosines with conserved methylation status (both methylated/neither methylated) among all mapped aligned cytosines in the coding sequences (CDS) of paralogs, for each of the cytosine contexts respectively. Gene body is typically only methylated in the CG context, so methylation on cytosine in CG context is most abundant in gene body [55,77] and CG methylation in genic region is associated with up-regulation of gene expression [78]. Whereas, CHG and CHH methylation is often associated with RNA-directed DNA methylation and represses gene transcription [5]. Therefore, we focused on CG methylation in the following analyses. Our results showed that the methylation levels of all duplicate genes are bi-modally distributed for both species (Figure 1).

Further, to examine the consistency of gene body methylation level of duplicate genes across different tissues, we reprocessed the raw BS-seq data from embryos, endosperm, seedling roots, and seedling shoots in rice. We pair-wisely plotted the gene body methylation levels of rice duplicate genes among the four tissues with the contour plots. The methylation levels between any pair of the tissues are highly correlated (Spearman’s rank correlation coefficients of all pairs of tissues $>0.96$, $p<2.2e-16$, Figure 2). This result suggested that the majority of duplicate genes have similar methylation levels across different tissues with few genes with tissue-specific methylation pattern, which is consistent with previous results for all the genes [34,51]. Therefore, it is rational to use the methylation level from one tissue (e.g. the immature floral tissue of Arabidopsis and the panicle of rice) as the representative to study the general patterns of methylation for a large set of genes in the following analyses.

Gene body DNA methylation correlates with gene properties

To investigate whether the methylation divergence influences the divergence of duplicate genes, we first measured the relationship between DNA methylation and the properties of duplicate genes including gene structure, transcription and mutation rate.

DNA methylation and gene structure. Previous studies revealed inconsistent relationships between gene length/exon number and methylation. For example, in honeybee and silkworm, highly methylated genes were significantly longer than highly methylated ones; whereas in the sea squirt and sea anemone, highly methylated genes were significantly longer than highly methylated ones [79]. And in Arabidopsis body-methylated genes are longer and have more exons [33,80]. Using smoothing spline function in R with different degrees of freedom, we plotted the regression line between the methylation level and gene length/exon number for duplicate genes in Arabidopsis and rice, respectively [51]. The methylation level 0.6 was present as the inflection point dividing the positive and negative correlation with gene length/exon number (Figure 3). Thus we investigated this relationship by separating the duplicate genes into two groups with methylation level $<0.6$ and $\geq 0.6$. We found that methylation level positively and negatively correlates to gene length/exon number for the two groups of genes, respectively. Total number of duplicate genes in Arabidopsis and rice are 6498 and 4992, respectively. Spearman’s rank correlation $p<2.2e-16$ and $R^2$ of linear regression model in the two methylation intervals are in the range of 0.14–0.38. The linear coefficients are statistically significant with $p$ value $<2e-16$ (Table S2). However, the data size of duplicate genes with methylation level $\geq 0.6$ was small with 247 in Arabidopsis and 631 in rice, and 26% – 43% of them fell into the range of methylation level $= 1$ (Figure 3), which is corresponding to the methylation level distribution of duplicate genes (Figure 1). Therefore, given the biased sampling, it needs caution to interpret the correlation in this interval. Further, to explore the relative association of gene length and exon numbers with methylation level and their interactions, we examined the partial correlation among the methylation level, gene length and exon number with the “ppcor” package in R. We found gene length and exon number are highly correlated with Spearman’s rank correlation coefficient around 0.7 and $p=0$. Gene length is more closely correlated to methylation level than exon number.

Our results show that the gene length/exon number does not linearly correlate to methylation level and immediately methylated genes are longer and contain more exons in plants. The underlying mechanisms of gene body methylation remain to be explored. However, our analyses imply that gene body methylation may prevent transcriptional initiation from cryptic promoters within genes. Therefore, longer genes with more exons require higher methylation level to ensure their precise expression. Thus methylation level is positively correlated with gene length/exon number.
number below certain threshold. Meanwhile, extreme high methylation level could hinder transcriptional elongation, so shorter genes can be associated with higher methylation level as well. Thus methylation level is negative correlated with gene length/exon number after gene body methylation level passes certain threshold.

**DNA methylation and gene expression.** We measured the correlation between gene body methylation and gene expression in terms of expression level and specificity in *Arabidopsis*. Expression level was measured with RNA-seq in the immature floral tissue where the methylation was measured. Whereas, the expression specificity was computed based on 56 tissues/conditions. Previous studies showed that moderately methylated genes had higher expression level, and extremely high or low methylation accompanied lower gene expression [27]. We plotted correlation between methylation level and gene expression level/specificity. Methylation level 0.5 appears as the turning point of the curves, which was determined based on smoothing spline regression (Figure 4). Therefore, we separated the duplicate genes into two groups with methylation level <0.5 and ≥0.5. For the ones with methylation level <0.5, expression level positively and expression specificity negatively correlate with methylation level; and for the ones with methylation level ≥0.5, expression level negatively and expression specificity positively correlate with methylation level. Spearman’s rank correlation \( p < 0.05 \) and \( R^2 \) of linear regression model in the two methylation intervals are in the range of 0.0079–0.2873 (Table S3). The linear coefficients are statistically significant with \( p \) value 1e-5 (Table S3). Additionally, the expression level and specificity strongly negatively correlate with each other (Spearman’s rank correlation coefficient: \( -0.6845 \) \( p < 2.2e-16 \)).

The small \( R^2 \) between gene body methylation and expression may be due to regulation of gene expression by multiple factors, so the contribution of gene body methylation to expression is relatively small and/or mirrored by other confounding factors. Thus, we examined the general correlation between expression levels/specificity and other factors including small RNA abundance (we focused on the 21nt and 24nt small RNA here, see explanation in the later section), promoter methylation, gene length, exon number, and mutation rate. Mutation rate was measured by theta (\( \theta \)) estimated from SNPs [81]. We found that expression level was significantly negatively associated with the 21nt and 24nt small RNA abundance, promoter methylation level, and mutation rate (\( p < 2.2e-16 \)), but was positively associated with gene length and exon number (\( p < 2.2e-16 \), Table S4). Expression specificity was significantly negatively associated with gene length and exon number (\( p < 2.2e-16 \), but was positively associated with 21nt and 24nt small RNA abundance, promoter methylation level, and mutation rate (\( p < 1e-7 \), Table S4).

To control for the effects of the above factors and test whether gene body methylation exclusively correlates to gene expression, we computed the partial correlation between methylation level and expression level/specificity, considering all the above factors simultaneously using the “ppcor” package of R [74,75]. Our analyses show that the partial correlations between methylation and expression are similar but weaker than the above general correlations (Spearman’s rank correlation coefficients. Methylation level <0.5: vs. expression level: 0.1375, \( p = 6.760e-16 \); vs. expression specificity: \( -0.1694 \), \( p = 3.9333e-17 \). Methylation level ≥0.5: vs. expression level: \( -0.0898 \), \( p = 0.1102 \); vs. expression specificity: 0.0934, \( p = 0.1858 \)). These results suggest that gene body methylation is indeed associated with gene expression, and the associations for duplicate genes are consistent with the overall association for the non-TE genes revealed by the previous studies [25–28,33,77].

**Figure 1. Histograms of methylation level of all duplicate genes in Arabidopsis and rice.**

![Histograms of methylation level of all duplicate genes in Arabidopsis and rice.](image-url)
DNA methylation and gene mutation rate. To investigate the association between methylation level and sequence changes of duplicate genes, we looked into the correlation between methylation level and DNA sequence mutation rate, $h_w$, in Arabidopsis as the SNP data corresponding to the current reference genome was only available in Arabidopsis. Given the inflection point of the regression line between $h_w$ and methylation level as 0.5 (Figure 5), we also split the duplicate genes into two groups with methylation level $<0.5$ and $\geq0.5$. Methylation level negatively and positively correlates with $h_w$ for the two groups of genes, respectively. Spearman’s rank correlation $p<1e^{-7}$ and $R^2$ of linear regression model in the two methylation intervals are in the range of 0.0070–0.4304. The linear coefficients are statistically significant, $p$ value $< 1e^{-7}$ (Table S5). However, the correlation of $h_w$ and other factors including exon number, gene length, expression, small RNAs and promoter methylation level could be the cofounding factors between methylation level and $h_w$ (Table S6). After controlled all these factors simultaneously, we still observed the consistent partial correlations between $h_w$ and methylation level (Spearman rank correlation coefficients: Methylation level $<0.5$: 0.0004, $p = 9.7994e-01$. Methylation level $\geq0.5$: 0.2241, $p = 4.6118e-05$). This result suggests that highly methylated genes have higher mutation rates and intermediate methylated genes have lower mutation rates. This is conceivable, since highly methylated genes may experience higher mutation rate due to the mutagenic property of the methylated cytosines, which frequently mutate to thymine [82–84]. To further test this speculation, we computed frequency of G/C→A/T mutation as the percentage of G/C sites with G/C→A/T mutation among all the G/C sites for Arabidopsis genes based on Arabidopsis SNP data and A. lyrata genome sequence (see Method and Materials). We found that when methylation level $<0.5$, G/C→A/T mutation frequency is negatively correlated with methylation level (Spearman rank’s correlation coefficient $= -0.1853$, $p<2.2e^{-16}$). Whereas, G/C→A/T mutation frequency is positively correlated with methylation level when methylation level $\geq0.5$ (Spearman rank’s correlation coefficient $= 0.1100$, $p = 0.01869$). Thus, this result supported the above conjecture, suggesting methylation and DNA mutation may be able to interact with each other.

Methylation divergence positively correlates with genetic divergence of paralogs

We estimated how divergence of methylation level/pattern was associated with genetic divergence of paralogs, measured by synonymous (Ks) and non-synonymous (Ka) substitution rate. Previous studies suggested that orthologous genes with high methylation levels in both species evolved more slowly than those with low methylation levels in both species, or those with high methylation level in one species but low methylation level in the other species [33,79]. Wang et al (2013) found that the
methylation level divergence of paralogs was positively correlated with Ks given the gene physical distance (measured by the number of genes between two paralogs) and duplication mechanisms [51]. Consistent to previous study [51], we observed that the physical distance of paralogs negatively correlated with the conservation of methylation level and pattern (Spearman’s rank correlation coefficients. Arabidopsis methylation level: $-0.0582$, $p = 4.375\times 10^{-03}$ and methylation pattern: $-0.1068$, $p = 1.937\times 10^{-07}$. Rice methylation level: $-0.0217$, $p = 0.3082$ and methylation pattern: $-0.0848$, $p = 7.708\times 10^{-05}$). Further, we found that the physical distance positively correlated with Ks or Ka (Spearman’s rank correlation coefficients. Arabidopsis Ks: $0.4639$, $p = 2.2\times 10^{-16}$ and Ka: $0.0586$, $p = 4.076\times 10^{-4}$. Rice Ks: $0.2799$, $p = 2.2\times 10^{-16}$ and Ka: $0.0033$, $p = 0.8775$).

Physical distance could be a confounding factor for the relationship between Ks or Ka and methylation conservation. Therefore, we computed the partial correlation between Ks or Ka and methylation level/pattern conservation given physical distance. We found that methylation level conservation negatively correlates with both Ks and Ka in both genomes, but methylation pattern conservation only negatively correlates with Ks and Ka in rice (Spearman’s rank correlation coefficients, $p<0.05$, Table 1; also see Figure S1). Theoretically, Ks implies the divergent time or neutral DNA evolutionary rate, and Ka suggests the amino acid changes. Thus, these results suggest that methylation level divergence of paralogs may increase along with evolutionary time and correlate with the DNA and protein sequence divergence. For example, genetic variables could give rise to methylation divergence or vice versa as we showed the correlation between methylation level and mutation rate. Further, no significant correlations between methylation pattern divergence and Ks in Arabidopsis could be explained by the process of recurrent methylation and demethylation at specific sites/regions in short period [85].

Methylation divergence is associated with the expression divergence of paralogs

As shown by the above analyses that gene body methylation indeed correlated with gene expression, we examined whether divergence of gene body methylation contributes to expression divergence of paralogs. We divided paralogs into four categories based on the conservation of methylation level/pattern in gene body: conserved highly methylated (CHM), conserved lowly methylated (CLM), conserved non-methylated (CNM), and non-conserved methylated (NCM) (see Materials and Methods). The numbers of paralogous pairs in the four categories for the two species were listed in Table 2.

We further test whether the patterns of methylation divergence of duplicate genes were consistent across different tissues in rice. Interestingly, for methylation level divergence, most duplicate gene pairs (83.16%–89.50%) were classified into the same categories (either conserved or non-conserved) for any pair of the tissues (e.g. embryos, endosperms, seedling roots, and seedling shoots). For methylation pattern, similarly, the majority of duplicate gene pairs (88.22%–92.63%) were in the same categories.
for any pair of the tissues. Binomial tests indicate that the proportion of these observed gene pairs in either methylation level or pattern are significantly larger than random expectation (p value, 2.2e-16). This result suggested that duplicate genes in general maintained the same patterns of methylation divergence across different tissues, which is consistent with the observation in human tissues [48] and was first reported in plants here. It also implies that it is feasible to use the methylation divergence of duplicate genes in one tissue to study the general relationship between methylation divergence and expression divergence.

We computed the expression level fold change and expression specificity difference between the two paralogous genes as the two proxies of their expression divergence, and compared the two proxies among the four categories. The expression level fold change was determined by the absolute quotient of expression level of two paralogous genes and the paralog with higher expression level was always assigned as numerator. We found the expression level and specificity changes of CHM paralogs were significantly lower than those of CLM, CNM and NCM paralogs for the two species in both methylation level and pattern (Figure 6. Wilcoxon rank sum test, p<0.05, the detailed p values see Table S7 and sample size see Table 2).

Cofounding factors that associated with methylation and expression divergence of paralogs

As shown previously, expression level/specificity could be affected by other factors, e.g. nucleotide substitution, difference in gene structure (gene length and exon number), cis-regulatory motifs, promoter methylation, small RNA abundance, and physical distance (Table S4). Thus expression divergence of paralogs could also result from these factors. We controlled these factors one by one and finally altogether, to determine whether and how the methylation divergence was associated with expression divergence of duplicate genes. Due to the similar behavior of CNM and CLM (Figure 6) and CNM paralogs may be not regulated by methylations, we only performed the comparison among CHM, CLM, and NCM in the following analyses.

The effect of gene structure. To control for the effect of gene structure difference, we choose the duplicate genes with the same ranges of exon number and gene length, which are not significantly different among the three categories in Arabidopsis and rice. We found the expression level changes of the CHM paralogs are significantly lower than those of the CLM and NCM ones for methylation level and pattern in both genomes (Wilcoxon rank sum test, p<0.05, Table S8).

The effect of nucleotide substitutions. To control for the effects of nucleotide substitutions, we calculated the nucleotide substitution rate (K) in CDS regions, for each of the paralogs, in Arabidopsis and rice. We choose the paralogs with the same range of K values (e.g. in the range of 0.05 to 0.2), which were not significantly different among the three categories (Table S9). We found the expression level differences/specificity differences of the CHM duplicate genes are significantly lower than those of the CLM and NCM ones in methylation level and pattern for both species (Figure 7). Wilcoxon rank sum test shows p<0.05 except the comparison of expression specificity differences between CHM
and CLM for methylation level \(p = 0.05418\), and between CHM and CLM for methylation pattern \(p = 0.05397\) (Table S9).

**The effect of cis-regulatory difference.** To control for the effect of *cis*-regulatory difference, we extracted the *cis*-regulatory binding sites of *Arabidopsis* duplicate genes, based on all the *cis*-regulatory binding sites in *Arabidopsis* genome downloaded from AGRIS [http://arabidopsis.med.ohio-state.edu/downloads.html](http://arabidopsis.med.ohio-state.edu/downloads.html). We picked the paralogous pairs whose two copies have the same types of *cis*-regulatory binding sites, we still observed that the expression level divergence of the CHM paralogs is significantly lower than that of the NCM ones in methylation level (Wilcoxon rank sum test, \(p = 0.0565\) with 101 CHM paralogs and 103 NCM paralogs). However, because the false positive rate of *cis*-regulatory binding site mapping is high, this analysis might not be able to completely and precisely control the effect of *cis*-regulatory difference.

**The effect of promoter methylation difference.** To control for the effect of promoter methylation difference, we choose the paralogous pairs with conserved methylation level (using Fisher test, see Materials and Methods) in their promoter regions in *Arabidopsis* and rice. Among them, we found that the expression level/specificity changes of the gene body CHM paralogs were significantly lower than those of the CLM and NCM ones in methylation level and pattern for both genomes.

![Figure 5. The plot of methylation level vs. mutation rate \(\theta_w\) for duplicate genes in *Arabidopsis*. The red, blue, and green lines are corresponded to smoothing spline regression from degree of freedom as 4, 6 and 8, respectively. doi:10.1371/journal.pone.0110357.g005](http://www.plosone.org/)

| Table 1. The partial Spearman’s rank correlations between methylation level/pattern conservation and substitution rates in *Arabidopsis* and rice. |
|---|---|---|---|---|
| **Rice** | Ks~level | Ka~level | Ks~pattern | Ka~pattern |
| Coefficient | \(-0.2218\) | \(-0.2175\) | \(-0.1175\) | \(-0.0702\) |
| \(p\) value | \(1.242e-26\) | \(1.241e-25\) | \(3.753e-08\) | \(0.0011\) |

<table>
<thead>
<tr>
<th><strong>Arabidopsis</strong></th>
<th>Ks~level</th>
<th>Ka~level</th>
<th>Ks~pattern</th>
<th>Ka~pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>(-0.1644)</td>
<td>(-0.1108)</td>
<td>(-0.0230)</td>
<td>(0.0212)</td>
</tr>
<tr>
<td>(p) value</td>
<td>(3.386e-16)</td>
<td>(4.872e-08)</td>
<td>(0.2634)</td>
<td>(0.3022)</td>
</tr>
</tbody>
</table>

Note: Correlation coefficient and \(p\) value were computed with Spearman’s rank sum correlation. doi:10.1371/journal.pone.0110357.t001
**Table 2.** The number of paralogous genes with different gene body methylation conservation.

<table>
<thead>
<tr>
<th>Paralogs</th>
<th>For methylation level</th>
<th>For methylation pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arabidopsis</td>
<td>Rice</td>
</tr>
<tr>
<td>CHM</td>
<td>633</td>
<td>403</td>
</tr>
<tr>
<td>CLM</td>
<td>627</td>
<td>402</td>
</tr>
<tr>
<td>CNM</td>
<td>758</td>
<td>159</td>
</tr>
<tr>
<td>NCM</td>
<td>797</td>
<td>729</td>
</tr>
</tbody>
</table>

Note: CHM-Conserved Highly Methylated; CLM-Conserved Lowly Methylated; CNM- Conserved Non-Methylated; NCM- Non-Conserved Methylated.

doi:10.1371/journal.pone.0110357.t002

**Figure 6.** Relationship between expression and methylation divergence of paralogs in *Arabidopsis* and rice. A: Distribution of expression fold changes of *Arabidopsis* paralogs in the three methylation level conservation categories. B: Distribution of expression fold changes of *Arabidopsis* paralogs in the three methylation pattern conservation categories. C: Distribution of expression fold changes of rice paralogs in the three methylation level conservation categories. D: Distribution of expression fold changes of rice paralogs in the three methylation pattern conservation categories. E: Distribution of expression specificity changes of *Arabidopsis* paralogs in the three methylation level conservation categories. F: Distribution of expression specificity changes of *Arabidopsis* paralogs in the three methylation pattern conservation categories. CHM: conserved highly methylated paralogs. CLM: conserved lowly methylated paralogs. CNM: conserved non-methylated paralogs. NCM: non-conserved methylated paralogs. Significant levels (*) are assigned to CLM, CNM, or NCM if their expression changes are significantly higher than those of CHM by Wilcoxon rank sum test. Significant levels were shown as ‘*’*: p<0.05, ‘**’*: p<0.01, and ‘***’*: p<0.001.

doi:10.1371/journal.pone.0110357.g006
The effect of post-transcription regulation. To control for the effect of post-transcriptional regulation (e.g., regulation of the small RNAs), we mapped small RNA sequences to Arabidopsis and rice genomes and estimated the numbers of the small RNAs mapped to each duplicate gene normalized by gene length. Because small RNAs with different length may regulate gene expression differently [86]. To simplify and standardize the experiment, we focused on 21nt and 24nt small RNAs, which are the most abundant small RNA species in Arabidopsis and rice genomes.

Table 3. Comparison of expression divergences (p value) in three gene body methylation conservation categories after controlling promoter methylation divergence.

<table>
<thead>
<tr>
<th>Species</th>
<th>Expression level divergence</th>
<th>Expression specificity divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CHM&lt;CLM</td>
<td>CHM&lt;NCM</td>
</tr>
<tr>
<td></td>
<td>3.432e-06</td>
<td>7.388e-07</td>
</tr>
<tr>
<td>Methylolation level</td>
<td>4.929e-04</td>
<td>0.02239</td>
</tr>
<tr>
<td>Methylolation pattern</td>
<td>1.831e-05</td>
<td>1.458e-05</td>
</tr>
<tr>
<td>Rice</td>
<td>CHM&lt;CLM</td>
<td>CHM&lt;NCM</td>
</tr>
</tbody>
</table>

Note: p value was computed with Wilcoxon rank sum test. doi:10.1371/journal.pone.0110357.t003
Exonic regions contribute more to gene body methylation divergence than intrinsic regions

To estimate whether the methylation level divergence in exonic and intrinsic regions contribute equally to the gene body methylation divergence, we divided the paralogs into the following categories using previous approaches: 1) conserved methylation level (CM) in exonic regions; 2) non-conserved methylation level (NCM) in exonic regions; 3) CM in intrinsic regions; 4) NCM in intrinsic regions. With Fisher test, we found that the proportion of exonic NCM among exonic NCM+CM was significantly higher than that of intrinsic regions in Arabidopsis and rice (Table 5, Fisher test, Arabidopsis: $p<2.2e-16$; rice: $p=2.212e-10$), suggesting exonic methylation divergence may contribute more to gene body methylation divergence than intrinsic methylation. Further, we found exonic methylation divergence is more closely correlated with expression divergence than intrinsic methylation divergence (Table S11), suggesting the exonic methylation divergence may play a more important role in the association with gene expression compared to intrinsic methylation divergence, for example, through regulating the alternative-splicing of transcripts.

Dispersed duplicate genes tend to have more divergent DNA methylation level/pattern

At the end, we explored the potential mechanisms contributing to the differential methylation between paralogs. Previous studies showed that DNA-based dispersed duplications in Arabidopsis tend to differ in DNA methylation in promoter regions than other gene duplication mechanisms [89]. In rice, different duplication mechanisms generate paralogs with different physical distance; which are associated with different body methylation level conservation [51]. Our analyses indicated that DNA-based dispersed gene duplication (Fisher test: methylation level: $p=2.699e-05$; pattern: $p=0.0074$) and RNA-based retrotransposition (Fisher test: methylation level: $p=0.06467$) generated a higher proportion of paralogs with divergent methylation level and/or pattern than tandem and segmental duplication in rice. We also found the methylation level (Wilcoxon rank sum test, $p=3.248e-06$) and pattern (Wilcoxon rank sum test, $p=1.689e-9$) of segmental duplication is more conserved than those of the dispersed genes generated by single events, such as tandem duplication, dispersed duplication in rice. However, we did not observe such pattern in Arabidopsis, which might be due to the intrinsic methylation property of Arabidopsis genome. And the underlying mechanism remains further to be explored.

Previous studies have demonstrated that various mechanisms contributed to the methylation variation in trans-generational epialleles, and indicated that DNA methylation could spread from highly methylated transposable elements (TEs) and repetitive sequences to the nearby genes [3,4,77,85,90,91]. Given our discovery that dispersed duplications generate a higher proportion of divergently methylated paralogs in rice, we asked whether the dispersed paralogs are more frequently associated with asymmetric TE/repeat sequence distribution, namely one duplicate copy flanked with TE/repeat sequence but not the other in rice. This could lead to the one flanked by TE/repeat sequence being more likely to be methylated, while the other without flanking TE/repeat sequence was not. Interestingly, we found that DNA-level dispersed duplication and retrotransposition paralogs were more frequently associated with asymmetric TE/repeat sequence distribution than tandem and segmental duplication ones in rice (Fisher test, $p<0.05$). The same trend was observed when we used different cutoff length of paralog flanking regions to search for the nearby TEs (Table S12).
Discussion and Conclusions

In human, an exclusive relationship of gene body DNA methylation with evolutionary and expression divergence of paralogs could not be revealed [48]. Our analysis in plants added a different aspect regarding with the property and significant role of gene body DNA methylation in gene duplication. Gene body methylation is a pervasive phenomenon in various species and their pattern is highly conserved in orthologs from highly diverged species [23,28,34,79,92,93]. Furthermore, as demonstrated previously, gene body DNA methylation in plant genomes has been considered as the main mode of DNA methylation and links with gene transcription [34]. Expectedly, the divergence of gene body methylation could play a functional role in influencing the evolution and divergence of paralogs given the interplay between gene transcription and DNA methylation [26,29,30,94]. Specifically, paralogs initially tended to maintain same expression pattern and retained similar function; and gene body DNA methylation could regulate gene expression. Therefore, the divergence of gene body DNA methylation could provide a fast trajectory to differentiate gene expression of paralogs, leading to the divergence and thus the preservation of duplicate genes [52].

To validate the association between gene body DNA methylation and divergence of plant duplicate genes and the contribution of methylation to the divergence of plant duplicate genes, we analyzed a large number of reliable paralogs spanning different evolutionary ages from two distantly related plant species, Arabidopsis and Oryza, which represent dicots and monocots, respectively. We first showed that divergence of methylation level was positively correlated to the genetic evolution of paralogs in terms of Ks and Ka. This result suggested that methylation level variation accumulates along evolutionary time, and companies with neutral DNA mutation and amino acid changes. These postulations are supported by the previous discoveries that CG-SMP (single methylation polymorphism)-based phylogenetic tree is correlated to the SNP-based phylogenetic tree among multiple populations of Arabidopsis or closely related rice species [27,77].

Differential gene body DNA methylation covaries with gene expression divergence between duplicate genes. It has been hypothesized that epigenetic modification of paralogs could facilitate functional divergence of duplicate genes [52]. Consistently, we demonstrated that conserved highly methylated duplicate genes have lower expression level and specificity divergence than conserved lowly- or non-conserved methylated duplicate genes. Remarkably, this pattern still robustly exists even after we controlled other factors known to affect gene expression divergence. Our observation is reflecting with previous transgenerational epigenetic variation studies, which indicated that epigenetic variation could contribute to phenotypic diversity [85,90] and that de novo genes could originate through change gene transcription [34]. Expectedly, the divergence of gene body methylation could play a functional role in influencing the evolution and divergence of paralogs given the interplay between gene transcription and DNA methylation [26,29,30,94]. Specifically, paralogs initially tended to maintain same expression pattern and retained similar function; and gene body DNA methylation could regulate gene expression. Therefore, the divergence of gene body DNA methylation could provide a fast trajectory to differentiate gene expression of paralogs, leading to the divergence and thus the preservation of duplicate genes [52].

To validate the association between gene body DNA methylation and divergence of plant duplicate genes and the contribution of methylation to the divergence of plant duplicate genes, we analyzed a large number of reliable paralogs spanning different evolutionary ages from two distantly related plant species, Arabidopsis and Oryza, which represent dicots and monocots, respectively. We first showed that divergence of methylation level was positively correlated to the genetic evolution of paralogs in terms of Ks and Ka. This result suggested that methylation level variation accumulates along evolutionary time, and companies with neutral DNA mutation and amino acid changes. These postulations are supported by the previous discoveries that CG-SMP (single methylation polymorphism)-based phylogenetic tree is correlated to the SNP-based phylogenetic tree among multiple populations of Arabidopsis or closely related rice species [27,77].

Differential gene body DNA methylation covaries with gene expression divergence between duplicate genes. It has been hypothesized that epigenetic modification of paralogs could facilitate functional divergence of duplicate genes [52]. Consistently, we demonstrated that conserved highly methylated duplicate genes have lower expression level and specificity divergence than conserved lowly- or non-conserved methylated duplicate genes. Remarkably, this pattern still robustly exists even after we controlled other factors known to affect gene expression divergence. Our observation is reflecting with previous transgenerational epigenetic variation studies, which indicated that epigenetic variation could contribute to phenotypic diversity [85,90] and that de novo genes could originate through change gene transcription [34]. Expectedly, the divergence of gene body methylation could play a functional role in influencing the evolution and divergence of paralogs given the interplay between gene transcription and DNA methylation [26,29,30,94]. Specifically, paralogs initially tended to maintain same expression pattern and retained similar function; and gene body DNA methylation could regulate gene expression. Therefore, the divergence of gene body DNA methylation could provide a fast trajectory to differentiate gene expression of paralogs, leading to the divergence and thus the preservation of duplicate genes [52].

To validate the association between gene body DNA methylation and divergence of plant duplicate genes and the contribution of methylation to the divergence of plant duplicate genes, we analyzed a large number of reliable paralogs spanning different evolutionary ages from two distantly related plant species, Arabidopsis and Oryza, which represent dicots and monocots, respectively. We first showed that divergence of methylation level was positively correlated to the genetic evolution of paralogs in terms of Ks and Ka. This result suggested that methylation level variation accumulates along evolutionary time, and companies with neutral DNA mutation and amino acid changes. These postulations are supported by the previous discoveries that CG-SMP (single methylation polymorphism)-based phylogenetic tree is correlated to the SNP-based phylogenetic tree among multiple populations of Arabidopsis or closely related rice species [27,77].

Differential gene body DNA methylation covaries with gene expression divergence between duplicate genes. It has been hypothesized that epigenetic modification of paralogs could facilitate functional divergence of duplicate genes [52]. Consistently, we demonstrated that conserved highly methylated duplicate genes have lower expression level and specificity divergence than conserved lowly- or non-conserved methylated duplicate genes. Remarkably, this pattern still robustly exists even after we controlled other factors known to affect gene expression divergence. Our observation is reflecting with previous transgenerational epigenetic variation studies, which indicated that epigenetic variation could contribute to phenotypic diversity [85,90] and that de novo genes could originate through change gene transcription [34].
the DNA methylation status (Silveira, et al. 2013). However, two additional considerations should be explored further. First, although we quantitatively controlled for other factors, we could not control for all the factors completely. For example, the same amount of nucleotide substitutions occurring in different gene regions might also contribute to expression divergence of paralogs, and the data available for cis- and post-transcription regulation are limited. Also, gene body methylation might interplay with other mechanisms together and contribute to the divergence of duplicate genes.

Methylation level measures the overall methylation density of genes. Whereas methylation pattern measures methylation property of specific cytosine sites of genes. Therefore, methylation level and pattern divergence could behave differentially in affecting the expression divergence. We identified the substantial roles of DNA methylation level for expression and genetic divergence of *Arabidopsis* duplicate genes. This result supports the hypothesis that gene body methylation might be maintained as a property of regions with a minimal/maximal methylation level over the whole region while experiencing site to site methylation stochasticity, driven by natural selection [34]. Alternatively, methylation changes in specific sites could also affect gene expression. For example, it was previously found that enhancer access could be modulated by a single methylation event [24].

The methylation divergence of duplicate genes may be the outcome of spontaneous methylation variation, which might be impacted by natural selection [see reviews by [95–97]]. A few factors could contribute to the methylation divergence of paralogs. For example, methylation divergence from exons and introns could differently contribute to the divergence of gene body methylation in duplicate genes. We found that more methylation divergence occurs in the exonic regions than intronic regions. Previously, two studies reported conflicting results about the involvement of exons and introns in differentially methylated regions (DMR) among different *Arabidopsis* lines due to spontaneous epigenetic variation. Becker et al. (2011) showed that CG-DMRs in exons were more abundant than those in introns [85], while Schmitz et al. (2011) found the opposite pattern [90]. Previous studies also showed that methylation in exons may be associated with mRNA-splicing, thus exonic methylation divergence might be relevant to the mRNA-splicing divergence and expression divergence of paralogs [29,98].

We demonstrated that exonic methylation divergence is more closely correlated with expression divergence than intronic methylation divergence, which implies that the exonic methylation divergence might have more functional consequences than intronic methylation divergence. Conclusively, exonic methylation variation might have a higher chance to be maintained than intronic methylation variation, which can explain our observation that exonic regions bear more methylation divergence than intronic regions. However, it remains to be explored other functional roles of methylation in exonic and intronic regions. Furthermore, the genomic environments of duplicate genes can also impact the divergence of methylation among paralogs. Our analysis confirmed that the physical distance of paralogs negatively correlated with the conservation of methylation level and pattern, and that dispersed duplicate genes more tend to have divergent methylation level and pattern.

We found inconsistent patterns, for some analyses, in *Arabidopsis* and rice. This may be due to two factors. First, the limitation of availability and the quality of the data in two genomes are different. For example, the expression data of *Arabidopsis* were obtained from RNA-seq but those of rice were generated from DGE. Second, properties of methylation in the two genomes are
different. For example, the methylation level in rice is at least two-fold higher than that in Arabidopsis.

In conclusion, the correlation between the divergence of gene body DNA methylation and expression in plant duplicate genes implicates that gene body DNA methylation could serve as another avenue for duplicate genes to develop different expression and undergo different evolutionary fates. Our study indicates that gene body DNA methylation as one type of epigenetic modifications is an important facilitator that potentially drove divergence and evolution of plant duplicate genes.

Supporting Information

Figure S1  Plots of methylation level vs. Ka and Ks for duplicate genes in Arabidopsis and rice. Red line is the linear regression fit to the data, which were generated with lm function in R.

(TIF)

Figure S2  Plots of the methylation level vs. the expression level for duplicate genes in Arabidopsis and rice. Red line is the linear regression fit to the data, which were generated with lm function in R.

(TIF)

Table S1  Reprocessed error rate in un-methylated chloroplast genome and methylation data for Arabidopsis and rice.

(PDF)

Table S2  The correlation between methylation level and gene length/exon number.

(PDF)

Table S3  The correlation between methylation level and expression level/specificity in Arabidopsis.

(PDF)

Table S4  The correlation between methylation level and expression divergence with small RNA abundance rate controlled.

(PDF)

Table S5  The correlation between methylation level and theta in Arabidopsis.

(PDF)

Table S6  The correlation between theta and other factors.

(PDF)

Table S7  The relationship of methylation conservation and expression divergence that is corresponded to Figure 6.

(PDF)

Table S8  The relationship of methylation conservation and expression divergence with gene structure controlled.

(PDF)

Table S9  The relationship of methylation conservation and expression divergence with nucleotide substitution rate controlled.

(PDF)

Table S10  The relationship of methylation conservation and expression divergence with small RNA abundance controlled.

(PDF)

Table S11  The correlation between exonic/intronic methylation conservation and expression level divergence.

(PDF)

Table S12  The p-values corresponding to the different flanking region length cutoffs in searching nearby TEs.

(PDF)

Acknowledgments

We thank Dr. Xin Li from Johns Hopkins University School of Medicine, Mr. Mulhua Wang from University of Arizona, and Dr. Chengjun Zhang from University of Chicago for helpful discussions. We thank the grid computing service from Computing & information Technology of Wayne State University.

Author Contributions

Conceived and designed the experiments: JW CF. Performed the experiments: JW NCM. Analyzed the data: JW CF. Contributed reagents/materials/analysis tools: JW CF. Contributed to the writing of the manuscript: JW NCM CF.

References


