Title

CG gene body DNA methylation changes and evolution of duplicated genes in cassava

Permalink
https://escholarship.org/uc/item/1012p8j1

Journal
Proceedings of the National Academy of Sciences of the United States of America, 112(44)

ISSN
0027-8424

Authors
Wang, Haifeng
Beyene, Getu
Zhai, Jixian
et al.

Publication Date
2015-11-03

DOI
10.1073/pnas.1519067112

Peer reviewed

CG gene body DNA methylation changes and evolution of duplicated genes in cassava

Haifeng Wang[a,b], Getu Beyene[c], Jixian Zhai[b], Suhua Feng[b,d], Noah Fahlgren[c], Nigel J. Taylor[c], Rebecca Bart[c], James C. Carrington[c], Steven E. Jacobsen[b,d,e,1], and Israel Ausin[a,1]

[a]Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China; [b]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA 90095; [c]Donald Danforth Plant Science Center, St. Louis, MO 63132; [d]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, CA 90095; and [e]Howard Hughes Medical Institute, University of California, Los Angeles, CA 90095

Contributed by Steven E. Jacobsen, September 25, 2015 (sent for review June 27, 2015; reviewed by Rebecca A. Mosher and Steven D. Rounsley)

DNA methylation is important for the regulation of gene expression and the silencing of transposons in plants. Here we present genome-wide methylation patterns at single-base pair resolution for cassava (*Manihot esculenta,* cultivar TME 7), a crop with a substantial impact in the agriculture of subtropical and tropical regions. On average, DNA methylation levels were higher in all three DNA sequence contexts (CG, CHG, and CHH, where H equals A, T, or C) than those of the most well-studied model plant *Arabidopsis thaliana*. As in other plants, DNA methylation was found both on transposons and in the transcribed regions (bodies) of many genes. Consistent with these patterns, at least one cassava gene copy of all of the known components of *Arabidopsis* DNA methylation pathways was identified. Methylation of LTR transposons (*GYPSY* and *COPIA*) was found to be unusually high compared with other types of transposons, suggesting that the control of the activity of these two types of transposons may be especially important. Analysis of duplicated gene pairs resulting from whole-genome duplication showed that gene body DNA methylation and gene expression levels have coevolved over short evolutionary time scales, reinforcing the positive relationship between gene body methylation and high levels of gene expression. Duplicated genes with the most divergent gene body methylation and expression patterns were found to have distinct biological functions and may have been under natural or human selection for cassava traits.

cassava | DNA methylation | duplicate genes | gene expression

DNA methylation plays an important role in the regulation of the expression of genes and the maintenance of transposable element (TE) silencing. In contrast to animals, in which methylation is often restricted to the CG context, plants exhibit robust methylation in every possible context CG, CHG (H is A, T, or C), and CHH. Previous research has identified different pathways responsible for the maintenance and establishment of DNA methylation patterns. In *Arabidopsis thaliana*, METHYLTRANSFERASE1 (MET1), a homolog of mammalian Dnmt1, mainly maintains methylation at the CG context, whereas CHROMOMETHYLASE3 (CMT3) mainly maintains CHG methylation. DOMAINS REARRANGED METHYLTRANSFERASE2 (DRM2) and CHROMOMETHYLASE2 (CMT2) maintain CHH methylation in the chromosome arms and pericentromeric regions, respectively (1–3). On the other hand, establishment of DNA methylation is performed by DRM2 through a complex pathway termed RNA-directed DNA methylation (RdDM) (4).

To date, the majority of our knowledge about DNA methylation is derived from the model plant *Arabidopsis*. These studies have allowed the identification of different components involved in different methylation pathways, the genome-wide identification of methylation patterns, and the study of effects of DNA methylation on gene expression. The knowledge acquired from *Arabidopsis* can now be used as the basis for investigations of methylation in agronomically important plants. However, thus far very few crop species have been subjected to detailed DNA methylation studies (5). Cassava (*Manihot esculenta*) is cultivated for its starch-rich tuberous roots and is one of the world's most important staple crops, especially in tropical America, Africa, and Asia (6). Cassava is a source of carbohydrates for nearly a billion people, but it is especially important for a large portion of Africa, where it serves as a subsistence crop because of its ability to tolerate drought and grow on poor soils, conditions unsuitable for rice and maize (6, 7). The genome sequence of cassava has been described recently with an estimated genome size of roughly 760 million base pairs (7). We have used bisulfite sequencing (BS-seq) to examine DNA methylation in cassava at single-base pair resolution. Broadly, the pattern of DNA methylation of both protein-coding genes and TEs is similar to other plants, although DNA methylation levels in cassava are higher than those in *Arabidopsis*. LTR retrotransposons, such as *GYPSY* and *COPIA*, tend to be more heavily methylated than other TEs. Interestingly, differentially expressed gene pairs derived from the last genome duplication tend to show differential gene body methylation, with the highly expressed paralogs displaying significantly higher gene body methylation. We also find that the most differentially gene body-methylated paralogs have distinct biological functions compared with genes that have maintained similar gene body methylation patterns.

Results and Discussion

Genes Involved in Different DNA Methylation Pathways Are Conserved in Cassava. Detailed genetic studies in *Arabidopsis* have defined the key components involved in DNA methylation pathways controlled by the MET1, CMT3, CMT2, and DRM2 methyltransferases (3, 4). As a preliminary assessment of the functioning of these pathways in

Significance

Plant traits exhibit variation as a result of genetic and epigenetic change. Genetic variation is used for breeding and crop improvement. Epigenetic variation, especially differences in DNA methylation, also contributes to phenotype. For example, epigenetic alleles of plant genes exist in nature, which are identical in DNA sequence, but show heritable differences in DNA methylation and gene expression. Here we present whole-genome DNA methylation patterns of the agronomically important crop cassava (*Manihot esculenta*), which can serve as the basis for the study of epigenetic variation in this organism. We found that recently duplicated genes have evolved different DNA methylation and expression patterns that likely contribute to important agronomic traits.

Author contributions: H.W., N.J.T., R.B., J.C.C., S.E.J., and I.A. designed research; H.W., G.B., J.Z., S.F., N.F., and I.A. performed research; H.W. analyzed data; and H.W., S.E.J., and I.A. wrote the paper.

Reviewers: R.A.M., University of Arizona; and S.D.R., University of Arizona.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequencing data have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE73645).

[1]To whom correspondence may be addressed. Email: jacobsen@ucla.edu or israel.ausin@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1519067112/-/DCSupplemental.

cassava, we searched the cassava genome for homologs of each of the *Arabidopsis* genes. We found that the cassava genome contains at least one copy of every key factor involved in DNA methylation control (Table 1), suggesting that all canonical DNA methylation pathways are functional and conserved in cassava.

DNA Methylation Patterns in Cassava. To study genome-wide DNA methylation patterns in cassava at single-base resolution, we used whole-genome BS-seq. BS-seq libraries were constructed from genomic DNA extracted from leaves of the TME 7 cultivar of cassava and subjected to deep Illumina sequencing. To assess variability, three biological replicates were generated. Reads generated from each library were mapped independently to the most recent version (6.1) of the cassava genome. Mapping was performed using BSMAP (7, 8), such that 68.6%, 69.7%, and 69.6% of total reads could be uniquely mapped for each replicate library (*SI Appendix*, Table S1). To test the reproducibility of our results, we calculated Pearson correlation coefficients between these three replicates, and found the correlations to be ∼0.87–0.89 (*SI Appendix*, Table S2), indicating a high reproducibility within our libraries. The total coverage of the cassava genome for these libraries was 63-fold (*SI Appendix*, Table S1). Approximately 82% of the cytosines were covered by at least four reads (*SI Appendix*, Fig. S1) and more than 70% of genome was covered by at least 30 reads (*SI Appendix*, Fig. S2). DNA methylation browser tracks are available at phytozome.jgi.doe.gov/jbrowse/index.html?data=genomes%2FMesculenta_er.

Global DNA methylation profiles of chromosome 1 to chromosome 5 are shown in Fig. 1*A* .The remaining 13 chromosomes are shown in *SI Appendix*, Fig. S3. As expected, we found TE populations to be especially dense in what are likely pericentromeric regions and to be heavily methylated, whereas chromosome arms were gene-rich and showed lower methylation levels. The average percentages of methylation of CG, CHG, and CHH contexts were 58.7%, 39.5%, and 3.5%, respectively, much higher than those in *Arabidopsis* (24%, 6.7%, and 1.7% for CG, CHG, and CHH, respectively) (Fig. 1*B*) (9). By comparing two other crop species with reported deep methylation data, we found that methylation levels in cassava were higher than those in rice, but lower than those reported for soybean (Fig. 1*B*) (10, 11). Interestingly, in contrast to other plant species analyzed, in which CG methylation is the

Table 1. DNA methylation related genes in cassava

		Cassava (*Manihot esculenta*)		
Gene function	Name (*Arabidopsis*)	Amino acid length	Copy 1	Copy 2
MET1	*VIM1, -2, -3, -4, -5, -6*	645	Manes.14G168600	Manes.08G101100
	MET1, -2a, -2b, -3	1,534	Manes.13G155300	Manes.13G119400
CMT3	*SUVH4*	624	Manes.06G009100	
	CMT2	1,295	Manes.09G037800	
	CMT3	839	Manes.03G089100	
Pol IV recruit	*CLSY1/CLSY2*	1,256	Manes.10G00780	
	SHH1/SHH2	258	Manes.04G133600	
Pol IV	*NRPD1*	1,453	Manes.02G028200	
Pol IV+V	*NRPD2/NRPE2*	1,172	Manes.16G129400	Manes.03G009000
Pol IV+V	*NRPD4/NRPE4*	205	Manes.09G085000	
Pol V	*NRPE1*	1,976	Manes.04G159600	
Pol V	*NRPE5*	222	Manes.09G007600	
Pol V	*NRPE9B*	114	Manes.15G005400	
Pol V recruit	*DRD1*	888	Manes.04G086500	
	DMS3	420	Manes.10G072000	Manes.17G027400
	RDM1	163	Manes.15G031200	
	SUVH2/9	650	Manes.03G082600	Manes.15G046600
RdDM	*RDR2*	1,133	Manes.14G068000	
	DCL1	1,910	Manes.05G015200	
	DCL2	1,388	Manes.12G003000	Manes.12G002800
	DCL3	1,580	Manes.03G056500	
	DCL4	1,702	Manes.14G140300	
	HEN1	942	Manes.06G068000	
	AGO4	924	Manes.02G209900	Manes.18G121900
	KTF1	1,493	Manes.07G094600	
	IDN2	647	Manes.07G117100	
	IDL1/2	634	Manes.04G103800	
	SUVR2	740	Manes.12G036100	
	DMS4	346	Manes.12G056300	
	UBP26	1,067	Manes.18G079200	
	DRM2	626	Manes.17G113600	
	DRM3	710	Manes.03G210200	
	LDL1	844	Manes.11G098200	
	LDL2	746	Manes.03G115600	
	JMJ14	954	Manes.16G062600	
	HDA6	471	Manes.14G061800	
Others	*RDR6*	1,196	Manes.16G121400	
	MOM1	2,001	Manes.03G122500	
	MORC6	663	Manes.11G096200	
	DDM1	764	Manes.01G134600	Manes.02G092800

Amino acid length is for the longest protein.

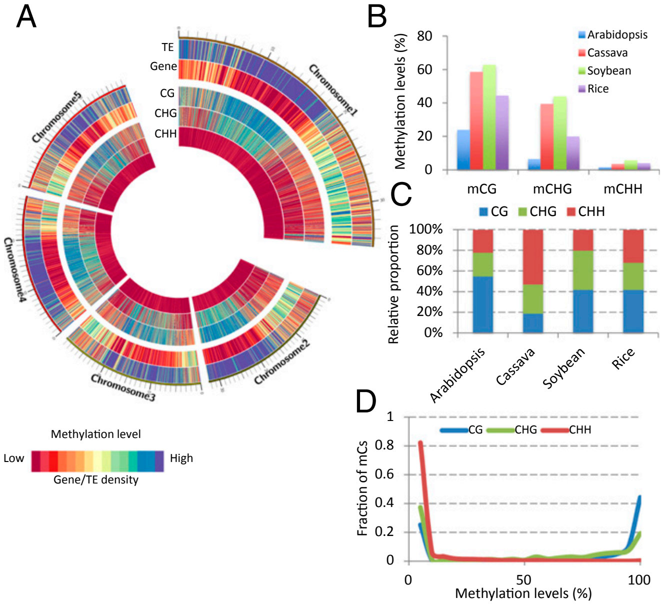

Fig. 1. Genome-wide DNA methylation profiles in cassava. (*A*) Circle plots of gene density, TE density, and methylation level of CG, CHG, CHH across five chromosomes of cassava. DNA methylation level is represented as a heatmap. Red color indicates low methylation level and low gene/TE density. Blue color indicates high methylation level and high gene/TE density. (*B*) Average methylation level of cassava in all three contexts. Data from *Arabidopsis*, soybean, and rice is also shown. (*C*) Relative proportion of mCs in all three sequence contexts. (*D*) The genome-wide distribution of methylation levels. Methylation levels were calculated by #C/(#C+#T) of individual cytosine, and each cytosine used in this analysis was covered by at least four reads. Methylation levels were divided into 5% bins, such that 100 indicates methylation level from 95% to 100%.

most abundant, cassava showed a very high proportion of CHH methylation relative to the other types (Fig. 1*C*). In *Arabidopsis*, CG sites show a bimodal distribution where sites tend to be either unmethylated or methylated at very high levels, approaching 100%, whereas CHG and CHH sites are rarely methylated at very high levels (9). This trend likely represents the different mechanisms by which these methylation types are maintained, where CG methylation is copied faithfully during the DNA replication process, whereas CHG and CHH methylation are perpetually targeted by histone methylation and noncoding RNAs (4). Interestingly, we found that cassava shows bimodal distribution patterns for both CG and CHG methylation, suggesting that CHG methylation is more robustly maintained in cassava than in *Arabidopsis* (Fig. 1*D*). For methylation of TEs, we observed that although there were a significant proportion of very short TEs with low levels of CG and CHG methylation, long TEs were almost always methylated at high levels (*SI Appendix*, Fig. S4). In summary, although there are general similarities between the methylation patterns of different plant species, cassava shows unique patterns, including a very high content of CHH methylation throughout the genome, and CHG methylation sites that are maintained at a very high level.

Methylation Patterns in Genic and TE Regions. Methylation patterns in protein-coding genes and TEs in cassava were characterized. CG methylation patterns in protein-coding genes are generally similar to those in *Arabidopsis*, rice, and soybean (9–13). Metaplot analysis of protein-coding genes showed that gene body methylation is almost exclusively in the CG context, and CG methylation levels are very low near transcriptional start sites (TSS) and transcriptional end sites (TES) (Fig. 2*A*). A small amount of non-CG methylation within protein coding genes was also found (Fig. 2 *B* and *C*). This is likely the result of a small portion of genes or pseudogenes possessing repeats or small TEs in their intronic sequences, because the levels of non-CG methylation were reduced when genes with intronic transposable elements are excluded (*SI Appendix*, Fig. S5).

For TE regions, high levels of methylation were seen in all three sequence contexts, consistent with previous studies in other plants (9–11). Interestingly, methylation of TEs was found to be, on average, higher than that in *Arabidopsis* for CG and CHG contexts (~90% vs. ~70% for CG and ~75% vs. ~40% for CHG) (Fig. 2 *D*–*F*) (9). This finding suggests CG and CHG methylation are more robustly maintained in cassava, perhaps because of the higher transposon load in the cassava genome. In addition, different types of TEs showed distinct levels of methylation. In particular, the *GYPSY* and *COPIA* LTR-type transposons displayed higher methylation levels compared with all other types of TEs in all three sequence contexts (Fig. 2 *G*–*I*), suggesting that methylation of LTR transposons could be especially important for repression of transposon activity. Consistent with this idea, a recent study showed that genome expansion of *Arabis alpina* was caused in part by the expansion of *GYPSY* retrotransposons, which could be a result of high transposition activity caused in turn by lower levels of DNA methylation of *GYPSY* retro-transposons (14). Repeats showed lower methylation levels than transposons (70%, 50%, and 5% methylation levels for CG, CHG, and CHH, respectively), which is consistent with results of *Arabidopsis* and other plant species (9, 10). Together, these data showed that methylation patterns in both protein-coding genes and TEs are generally consistent with those in other plant species (9, 10, 13, 15), but cassava shows a particularly high level of maintenance methylation at CG and CHG sites, especially in *GYPSY* and *COPIA* retro-transposons.

Gene Body Methylation Is Associated with Gene Activity. Nongenic methylation is usually associated with transcriptional repression at repetitive elements and transposons, and silencing can also be observed when methylation is present at gene promoters. Conversely, gene body methylation generally correlates with transcriptionally active genes (1, 16, 17). To assess the correlation between DNA methylation and gene expression, RNA levels were profiled by high-throughput RNA-sequencing (RNA-seq). In total, ~95 million raw reads were generated by paired-end 100-bp sequencing, with ~81 million reads uniquely mapping to the reference cassava genome (*SI Appendix*, Table S3). Correlations between the three biological replicates were very high (*SI Appendix*,

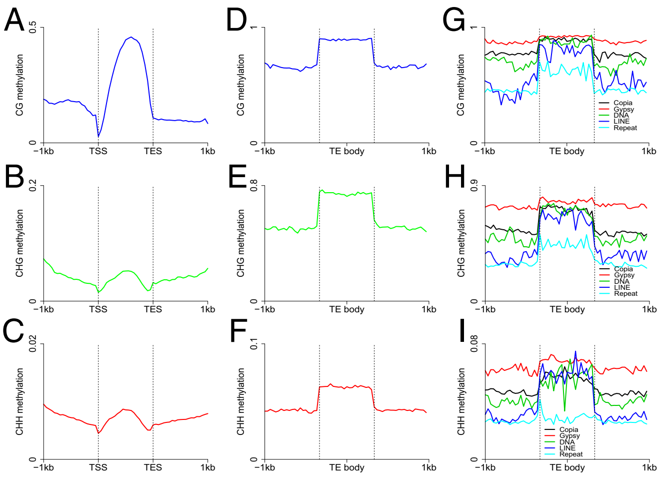

Fig. 2. DNA methylation patterns across genes (*A*–*C*) and TEs (*D*–*F*). (*G*–*I*) Metaplots representing DNA methylation patterns within different types of TEs, such as Copia, Gypsy, DNA-type, LINE, and simple repeats. In all cases, −1 kb indicates the upstream 1,000 bp of TSS, and 1 kb indicates the downstream 1,000 bp of TES. Upstream, gene body/TE, and downstream were divided into 20 proportionally sized bins.

Table S4). Reads were mapped to 23,297 of the 33,033 annotated protein-coding genes.

Genes were divided into four quartiles based on expression levels, from the first quartile (the most lowly expressed 25% of genes) to the fourth quartile (the most highly expressed 25% of genes). A positive correlation was observed between gene body CG methylation and gene-expression levels (Fig. 3*A*). Moreover, consistent with what has been found in other organisms (1, 11, 17), the highest methylation levels were not detected in the most highly expressed genes, but instead in those that are moderately highly expressed (the third quartile). For non-CG methylation, genes with different expression levels showed comparable low levels of methylation (Fig. 3 *B* and *C*). Furthermore, there were also very low levels of non-CG methylation present across gene bodies and flanking regions of the genes in all expression groups. A Spearman correlation coefficient was calculated between DNA methylation and expression levels across gene bodies and flanking regions in different sequence contexts, which confirmed that CG gene body methylation is positively correlated with expression, whereas CG methylation in flanking regions is negatively correlated with expression (Fig. 3*D*).

In summary gene body methylation shows a generally positive correlation with expression, whereas methylation upstream and downstream of the transcription unit is generally correlated with lower gene-expression levels.

DNA Methylation Variation Between Duplicated Genes. Virtually all angiosperms have undergone polyploidization (or whole-genome duplication, WGD). After WGD most duplicated genes are lost, but some may be retained by selective pressure (12). To explore the relationship between DNA methylation and gene expression of duplicated genes, an analysis of recently duplicated genes in the cassava genome was performed.

It was reported that a relatively recent WGD likely occurred in cassava (7). It is known that synonymous divergence levels (K_s) of duplicated paralogs can be used as a proxy to calculate the age of duplications (18–20). The K_s values of each duplicated gene pair were calculated, and duplicated genes likely resulting from the most recent WGD were identified. Fig. 4*A* shows that there is a significant peak of K_s values at around 0.4. The likely explanation for why so many similarly aged paralogs are found is that a relatively recent WGD occurred at around 10–13.3 million y ago (K_s from 0.3 to 0.4 based substitution rate 1.5×10^{-8}) (21) (*SI Appendix*, Fig. S6), after the divergence of cassava and poplar. Although this is a relatively recent WGD, it clearly precedes the domestication of cassava that occurred no more than 10,000 y ago (22).

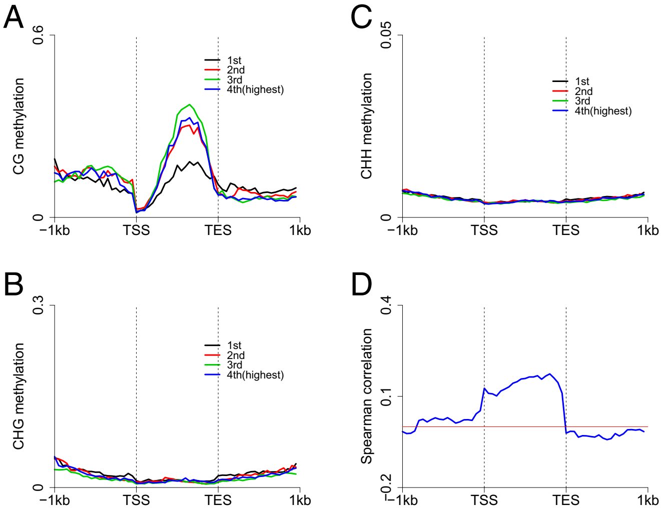

Fig. 3. Association between DNA methylation and expression. (*A*–*C*) Association between methylation and expression in CG, CHG, and CHH contexts. (*D*) Spearman correlation coefficient between CG methylation and expression across gene body and flanking regions. In all cases, −1 kb indicates the upstream 1,000 bp of TSS, and 1 kb indicates the downstream 1,000 bp of TES. Upstream, gene body/TE, and downstream were divided into 20 proportionally sized bins. Genes were divided into four groups of increasing expression levels, from first (lowest expression) to fourth (highest expression).

We extracted this set of duplicated paralogs and rank-ordered gene pairs according to the level of gene body methylation divergence between the pairs. We then plotted RNA expression levels to generate a heatmap (*SI Appendix*, Fig. S7). We found that for CG methylation, the biggest change in gene expression between the gene pairs was clearly present in the set of genes with the biggest differences in gene body methylation between the pairs (*SI Appendix*, Fig. S7). Conversely, we also classified duplicate gene pairs into either differentially or nondifferentially expressed pairs. A differentially expressed pair was defined by at least a twofold difference in expression levels. CG gene body methylation was found to be significantly higher for genes in the high-expression group compared with the low group (P value < 0.01; Wilcoxon rank sum test), whereas CHG and CHH body methylation did not show significant differences between these two groups across the gene body or flanking regions (*SI Appendix*, Fig. S8). The difference in CG gene body methylation became even more prominent when the fold expression change between paralogs was increased to fourfold (Fig. 4 *B*–*D*). We also performed an analysis of DNA methylation patterns of each gene pair within all three sequence contexts, rank ordered by expression fold-change. Fig. 4*E* shows that the higher the expression fold-change between paralogs, the greater the difference in CG methylation. However, this was not the case for non-CG methylation. Taken together, these analyses indicate that within duplicated genes, there is a strong positive correlation between the level of CG gene body methylation and levels of gene expression, suggesting that CG gene body methylation changes have evolved along with expression level changes on the time scale of the latest genome duplication in cassava.

To investigate whether gene pairs with more divergent expression levels and gene body methylation belong to specific gene classes, duplicated pairs were divided into three groups based on the expression fold-change. The first group consisted of duplicated genes with at least fourfold change of expression between duplicates, the second consisted of duplicated genes with at least a twofold difference in expression, and the third group were those duplicated genes with less than a twofold change between duplicates (Fig. 4*E*). Functional categories were examined among these three groups by using Gene Ontology (GO) term enrichment analyses. Intriguingly, within the first group of genes consisting of paralogous gene pairs in which only one gene copy is predominantly expressed and heavily body methylated, the most significant GO terms were found to consist of functional categories involved in carbohydrate metabolism. These included hexose metabolic process, glucose metabolic process, monosaccharide metabolic process, and others (*SI Appendix*, Fig. S9 and Table S5). The second and third group of genes in which the gene pairs showed more similar gene expression and gene body methylation showed enrichments in other categories, but were not as enriched in carbohydrate metabolism (*SI Appendix*, Tables S6 and S7). It is intriguing that the most differentially expressed and differentially gene body-methylated genes are highly enriched for genes involved in carbohydrate metabolism, given that cassava has been strongly selected for storage root production as a source of carbohydrates. One possibility is that these duplicate genes may have been under greater selection, such that one gene copy evolved preferentially over the other. To test this idea, K_a/K_s values were calculated, which is the ratio of the number of nonsynonymous mutations to synonymous mutations for each gene pair. Interestingly, group 1 genes that showed the most divergence between expression and gene body methylation also showed the highest K_a/K_s ratios compared with the other two groups (*SI Appendix*, Fig. S10). These results suggest that these carbohydrate metabolism genes have been under either natural or human selection.

Conclusion

To our knowledge, this work provides the first high-resolution genome-wide DNA methylation maps of the cassava genome. Cassava is one of the most important food security crops in the world, and given the important role that DNA methylation plays in the control of gene expression, these data should serve as an important resource for the scientific and agronomic community.

Although the general trends of cassava methylation patterns are similar to other plant species, cassava was found to have particularly high proportion of CHH methylation throughout the genome. In addition, very high levels of CHG methylation were observed, suggesting that cassava likely has a more robust maintenance methylation mechanism for CHG sites than *Arabidopsis*, which has a lower transposon content. As in other plants, genes are enriched for CG methylation, whereas TEs are enriched for all types of methylation. Because cassava is vegetatively propagated and the cultivar used in this study has not passed through meiosis for decades, one speculation is that some of the unusual properties of the methylation pattern could be attributed to many generations of clonal propagation.

Examination of DNA methylation within the recently duplicated genes generated by the latest WGD shows that the more highly gene body-methylated gene of the pair also shows the higher level of gene expression, suggesting that gene methylation and expression coevolved in cassava over short evolutionary time scales. Intriguingly, gene pairs with the highest difference in DNA methylation and expression are highly enriched for carbohydrate metabolism, and show higher K_a/K_s values that could possibly have resulted from human selection for beneficial crop traits. Alternatively, because the last WGD (10–13.3 million y ago) occurred long before human agriculture and cassava domestication (10,000 y ago), it is also possible that natural selection for carbohydrate storage and tuber development had a stronger influence on the observed enrichment for carbohydrate metabolism genes divergence in DNA methylation and expression than did human selection.

Methods

Library Construction and Sequencing. BS-seq libraries were prepared using the TruSeq DNA LT kit (Illumina), as described previously (23), except that the EZ DNA Methylation-Lightning Kit (Qiagen) was used for bisulfite conversion of

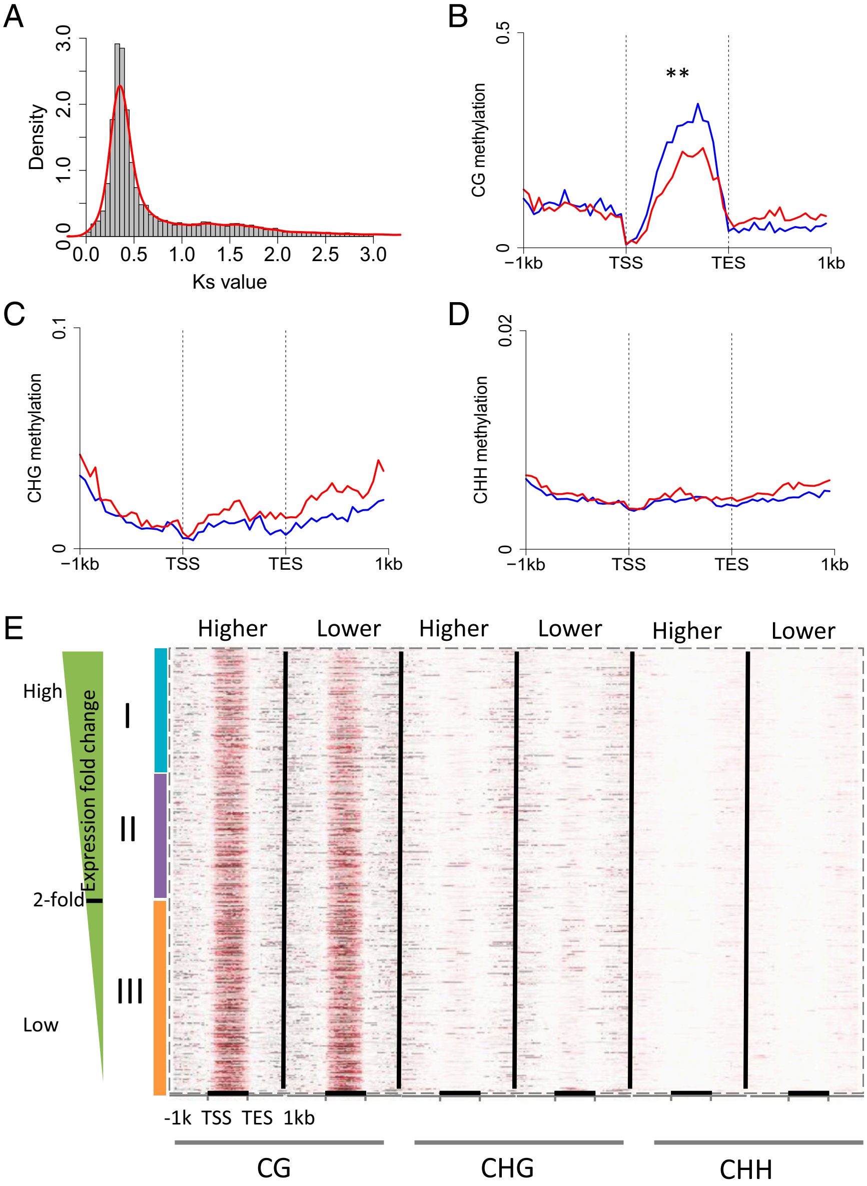

Fig. 4. Evolutionary analysis of DNA methylation of duplicated genes. (*A*) Graph showing K_s distribution. (*B–D*) Metaplots showing DNA methylation of protein-coding genes from high- and low-expression recent paralogous pairs. Higher-expression paralogs were selected by fourfold change of expression. (*E*) Heatmap of DNA methylation of each pair of genes. Recent paralogous pairs were ranked from high- to low-expression fold-change. In all cases −1 kb indicates the upstream 1,000 bp of TSS, and 1 kb indicates the downstream 1,000 bp of TES. Upstream, gene body, and downstream were divided into 20 proportionally sized bins. Methylation differences in *B* were tested by using the Wilcoxon rank sum test. $^{**}P < 0.001$.

the DNA. BS-Seq libraries were sequenced on a HiSeq 2000 system (Illumina) to obtain single-end 100-bp reads per the manufacturer's instructions.

Total RNA was extracted from the third or fourth fully expanded leaf of 7.5-wk-old TME 7 plants following the cetyltrimethylammonium bromide (CTAB) protocol (24). Genomic DNA was removed by TURBO DNA-free Kit (Ambion) and RNA quality and quantity were each assessed, respectively, by Agilent 2100 BioAnalyzer (Agilent Technologies) and NanoDrop 2000c (Thermo Scientific). One microgram of total RNA per sample was used for library preparation using the Illumina TruSeq sample preparation kit (v2) with polyA mRNA selection, as per the manufacturer's instructions (Illumina). Three libraries were pooled and sequenced using an Illumina HiSeq 2000 with paired-end reads of 101 bp at the Genome Technology Access Center of Washington University at St. Louis, MO.

BS-seq Data Analysis. Low-quality Illumina reads were filtered after which the remaining reads were aligned to cassava reference genome using BSMAP 2.87 (8). Only uniquely mapping reads were used to estimate methylation ratios. Methylation ratios were calculated as the number of Cs divided by Cs plus Ts (#C/#C+#T).

Reproducibility between replicates of BS-seq was calculated as methylation levels of total Cs in 2-kb regions. First, the reference genome was divided into 2-kb bins, and methylation levels were calculated as the average #C/(#C+#T) for all cytosines in each bin. We then we calculated Pearson correlation coefficients between replicates.

RNA-seq Data Analysis. We obtained a total of ~95 million paired-end 100-bp reads from three RNA-seq replicates. Total reads were aligned to the cassava reference genome using TopHat 2.0.11 using default parameters (25), then quantified using Cufflinks (26). Expression values were expressed in fragments per kilobase per million mapped reads (FPKM).

To estimate the correlation between replicates, we used the expression levels of individual genes estimated by FPKM. Genes with values under 0.5 FPKM were discarded and the remaining genes were used to calculate Pearson correlation coefficients.

GO Enrichment Analysis. GO enrichment analysis was performed using AgriGO online tools (bioinfo.cau.edu.cn/agriGO/analysis.php) with false-discovery rate correction (0.05).

Identification of Duplicated Genes and K_s Estimation. Duplicated genes were identified using MCScanX (27), an algorithm for detection of synteny and collinearity of genomes or subgenomes. Initially, the cassava proteome was subjected to search similarity using BLAST. A BLAST –m 8 output file was then provided as input to MCScanX. Simple linux "awk" command was used to extract those duplicated genes from collinearity regions from the MCScanX output file.

K_aK_s Calculator (v1.2) was used to calculate K_s values of individual gene pairs (28). Only duplicated pairs with less than 3 K_s value were used to plot the frequency distribution of K_s and to estimate large-scale gene duplication of cassava. K_s bin size was set at 0.05, and R scripts were used to draw the histogram and density plot.

Duplicated Genes Analysis. From MCScanX, 9,862 duplicated gene pairs were identified. Of these pairs, 4,169 showed twofold expression changes between members of a pair [excluding very lowly expressed genes (FPKM < 0.5 across three replicates)], and 2,333 showed at least a fourfold change in expression between pairs.

The difference of methylation levels across gene body and flanking regions between the higher-expression and lower-expression group was analyzed by the Wilcoxon rank test.

ACKNOWLEDGMENTS. We thank members of the S.E.J. laboratory for useful discussions. We also thank Raj Deepika Chauhan for production of plants used in these experiments. High-throughput sequencing was performed at Broad Stem Cell Research Center BioSequencing Core Facility of University of California, Los Angeles. This work was supported by a grant from the Bill & Melinda Gates Foundation. J.Z. is a Life Science Research Foundation postdoctoral fellow, sponsored by the Gordon and Betty Moore Foundation. S.E.J. is an investigator of the Howard Hughes Medical Institute.

1. Zemach A, et al. (2010) Local DNA hypomethylation activates genes in rice endosperm. *Proc Natl Acad Sci USA* 107(43):18729–18734.
2. Stroud H, et al. (2014) Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nat Struct Mol Biol* 21(1):64–72.
3. Matzke MA, Mosher RA (2014) RNA-directed DNA methylation: An epigenetic pathway of increasing complexity. *Nat Rev Genet* 15(6):394–408.
4. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11(3):204–220.
5. Ji L, Neumann DA, Schmitz RJ (2015) Crop epigenomics: Identifying, unlocking, and harnessing cryptic variation in crop genomes. *Mol Plant* 8(6):860–870.
6. FAO (2013) *Save and Grow: Cassava* (FAO, Rome).
7. Prochnik S, et al. (2012) The cassava genome: Current progress, future directions. *Trop Plant Biol* 5(1):88–94.
8. Xi Y, Li W (2009) BSMAP: Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* 10:232.
9. Cokus SJ, et al. (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452(7184):215–219.
10. Song QX, et al. (2013) Genome-wide analysis of DNA methylation in soybean. *Mol Plant* 6(6):1961–1974.
11. Li X, 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.
12. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.
13. Lister R, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133(3):523–536.
14. Willing E-M, et al. (2015) Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nature Plants* 1:14023.
15. Zhong S, et al. (2013) Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nat Biotechnol* 31(2):154–159.
16. Xiang H, et al. (2010) Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat Biotechnol* 28(5):516–520.
17. Zhang X, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* 126(6):1189–1201.
18. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16(7):1679–1691.
19. Cui L, et al. (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16(6):738–749.
20. Vanneste K, Baele G, Maere S, Van de Peer Y (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res* 24(8):1334–1347.
21. Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol Biol Evol* 17(10):1483–1498.
22. Olsen KM, Schaal BA (1999) Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. *Proc Natl Acad Sci USA* 96(10):5586–5591.
23. Du J, et al. (2014) Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. *Mol Cell* 55(3):495–504.
24. Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. *Focus* 12:13–15.
25. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–1111.
26. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–578.
27. Wang Y, et al. (2012) MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 40(7):e49.
28. Zhang Z, et al. (2006) KaKs_Calculator: Calculating *Ka* and *Ks* through model selection and model averaging. *Genomics Proteomics Bioinformatics* 4(4):259–263.