2380–2397 *Nucleic Acids Research, 2018, Vol. 46, No. 5* *Published online 22 January 2018*
doi: 10.1093/nar/gky025

The unique epigenetic features of Pack-MULEs and their impact on chromosomal base composition and expression spectrum

Dongyan Zhao[1,2], John P. Hamilton[2], Brieanne Vaillancourt[2], Wenli Zhang[3,4], Georgia C. Eizenga[5], Yuehua Cui[6], Jiming Jiang[1,2], C. Robin Buell[2,*] and Ning Jiang[1,7,*]

[1]Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA, [2]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA, [3]Department of Horticulture, University of Wisconsin, Madison, WI 53705, USA, [4]State Key Laboratory for Crop Genetics and Germplasm Enhancement, Nanjing Agriculture University, Nanjing, Jiangsu 210095, China, [5]USDA-ARS Dale Bumpers National Rice Research Center, 2890 Highway 130 East, Stuttgart, AR 72160, USA, [6]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA and [7]Program in Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI 48824, USA

Received October 27, 2017; Revised December 11, 2017; Editorial Decision January 04, 2018; Accepted January 18, 2018

ABSTRACT

Acquisition and rearrangement of host genes by transposable elements (TEs) is an important mechanism to increase gene diversity as exemplified by the ∼3000 Pack-*Mutator*-like TEs in the rice genome which have acquired gene sequences (Pack-MULEs), yet remain enigmatic. To identify signatures of functioning Pack-MULEs and Pack-MULE evolution, we generated transcriptome, translatome, and epigenome datasets and compared Pack-MULEs to genes and other TE families. Approximately 40% of Pack-MULEs were transcribed with 9% having translation evidence, clearly distinguishing them from other TEs. Pack-MULEs exhibited a unique expression profile associated with specificity in reproductive tissues that may be associated with seed traits. Expressed Pack-MULEs resemble regular protein-coding genes as exhibited by a low level of DNA methylation, association with active histone marks and DNase I hypersensitive sites, and an absence of repressive histone marks, suggesting that a substantial fraction of Pack-MULEs are potentially functional *in vivo*. Interestingly, the expression capacity of Pack-MULEs is independent of the local genomic environment, and the insertion and expression of Pack-MULEs may have altered the local chromosomal expression pattern as well as counteracted the impact of recombination on chromosomal base composition, which has profound consequences on the evolution of chromosome structure.

INTRODUCTION

Transposable elements (TEs) are DNA fragments that can move and amplify in the genome. With rare exception, TEs constitute a large fraction of plant and animal genomes. The majority of TEs are held in check by the host surveillance system, i.e. epigenetic silencing. Despite their contribution to genome size variation, it was hypothesized that the majority of TEs do not have a function although examples of domesticated TEs suggest that some TEs have evolved and are functional (1). Indeed, TE domestication has been reported in animals and plants. In jawed vertebrates, the RAG1 and RAG2 proteins in the V(D)J recombination machinery were derived from an ancient transposon of the *Transib* superfamily (2). In *Arabidopsis thaliana*, the *FHY3* and *FAR1* genes regulate phytochrome A signaling and were derived from *Mutator*-like transposases (MULEs) (3–5). Sequence exaptation from other TE families, as shown by the overexpression of *DAYSLEEPER*, a *hAT*-like transposase, alters expression of many other genes thereby implicating a role for TEs in regulation of global gene expression (6). Aside from exaptation of the entire transposase, computational analyses demonstrated that partial TE sequences have been incorporated into protein-coding genes. In *Arabidopsis thaliana*, 7.8% of expressed genes contained sequences from TEs while 1.2% had translation evidence suggesting the exonization of TEs (7). Similarly, the RNA *Alu* TE elements in human are present in ∼10% of mature mRNAs (8). The extent of domesticated TEs, including en-

*To whom correspondence should be addressed. C. Robin Buell. Tel: +1 517 353 5597; Email: buell@msu.edu, Ning Jiang. Tel: +1 517 353 0381; Email: jiangn@msu.edu

tire or partial TE sequences in plant and animal genomes suggests that TEs have the potential to have an active role in gene evolution.

TEs are also capable of duplicating/capturing normal genes, a process termed transduplication. This phenomenon has been reported for almost all major TE families in plants, albeit the frequency varies among TE families (9–12). Gene duplication by TEs is significant as it has the potential to affect the expression level of the parental gene (the gene from which it is derived). Novel genes can also be created by TE transduplication thereby contributing to the overall gene reservoir of a species. Several TE families, including both DNA transposons and retrotransposons (RNA transposons), are associated with frequent transduplication. In rice, there are a total of 1235 retrogenes that were created by retrotransposons, and a large portion of the retrogenes (42%) have recruited new exons from flanking sequences, leading to the formation of chimeric open reading frames (13). Using homology and structure-based approaches, it was estimated that over 60% of the 1194 intact Helitrons in maize have acquired fragment(s) of nuclear genes (12). A more recent study demonstrated that Helitrons generated ∼11 000 new transcripts in the maize genome; some were chimeric transcripts from different captured genes whereas other transcripts were derived from transcriptional fusion events with nuclear genes in their vicinity (14). Pack-MULEs, belonging to the MULE family, exhibit abundant gene-capture events with ∼3000 Pack-MULEs which have captured over 1500 gene/gene fragments annotated in the Nipponbare rice reference genome (9,15,16). A previous study using full-length cDNA sequences, Massively Parallel Signature Sequencing data, and proteomic data revealed that 22% of rice Pack-MULEs were transcribed and a mere 1% were translated (15). Although these datasets were incomplete with respect to coverage of the rice transcriptome and proteome, they indicate that a subset of Pack-MULEs may be functionally relevant.

Within a genome, TEs and protein-coding genes typically bear distinct epigenetic marks and states of chromatin (17–20). While TEs are usually highly methylated in all cytosine contexts (i.e. CG, CHG and CHH), protein-coding genes are rarely methylated. Interestingly, ∼33% of expressed protein-coding genes in *A. thaliana* are highly methylated in the CG context within the gene body but not in their promoter regions suggesting differential roles of CG methylation in promoters vs. gene bodies (21). A recent study showed differential methylation levels between the terminal inverted repeat and internal regions of *Mutator*-like transposable elements which capture ectopic genomic sequences (22). TEs are often enriched in repressive histone marks (e.g. H3K9me2) and depleted in active histone marks (e.g. H3K4me3) (23), the opposite pattern observed in actively transcribed protein-coding genes. Thus, epigenetic features such as DNA methylation and histone marks are likely to have a role in the formation, evolution, and regulation of sequences derived from these elements. With access to significant improvements in genomics technology, throughput, and resolution, we assessed the transcription, translation, epigenetic and chromatin state of rice protein-coding genes, Pack-MULEs, and their parental genes to understand the impact of gene-capture by TEs at the whole genome level and provide insight into their evolution. Our results indicate that a subset of Pack-MULEs are associated with signatures of active protein-coding genes including bearing active histone marks, enrichment of DNase I hypersensitive sites (DHSs), low DNA methylation, and exhibiting a high frequency of transcription and translation, suggesting that these Pack-MULEs have the potential to contribute to the functional components of the rice genome. In addition, we demonstrate that Pack-MULEs may have influenced the chromosomal base composition and expression patterns in the rice genome.

MATERIALS AND METHODS

Pack-MULEs, protein-coding and other TE gene datasets used in this study

Genome features analyzed in this study are depicted in Figure 1. The annotated Pack-MULEs (2967, Supplementary Dataset S1) in the rice Nipponbare genome were described previously (16) with an additional 43 newly identified Pack-MULEs in this study. Protein-coding (39 049) and TE (16 937) gene sets (MSU Release 7) were downloaded from The Rice Genome Annotation Project (24). A manually curated custom rice repeat library was generated and used to mask the cDNA sequences of all protein-coding genes. A protein-coding gene was excluded if over half of its cDNA sequence was masked by repeat sequences, or if it overlapped with Pack-MULE sequences. The filtered protein coding genes (34 498) were further separated into 'Pack-MULE parental genes' from which Pack-MULEs acquire sequences (1560) and 'other protein- coding genes' (32 938). The cDNA sequences of the 16 937 TE genes were also masked by the custom repeat library, and if 50% or more of the cDNA sequence was masked, the relevant gene was confirmed as a TE gene, which led to 15 461 'other TE genes'. Auto-MULEs (476 MULEs that contain entire or partial transposase sequence) were obtained from a previous study (25). LTR/gypsy elements were from a previous study (26).

Calculation of expression abundance of different gene datasets

We used a suite of large-scale datasets including 61 203 full-length cDNAs (fl-cDNA) that were downloaded from NCBI and kindly provided by Joshua C. Stein (Cold Spring Harbor Laboratory). This is in addition to 45 mRNA-seq samples of various tissues from different developmental stages, and grown under normal and abiotic/biotic stress conditions, and three TRAP-seq samples from *Oryza sativa* cv. Nipponbare (Supplementary Table S1). RNA-seq reads were cleaned using Trimmomatic (v0.32) with the parameters LEADING:5 TRAILING:5 SLIDINGWINDOW:4:10 MINLEN:30 (27). The cleaned reads were then mapped to the Os-Nipponbare-Reference-IRGSP-1.0 pseudomolecules with TopHat (v1.4.1) (28) using a minimum intron size of 5 bp and a maximum intron size of 15 kb. Transcript abundances were generated for the MSU Release 7 representative gene models using Cufflinks (v1.3.0) (29) with a GFF3 file.

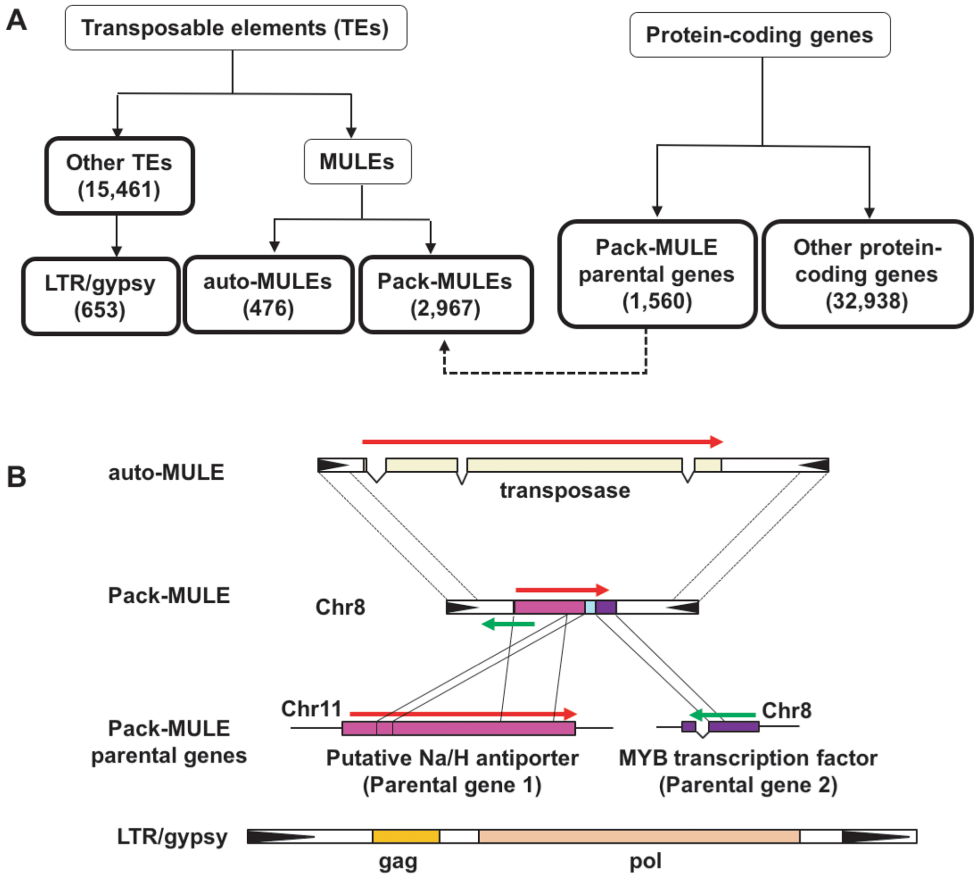

Figure 1. Transposable elements (TEs) and protein-coding genes used in this study. (**A**) TE and protein-coding gene datasets and their relationship. Numbers in parenthesis are the size of each dataset. (**B**) Diagram of the structure of an auto-MULE, a Pack-MULE and its parental genes, and an LTR gypsy retrotransposon. Colored boxes denote open reading frames and white boxes denote non-coding regions; introns are depicted as 'V' shape lines connecting colored boxes. Black triangles denote terminal inverted repeats (TIRs) of an auto-MULE and a Pack-MULE, and long terminal repeat (LTR) for LTR/gypsy retrotransposon. Homologous sequences are connected by solid or dashed lines; light blue boxes represent exons where the origin of the sequence is unclear. Red and green arrows indicate transcribed regions and their orientation.

To provide equivalent estimations of expression between Pack-MULEs and protein-coding genes, the transcribed regions of Pack-MULEs were determined using *ab initio* transcripts predicted using Cufflinks (v1.3.0) (29). Briefly, the cleaned reads of 13 representative mRNA-seq and three TRAP-seq datasets (Supplementary Table S1) were aligned to the Os-Nipponbare-Reference-IRGSP-1.0 pseudomolecules using TopHat (v1.4.1) (28) in the uniquely mapping mode (-g 1) and requiring a minimum intron size of 5 bp and a maximum intron size of 15 kb. *Ab initio* transcripts and their associated genomic features (gtf) were generated using Cufflinks with no reference annotation provided (omitting the –GTF option). Genomic features of transcripts that overlapped with Pack-MULEs were extracted and redundant transcripts were merged requiring a minimum exon length of 40 bp and minimum transcript length of 200 bp to be considered as a valid transcript for a Pack-MULE. A genomic feature file (GFF) was generated for Pack-MULEs with qualified transcripts. For those with no expression in any of the 16 expression datasets used for determination of transcribed regions, the entire regions of those Pack-MULEs were used as the base for calculation of expression abundance. The final GFF file was used to calculate expression abundance of Pack-MULEs using Cufflinks (v1.3.0) (29).

Analysis of DNase-hypersensitivity sites

DNase I hypersensitive sites (DHSs) were from a previous study (30) and the locations of DHS peaks around each protein-coding or TE gene were categorized into five groups (0.2–1 kb and 0.2 kb of the 5′ flanking sequences, gene/TE body, 0.2 kb and 0.2–1 kb of the 3′ flanking sequences).

Construction of bisulfite-sequencing (BS-seq) and chromatin-immunoprecipitation-sequencing (ChIP) libraries and data analyses

Rice cultivar Nipponbare was used for BS-seq and ChIP-seq analyses. For shoot tissue, plants were grown in a greenhouse with a temperature of 32–35°C and 12h light/dark cycle. Immature panicles at late R1 or early R2 growth stage (31) were harvested from plants grown in irrigated rice fields at Stuttgart, Arkansas in September (∼12 h daylight). Shoot and panicles (≤5 cm in length) were collected for DNA isolation using DNeasy plant mini

kit (Qiagen, Hilden, Germany). A total of 5 μg of DNA was fragmented by sonication to a mean size of 250 bp. Sonicated DNA was purified using the MinElute PCR Purification Kit (Qiagen, Hilden, Germany) and eluted twice with 17 μl of EB each. A total of 3 μg of sonicated DNA spiked with fragmented unmethylated lambda DNA at a final concentration of 0.5% (Promega, Madison, WI, USA) was end repaired using the End-It Kit (Epicentre, Madison, WI, USA), followed by 3′-end addition of dA using Klenow (3′→5′ exo$^-$) (NEB, Ipswich, MA, USA). A-tailed DNA fragments were ligated with methylated Illumina DNA adapters using LigaFast (Promega, Madison, WI, USA). Ligated DNA ranging from 150 to 220 bp was sized from a 2% agarose gel and purified using MinElute Gel Extraction Kit (Qiagen, Hilden, Germany). The bisulfite conversion of gel purified DNA fragments was conducted according to the manufacturer's instruction from EZ DNA methylation-lightning kit (ZYMO Research, Irvine, CA, USA). Bisulfite-treated DNA was PCR amplified with 10–12 cycles using Pfu Turbo Cx hotstart DNA polymerase (Stratagene, Santa Clara, CA, USA) and PE PCR Primer 1.0 (5′-AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT) and PE PCR Primer 2.0 (5′-CAAGCAGAAGACGGCATAC GAGATCGGTCTCGGCATTCCTGCTGAACCGCT CTTCCGATCT). PCR amplified DNA was purified by running a 2% agarose gel and the resulting BS-seq library was sequenced in paired-end mode on the Illunima HiSeq 2000 and 2500 platforms (Supplementary Table S2).

The BS-seq reads were cleaned using Cutadapt (v1.2.1; -m 41 –q 10) (32) to remove adapters and low quality bases. The cleaned reads were mapped in single-end mode to the Os-Nipponbare-Reference-IRGSP-1.0 pseudomolecules and methylation calls were generated using Bismark (v0.10.0) (33) for the CG, CHG, and CHH contexts.

ChIP experiments and barcoded ChIP-seq library preparation were performed following published protocols (34) using ChIP-grade commercial antibodies against H3K4me3 (07-473, Millipore, MA, USA), H4K12ac (07–595, Millipore, MA, USA), and H3K9me2 (07-441, Millipore, MA, USA). All BS-seq and ChIP-seq libraries were sequenced on the Illumina HiSeq 2000 and 2500 platforms (Supplementary Table S2). The ChIP-Seq reads were cleaned using Cutadapt (32) (v1.2.1; -m 71 –q 10) to remove adapters and low quality bases. The cleaned reads were mapped using Bowtie (v1.1.0) (35) to the Os-Nipponbare-Reference-IRGSP-1.0 pseudomolecules with the parameters: -v 2 -k 1 -m 1 –best. Putative ChIP-enriched-regions associated with each histone mark were called with SICER (v1.1) (36) using a window size of 200 bp and a gap size of 200 bp.

Estimation of recombination rate and pericentromeric regions

The recombination rate in rice was adapted from previous studies (37,38). Recombination rates in exact 2 Mb window (centiMorgan/Mb) except the last window on each chromosome were calculated. If two markers spanned more than 2 Mb sequence, the genetic distance (cM) between them was assigned to each 2 Mb window based on their sequence proportion in each window. The positions (Mb) of centromeres in the rice chromosomes were from the Rice Genome Annotation Project (24). The pericentromeric regions were defined as regions flanking the centromeres with recombination rate less than 2 cM/Mb. In this way, the pericentromeric regions ranged from 4 to 8 Mb among 12 chromosomes.

Detection of the presence and absence of Pack-MULEs between the genome of Nipponbare and that of other cultivars and wild relatives

In addition to the genome of Nipponbare, five other genomes were used for comparison. The genome assembly of three *indica* cultivars and *O. punctata* was downloaded from NCBI. The accession numbers were CM003910–CM003921 for Zhenshan 97, CM003922–CM003933 for Minghui 63, CP018157–CP018168 for Shuhui498 and CM002488–CM002499 for *O. punctata*. The genome assembly of *O. meridionalis* (v1.3) was downloaded from Ensembl Genomes 37 (https://plants.ensembl.org/Oryza_meridionalis/Info/Index). The sequence identify cutoff for an orthologous position is 95% for *indica*, 90% for *O. meridionalis* and 85% for *O. punctata* with an *E* value < 10^{-10} (BLASTN). To detect the presence of Pack-MULEs in these other Oryza genomes, the junction sequence containing both flanking and TIR (100 bp flanking plus 100 bp TIR) from Nipponbare was used to search the other genome. Both the 5′ and 3′ junctions were used for search. If the junction sequence had a match (see above for cutoff requirement) and the alignment between the sequence from Nipponbare and the other genome crossed the junction point, extended at least 30 bp on each side, and was on the same chromosome with the same orientation, the relevant element was considered to be present in the other genome. As the flanking plus element junction is unique, no multiple mapping of junction sites were observed. For the remainder of the Pack-MULE loci, a junction sequence resembling the 'empty site' was constructed by combining the 5′ and 3′ flanking sequence, 100 bp on each side excluding one copy of TSD. Thereafter the 'empty site' sequences were used to search the other genome, and the criteria for defining 'absence' is similar to that for 'presence', i.e., if the 'empty site' sequence had a match and the alignment between sequence from Nipponbare and the other genome crossed the junction point, extended at least 30 bp on each side, and the match was on the same chromosome with same orientation, the relevant element was considered to be absent from the other genome. In addition, the match had to be unique on the chromosome. If a Pack-MULE was neither in the 'presence' nor 'absence' group, it was classified as 'ambiguous', which indicates the status of the locus is uncertain. Those included if the insertion site was absent or located in sequencing gaps, the insertion site was too divergent to detect, the insertion site mapped to multiple locations, and the orientation of the insertion site is flipped or mapped to another chromosome. The Pack-MULEs that were present in the other genomes but absent from Nipponbare were not surveyed.

Statistical analyses

Kolmogorov–Smirnov (KS), Student's t, and χ^2 tests were conducted using SAS/9.4. Pearson's r analyses were conducted using R/3.2.3.

Full methods and associated references are in the Supplementary Data—Supplementary Materials and Methods.

RESULTS

Frequent transcription and translation of Pack-MULEs

A total of 2967 Pack-MULEs, which captured gene or gene fragments from 1560 unique protein-coding genes (Pack-MULE parental genes) were annotated in Os-Nipponbare-Reference-IRGSP-1.0 pseudomolecules (24) (Figure 1). Based on 45 diverse mRNA-sequencing (mRNA-seq) datasets and three Translating Ribosome Affinity Purification RNA-sequencing (TRAP-seq) datasets from a range of developmental stages and treatments of *Oryza sativa* cv. Nipponbare (Supplementary Table S1; 4 callus samples, 18 reproductive and 26 vegetative samples) (37,39–44), 92.8% of Pack-MULE parental genes ($n = 1560$) and 81.9% of all other protein-coding genes ($n = 32\ 938$) were expressed, consistent with the hypothesis that Pack-MULEs capture *bona fide* genes that are often expressed (Supplementary Figure S1A) (16). For Pack-MULEs, 40.1% were expressed, which is significantly higher than the previously estimated 22% (15), and is ∼3-fold higher than that of other transposable element (TE) genes (10.2%) (Supplementary Figure S1A). Classification of the 45 mRNA-seq and three TRAP-seq samples into 16 tissues revealed that Pack-MULE parental genes and other protein-coding genes had a broader breadth of expression with the majority of genes (∼65%) expressed in eight or more tissues ($\chi^2 = 5613.2044$, $P < 0.0001$) compared to Pack-MULEs and other TE genes (Supplementary Figure S1B). Specifically, Pack-MULE parental genes are more widely expressed than other protein-coding genes (10 versus 8 tissues, median value), but there are less constitutively expressed genes among parental genes than other protein-coding genes (9.6% versus 14.4%). Pack-MULEs showed a slightly broader expression than other TE genes (Supplementary Figure S1B) with 16.2% of the Pack-MULEs (or 38% of the expressed Pack-MULEs) exclusively expressed in reproductive tissues, significantly higher than all of the other gene sets (10.6% of Pack-MULE parental genes, 12.0% of other protein-coding genes and 5.2% of other TE genes; $\chi^2 = 643.96$, $P < 0.0001$, Supplementary Table S3). This contrasts with vegetative tissue-restricted expression, which was similar among Pack-MULEs (4.4%), Pack-MULE parental genes (4.3%), and other protein-coding genes (4.5%) ($P > 0.6$) and distinct from other TE genes with a limited number with vegetative-restricted expression (1.2%) (Supplementary Table S3). It is unlikely that the enriched expression in reproductive tissues of Pack-MULEs is due to the expression specificity of their parental genes as no significant correlation was observed between expression of Pack-MULEs and Pack-MULE parental genes (Pearson's r analysis, $P > 0.41$). This is in contrast to the retrogenes in rice, which demonstrate similar tissue specificity to their parental genes (45).

Not all transcripts are translated and evidence of translation would suggest potential function of Pack-MULEs at the protein level. TRAP-seq is a method to determine mRNAs associated with ribosomes by immuno-precipitation of ribosome-RNA complexes followed by high-throughput sequencing of the mRNAs. The ratio between TRAP-seq expression measured in fragments per kb exon model per million mapped reads (FPKM) and mRNA-seq FPKM within the same tissue, referred to as Translatome Enrichment Index (TEI), was used to determine the enrichment of mRNAs on ribosomes as previously described (37). Analysis of TRAP-seq data revealed that 72.9% of the Pack-MULE parental genes and 60.2% of other protein-coding genes had translation evidence compared to 9.0% of Pack-MULEs and 1.2% of other TE genes (Figure 2). Although Pack-MULEs tend to have low levels of transcription (median mRNA-seq FPKM of 1.69 versus 7.65 for Pack-MULE parental genes) and translation (median TRAP-seq FPKM of 2.99 versus 7.33 for Pack-MULE parental genes), their TEI was 1.43 which is significantly higher than their parental genes (TEI = 1.07; Kolmogorov-Smirnov (KS) test, $P = 0.0017$), other protein-coding genes (TEI = 0.87; KS test, $P < 0.0001$), and other TE genes (TEI = 0.52; KS test, $P < 0.0001$), suggesting that Pack-MULEs have higher translation efficiency than other TE or protein-coding genes (Supplementary Figure S1C). Interestingly, the number of Pack-MULEs translated in panicles was over twice (179) that in the shoots (75) and calli (79). This pattern was not observed with parental genes, other protein-coding genes, or other TE genes suggesting that Pack-MULEs may be preferentially transcribed and translated in reproductive tissues (Figure 2, Supplementary Table S3).

In summary, the transcription and translation profiles of Pack-MULEs are distinct from other TE genes and protein-coding genes. For subsequent analyses, unless specified, we analyzed and compared: expressed Pack-MULEs (1189) which have an mRNA-seq or TRAP-seq FPKM ≥ 1 in at least one sample and/or have proteomic evidence based on previous proteomic studies (46–51); non-expressed Pack-MULEs (1070) which have an FPKM ≤ 0.1 in any expression dataset; the remaining Pack-MULEs (708) with $0.1 <$ FPKM < 1 were excluded from downstream analyses unless specified.

Enrichment of DNase I hypersensitive sites within Pack-MULEs and their proximal regions

DNase I hypersensitive sites (DHSs) are genomic regions depleted in nucleosomes, enriched with *cis*-regulatory sequences and are associated with most promoters and enhancers (52,53). Using a previously described rice DHSs dataset (30), we examined the presence of DHSs in the gene bodies and flanking sequences of Pack-MULEs, Pack-MULE parental genes, other protein-coding genes, and other TE genes. As shown in Figure 3A, the majority (82.1%) of Pack-MULE parental genes harbored at least one DHS within the gene body and 1 kb flanking sequence, suggesting the overall openness of the local chromatin. Similar fractions (66.8% versus 65.7%) of Pack-MULEs and other protein-coding genes harbored at least one DHS. By contrast, only 9.5% of other TE genes were associ-

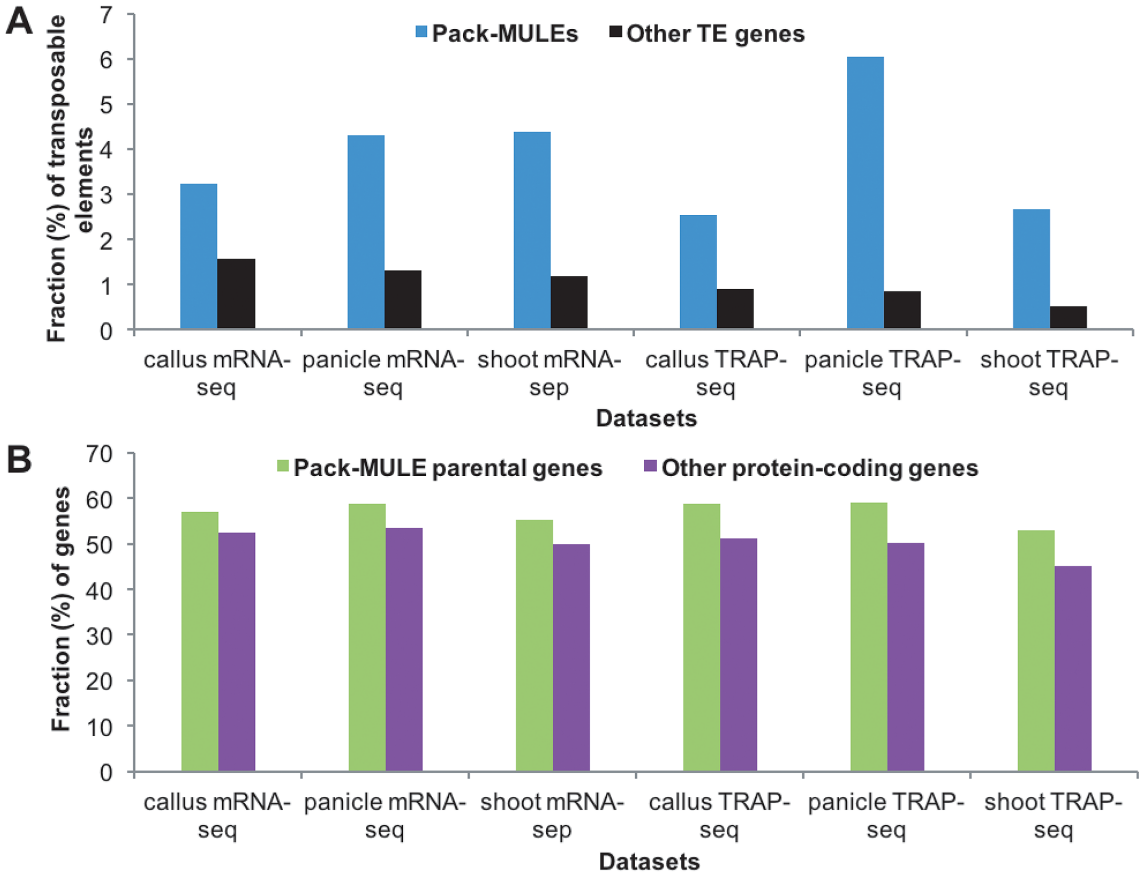

Figure 2. Transcription and translation profiles of Pack-MULEs, other TE genes, parental genes, and other protein-coding genes. (**A**) Fraction of Pack-MULEs and other TE genes with transcription and translation evidence in callus, panicle, and shoot. (**B**) Fraction of Pack-MULE parental genes and other protein-coding genes with transcription and translation evidence in callus, panicle and shoot.

ated with a DHS, which clearly distinguishes Pack-MULEs from other TE genes. For both Pack-MULE parental genes and other protein-coding genes, an enrichment of DHSs 0.2 kb upstream of the transcription start site (TSS) was observed, coincident with the localization of promoter sequences upstream of the TSS and active transcription. Moreover, the DHS density immediately upstream of Pack-MULE parental genes was higher than that of other protein coding genes (t test, $P < 0.0001$; Figure 3B). In contrast, few DHSs were observed within or flanking other TE genes. Interestingly, for both expressed and non-expressed Pack-MULEs, an enrichment of DHSs in the upstream and downstream 0.2 kb flanking sequences was observed; even for non-expressed Pack-MULEs, the DHS density flanking the elements was significantly higher than that of the genome average (1.15 versus 0.16, t test, $P < 0.0001$). The terminal inverted repeats (TIRs) of *MuDR* (a maize MULE element that encodes transposase) serves as the promoter for transcription of the element (54) and a higher density of DHSs was observed in the TIRs of expressed Pack-MULEs compared to non-expressed Pack-MULEs (KS test, $P < 0.0001$), suggesting that TIRs function as promoters and direct transcription of Pack-MULEs (Figure 3B). Additionally, within Pack-MULE internal sequences, significantly more DHSs were present inside expressed Pack-MULEs relative to non-expressed Pack-MULEs, parental genes, and other protein-coding genes (t test, $P < 0.0001$; Figure 3B). To determine whether transcription orientation affected DHS enrichment in TIRs, Pack-MULEs with cDNA evidence were analyzed. As shown in Figure 3B, TIRs upstream of the TSS seem to contain slightly more DHSs than TIRs downstream of the transcription termination site (TTS); however, the difference is not significant (t test, $P = 0.3284$).

The local genomic context can influence chromatin state and Pack-MULEs within 1 kb of protein-coding genes exhibited higher DHS density in their 0.2 kb flanking sequences compared with Pack-MULEs lacking protein-coding genes in their vicinity (t test, $P < 0.0001$; Figure 3C), suggesting a positive effect of the presence of other protein-coding genes on the 'openness' of chromatin near Pack-MULEs. To determine whether Pack-MULEs impact the local chromatin state, we compared protein-coding genes with and without Pack-MULEs at their 5′ or 3′ end, within 1 kb flanking sequences. As shown in Figure 3C, protein-coding genes with adjacent Pack-MULEs at their 5′ ends exhibited significantly higher DHS density in the 5′ end regions, including upstream 0.2 kb and 0.2–1 kb flanking sequences compared with those without adjacent Pack-MULEs (t test, $P < 0.0001$). Similarly, protein-coding genes with Pack-MULEs at their 3′ ends showed significant DHS enrichment in the 3′ end regions, including the downstream

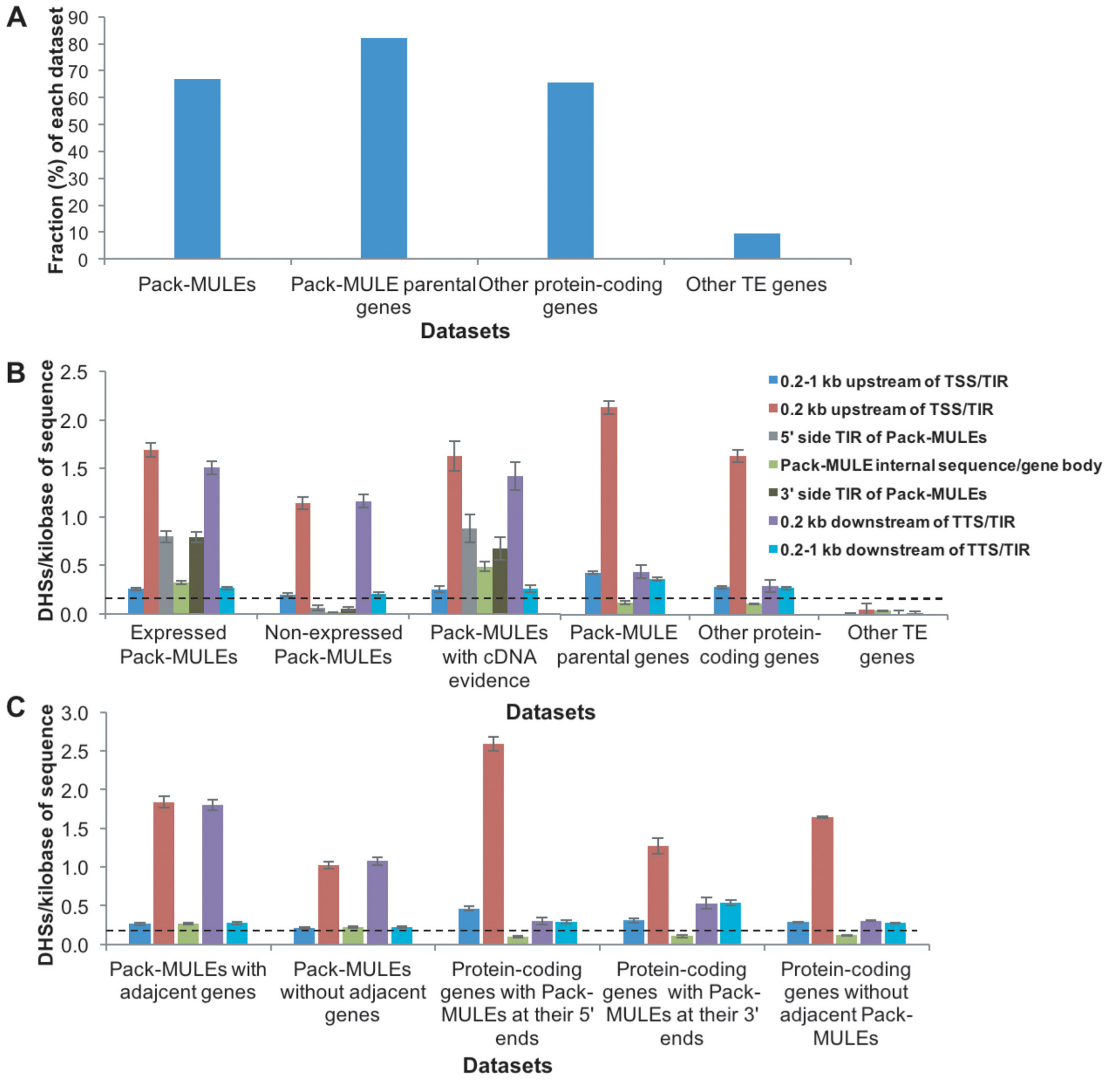

Figure 3. The density of DNase I hypersensitive sites for TEs, Pack-MULEs, protein-coding genes, and their flanking sequences. (**A**) Fraction of TE/protein-coding genes with at least one DHS in the gene body and 1 kb flanking sequence. (**B**) DHS density for Pack-MULEs with and without expression evidence, Pack-MULEs with cDNA evidence, Pack-MULE parental genes, other protein-coding genes and other TE genes. (**C**) DHS density for Pack-MULEs with and without protein-coding gene(s) in their 1 kb flanking sequences, protein-coding genes with and without adjacent Pack-MULEs at 5′ or 3′ end. For (B) and (**C**), the y-axis represents a weighted measure of DHSs based on the length of the sequences (DHSs per kilobase sequence). The horizontal dashed lines indicate the genome average number of DHSs per kilobase of sequences. Error bars are standard errors (SEMs).

0.2–1 kb sequences in comparison to those without adjacent Pack-MULEs (*t* test, $P < 0.0001$, Figure 3C), suggesting that the presence of Pack-MULEs may have contributed to the 'openness' of local chromatin (also see discussion). To determine whether Pack-MULE internal sequences have a role in DHS enrichment, we compared Pack-MULEs with auto-MULEs (MULEs with partial or complete transposase sequence) since they both share similar transposon terminal regions or TIRs and therefore should have similar target specificity. Negligible DHSs were found within auto-MULEs with less DHS enrichment in the 0.2 kb flanking sequences compared to that of Pack-MULEs (*t* test, $P < 0.001$; Supplementary Figure S1D), suggesting that the non-TE internal sequences of Pack-MULEs may contribute to the 'open' chromatin status.

Epigenetic marks associated with Pack-MULEs, their parental genes, and other genes/TEs

To determine the epigenetic status of Pack-MULEs, we conducted whole-genome bi-sulfite sequencing (BS-seq) and chromatin immuno-precipitation with antibodies targeting three modified histones followed by high-throughput sequencing (ChIP-seq) using both shoot and panicle tissues. Of the expressed Pack-MULEs, 72.6% had low CG methylation in their internal regions as defined by <10% of CGs methylated in panicles, similar to parental genes (73.0%) yet substantially higher than other protein-coding genes (43.4%), non-expressed Pack-MULEs (11.5%), and auto-MULEs (1.5%) (Figure 4A). For CHG methylation, the majority of parental genes (96.6%) were lowly methylated (<10% CHG methylation), followed by other protein-coding genes (82.4%), and expressed Pack-MULEs (75.6%)

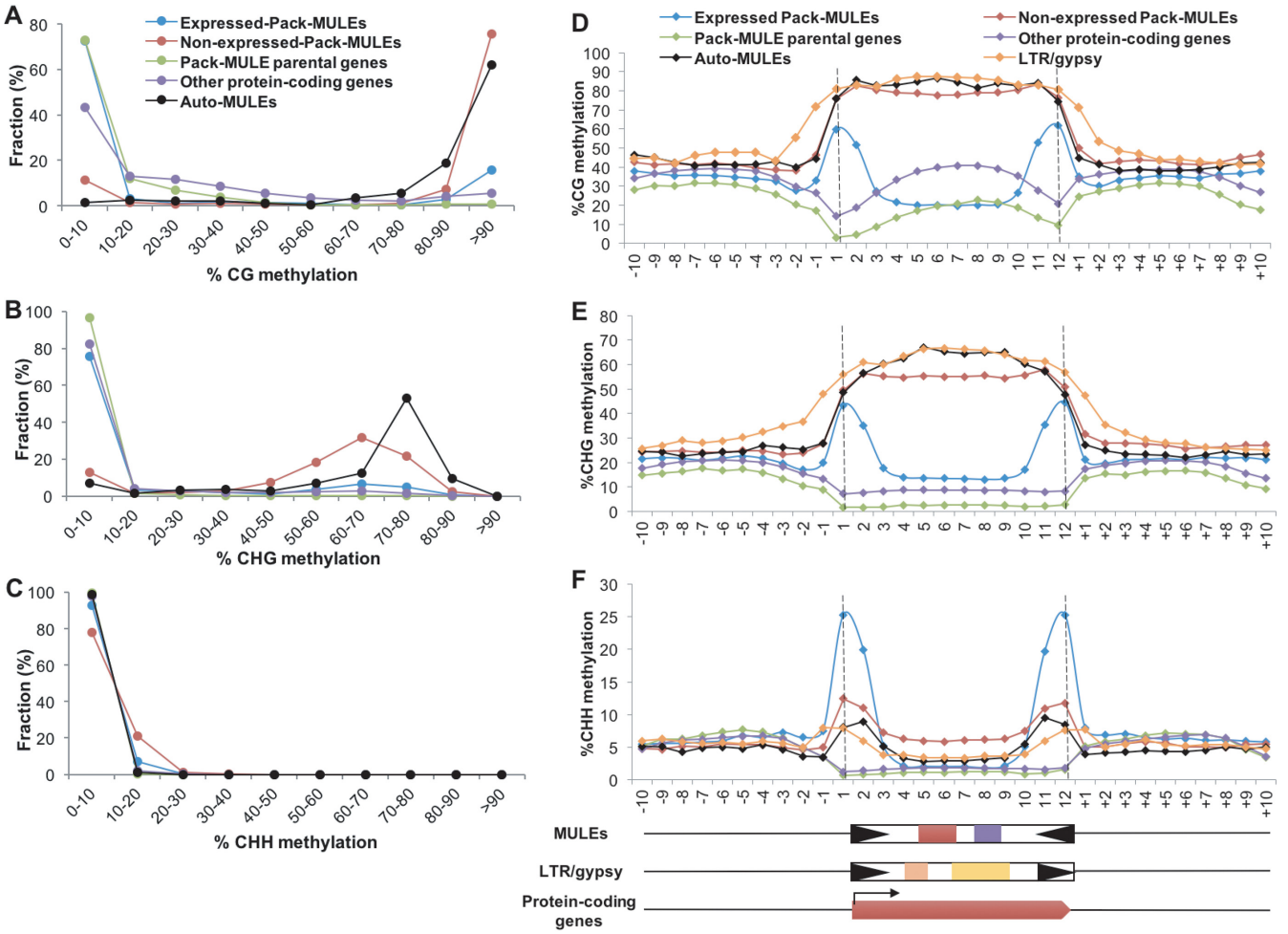

Figure 4. DNA methylation in rice young panicles. (A–C). Percent DNA methylation of internal regions of expressed-Pack-MULEs, non-expressed-Pack-MULEs, Pack-MULE parental genes, other protein-coding genes, and auto-MULE internal regions in CG (**A**), CHG (**B**) and CHH (**C**) contexts. (D–F) Percent DNA methylation along the length of expressed-Pack-MULEs, non-expressed-Pack-MULEs, Pack-MULE parental genes, other protein-coding genes, auto-MULEs, and LTR/gypsy elements in CG (**D**), CHG (**E**) and CHH (**F**) contexts. The entire sequence of Pack-MULEs, auto-MULEs, and LTR/gypsy elements was divided into 12 bins, with 2 bins for each TIR/LTR on both ends and 8 bins for the internal sequences. For protein-coding genes, their body sequence was divided into 12-equal-sized bins. For all datasets, the 1 kb flanking sequence was divided into 10 bins (100 bp/bin). Regions between dashed lines denote the boundary between gene/TE body and flanking sequence.

while only a small fraction of non-expressed Pack-MULEs (12.9%) and auto-MULEs (6.9%) were lowly methylated (Figure 4B). For CHH methylation, the majority of all gene sets were lowly methylated (<10% CHH methylation) with non-expressed Pack-MULEs having the highest CHH methylation (Figure 4C).

To better understand the distribution of methylation, we binned genes, Pack-MULEs, and other TEs into 12 bins and binned the 1 kb flanking regions into 10 bins. Previous studies in maize demonstrated that TE families differ in the extent of methylation dispersion with Long Terminal Repeat (LTR) elements exhibiting the most dramatic dispersion (55). In rice, LTR elements exhibited a gradual decline of methylation level based on distances from their LTRs that contrasted with a sharp boundary in methylation observed in all three cytosine contexts from the TIRs to the immediate flanking sequence for Pack-MULEs and auto-MULEs (Figure 4D–F), suggesting that methylation of MULEs does not significantly influence the methylation of their flanking sequences. As shown in Figure 4, Pack-MULE parental genes had the lowest CG and CHG methylation across gene body and their flanking sequences. Expressed Pack-MULEs had similar CG methylation and slightly higher CHG methylation in their internal regions compared to Pack-MULE parental genes. In contrast, non-expressed Pack-MULEs had high CG and CHG methylation in their internal regions that was only slightly lower than that of auto-MULEs (Figure 4D and E). Due to their repetitive nature, it is not surprising that the TIRs of Pack-MULEs were highly methylated compared with their internal regions and if TIR regions were excluded, expressed Pack-MULEs exhibited similarly low CHH methylation as Pack-MULE parental genes and other protein-coding genes, including gene bodies and their flanking sequences (Figure 4F). Expressed Pack-MULEs exhibited lower CG and CHG methylation than non-expressed Pack-MULEs in both TIRs and internal regions; however, dramatically higher CHH methylation was observed in the

TIRs of expressed Pack-MULEs than that of non-expressed Pack-MULEs or auto-MULEs (Figure 4). A previous study suggested that TEs in genic regions are frequently highly methylated in the CHH context, a mechanism by which transcription initiation in TEs is inhibited (56). Indeed, the CHH methylation level of TIRs was slightly higher for Pack-MULEs that are within 1 kb of protein-coding genes than those that are not close to other genes (mean CHH methylation 19.4% versus 17.7%; KS test, $P < 0.01$), confirming that genomic environment influences the CHH methylation of Pack-MULE TIRs or the surveillance machinery is more robust in genic regions.

In addition to DNA methylation, covalent modifications of histone proteins impact gene expression (23). The presence of active histone marks coincident with the lack of repressive marks generally corresponds to a low-level of DNA methylation indicating a high expression potential. We examined three histone marks using ChIP-seq: H3K4me3 and H4K12ac associated with open chromatin and active transcription and H3K9me2 associated with closed chromatin and a repressive state of transcription (57). For this analysis, if a gene overlapped with the peak of a histone modification enriched region, the gene was considered to harbor the histone modification. A total of 65.9% of expressed Pack-MULEs, 89.7% of Pack-MULE parental genes, and 65.5% of other protein-coding genes harbored the active histone mark, H3K4me3, in their gene bodies while a limited number of expressed Pack-MULEs (7.6%), Pack-MULE parental genes (0.7%), and other protein-coding genes (3.6%) harbored the repressive mark, H3K9me2, reinforcing the findings that expressed Pack-MULEs resemble regular protein-coding genes and maintain expression capacity (Figure 5). In contrast, only 9.5% of non-expressed Pack-MULEs harbored H3K4me3 yet 33.7% were associated with the repressive histone mark (H3K9me2), a trend similar to that of other TE genes (3.3% with H3K4me3 and 34.5% with H3K9me2). In contrast to the internal region of Pack-MULEs, the TIRs of Pack-MULEs were depleted in the repressive mark (H3K9me2) despite their high methylation rate (Figure 4D–F) and expression status (bins 1–2 and 11–12, Figure 5C). Interestingly, expressed Pack-MULEs, parental genes, and other protein-coding genes also differ from non-expressed Pack-MULEs and other TEs in the enrichment of H4K12ac, another active histone mark. About 28.7% of expressed Pack-MULEs, 42.5% of Pack-MULE parental genes, and 35.4% of other protein-coding genes contained H4K12ac while only 6.4% of non-expressed Pack-MULEs and 7.7% of other TE genes contained this histone mark. The high frequency of active histone marks and lack of the repressive histone marks in expressed Pack-MULEs are reminiscent of protein-coding genes distinguishing them from non-expressed Pack-MULEs and other TEs (Figure 5). Similar patterns of methylation and histone marks were observed in shoots (Supplementary Figures S2 and S3).

The effect of terminal and internal sequences on Pack-MULE expression

The above data clearly indicate that the TIR and internal regions of Pack-MULEs have distinct epigenetic features. Although the internal regions of Pack-MULEs are low copy sequences, their TIRs are repetitive and have been classified into 122 different families (16). Consistent with previous analyses on the influence of TIRs on expression (15), expression frequency of Pack-MULEs is impacted by TIR family membership as shown in Supplementary Figure S4A with five TIR families that were the most over-represented in expressed Pack-MULEs and another five TIR families that were most enriched in non-expressed Pack-MULEs (t test, $P < 0.0001$). Analysis of 372 Pack-MULEs with full-length cDNA evidence showed that over half (58.6%) had transcription initiation within the TIR or not far from TIR ($\leq$200 bp). The remainder are within flanking sequence (12.6%) or from other regions inside the element (28.8%). The predominance of transcription initiation within or close to TIRs is consistent with the observation that promoters were located inside TIRs (54). It is worth noting that a small subset of Pack-MULEs contains regulatory/promoter sequences (the sequence upstream of 5′ UTR regions) from protein-coding genes and more expressed Pack-MULEs carry regulatory sequences than their non-expressed counterparts (82 versus 52, χ^2 test, $P < 0.01$). In summary, specific TIR families and the presence of regulatory regions favor the expression of Pack-MULEs.

The influence of duplication and insertion time on Pack-MULE expression

To test whether the age of duplicated fragments within Pack-MULEs is correlated with expression, the transversion rate (Tv) of Pack-MULE acquired regions was compared to their parental genes and used as an estimate for the time of gene duplication/acquisition by Pack-MULEs (16). Statistical analyses indicated that Pack-MULEs with an intermediate transversion rate (Tv = 1–3) were more likely to be expressed than Pack-MULEs either very young (Tv $\leq$ 1) or relatively old (Tv > 3) ($\chi^2 = 38.20$, $P < 0.0001$) (Supplementary Figure S4B). Interestingly, more Pack-MULEs are expressed in pistils and developing seeds compared to male organs (i.e. anther and sperm cells, Supplementary Figure S4C). Moreover, these differences are largely due to the over-representation of young and middle-aged (Tv < 3) Pack-MULEs in pistil and young seed transcriptomes (χ^2 test, $P < 0.05$; Supplementary Figure S4C) suggesting that Pack-MULEs with young and middle-aged duplications are components of female and primarily female-derived transcriptome but not male organs.

In addition, the insertion time of Pack-MULEs was estimated based on presence/absence variation (Pack-MULE insertion polymorphism) between Nipponbare (*O. sativa japonica*) and five other *Oryza* genomes (see Materials and Methods for details) including three *O. sativa indica* cultivars (Zhenshan 97, Minghui 63 and Shuhui 498) (58,59), which diverged from *japonica* ~0.5 million years (MY) (60), and two wild rice species *O. meridionalis* and *O. punctata* (Supplementary Table S5). *O. meridionalis* diverged from Asian rice about 3 MY (61) and are associated with AA genotype along with the *O. sativa* genomes (62). *O. punctata*, which is associated with BB genotype, is diverged from the AA genome ~6 MY (63). Based on the presence and

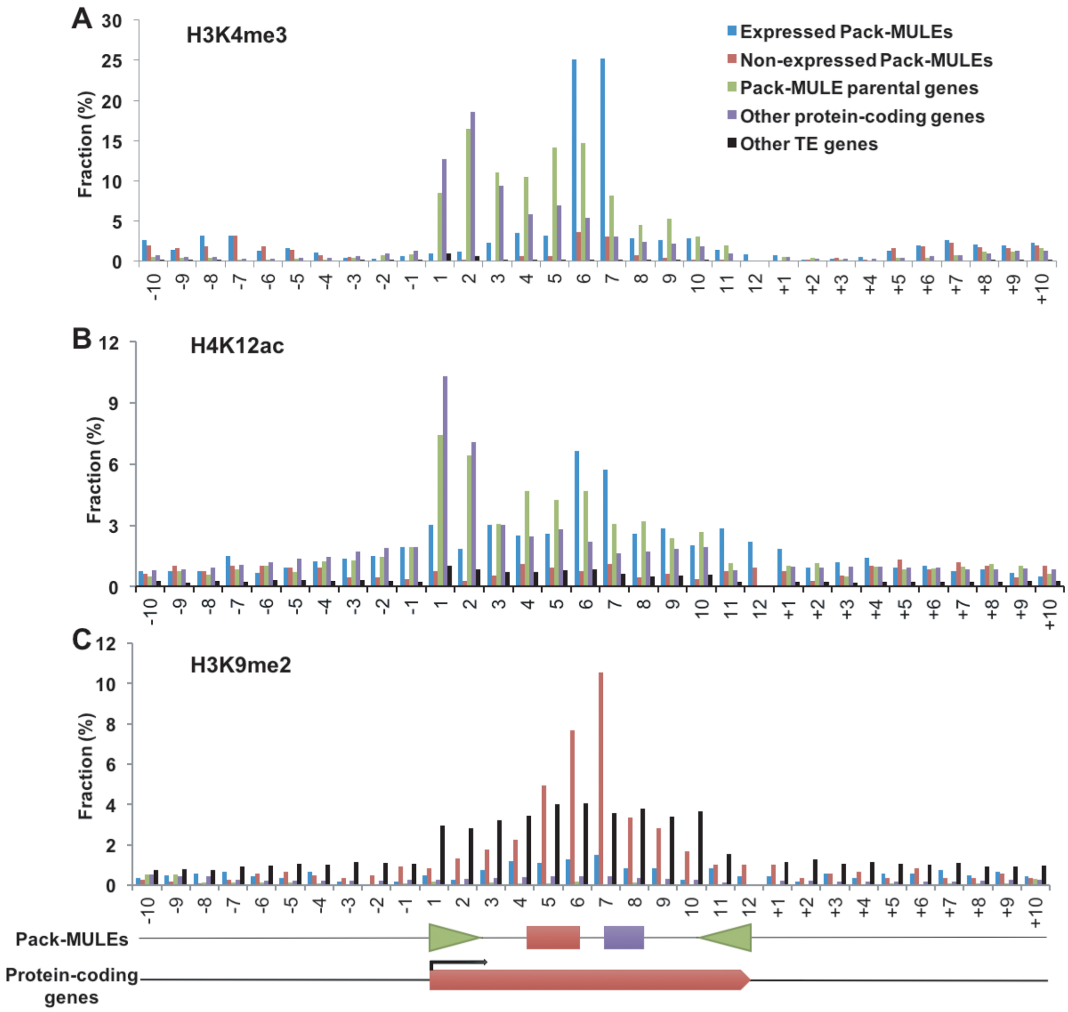

Figure 5. Fraction of TEs, Pack-MULEs, and protein-coding genes with histone modifications in rice young panicles. (**A**) H3K4me3: tri-methylation of lysine 4 on histone 3; (**B**) H4K12ac: acetylation of lysine 12 on histone 4; (**C**) H3K9me2: di-methylation of lysine 9 on histone 3 of expressed and non-expressed Pack-MULEs, Pack-MULE parental genes, other protein-coding genes, and other TE genes. Gene and TE body sequence was divided into 12-equal-sized bins and the 1 kb flanking sequence was divided into 10 bins (100 bp/bin), similar to that in Figure 4.

absence information in these comparator genomes, Pack-MULEs were classified into five groups. The most ancient group is shared between Nipponbare and *O. punctata* and only 21 Pack-MULEs are in this group, suggesting that the majority of Pack-MULEs in Nipponbare were formed in the last 6 MY. The second ancient group of Pack-MULEs (647 elements) is shared between Nipponbare and *O. meridionalis*, which diverged more than 3 MY ago. However, 577 Nipponbare Pack-MULEs are absent from *O. meridionalis* suggesting that approximately half of the Pack-MULEs were formed in last 3 MY. The third group (522 elements) is composed of Pack-MULEs absent from *O. meridionalis* but not polymorphic between Nipponbare and the *indica* cultivars, suggestive of an insertion time of 0.5–3.0 MY, representing a mixture of recent and middle-aged elements. The fourth group (110 elements) represent elements polymorphic between Nipponbare and any one of the *indica* cultivars, and not present in *O. meridionalis*, suggestive of an insertion time of 0.5 MY. The most recent group (50 elements) are elements only present in Nipponbare and not in any of the other genomes, which are very recent insertions.

As shown in Supplementary Table S6, the insertion time of Pack-MULEs is positively correlated albeit not perfectly proportional to the transversion rate between Pack-MULEs and their parental genes. This is because duplication and transposition represent different components of the Pack-MULE activity which are not necessarily coupled (see discussion). Consistent with transversion rates, elements inserted within 0.5–3 MY are much more frequently expressed than very young (<0.5 MY) and old elements (>3 MY) (Supplementary Figure S4D). In fact, the variation of expression frequency seems to be more dramatic among groups with different insertion time than those with different transversion rate (Supplementary Figures S4B and D), suggesting that the age of the insertion maybe more critical than the age of duplication. Moreover, the elevated expression is mostly enriched in pistil and seeds (Supplementary Figure S4E).

The chromosomal distribution of Pack-MULEs and its impact on expression

To assess whether the genomic context affects expression status, we compared the local recombination rates and observed that expressed Pack-MULEs are located in regions with slightly higher recombination rates compared with that of all Pack-MULEs, albeit the difference is not significant (KS test, $P = 0.2485$). This contrasts with expressed protein-coding genes, which were preferentially distributed in regions of the genome with increased recombination rates (KS test, $P = 0.0097$). For simplicity, we divided the genome into two regions: euchromatic chromosome arms with relatively high recombination rates (307 Mb, 82%) and pericentromeric regions with low recombination rates (67 Mb, 18% of the genome) (see Materials and Methods for details). As shown in Figure 6A and B, the pericentromeric regions are associated with fewer protein-coding genes than euchromatic chromosomal arms. Pack-MULEs demonstrate a similar distribution preference, albeit the difference between pericentromeric regions and euchromatic chromosomal arms is not as dramatic as that of protein-coding genes. Not surprisingly, protein-coding genes in pericentromeric regions are less likely to be expressed as indicated by the lower fraction of expressed protein-coding genes in pericentromeric regions than that on euchromatic chromosomal arms (Figure 6C, 82.2% versus 71.4%, KS test, $P < 0.0001$). In contrast, there is no obvious correlation between the fraction of expressed Pack-MULEs and the recombination rate or locations on chromosomes (Figure 6C, 41.5% versus 37.8%, KS test, $P = 0.2501$). Thus, the expression of Pack-MULEs is not as influenced by the macro-environment of the chromosome as seen with protein-coding genes. The less biased expression of Pack-MULEs implies that they can create newly expressed sequences in regions with few transcripts.

To provide a higher resolution of the impact of local genomic region on Pack-MULE expression, we used the presence of protein-coding genes within 1 kb flanking regions of Pack-MULEs as a proxy of chromatin state. Similar fractions (41.6% versus 39.1%) of Pack-MULEs with and without nearby protein-coding genes were expressed (χ^2 test, $P = 0.17$). Analysis of DHSs of Pack-MULEs located within pericentromeric regions revealed a higher DHS density in the 0.2 kb flanking sequences than the genomic average or average level in pericentromeric regions (0.69 vs. 0.16 or 0.07, KS test, $P < 0.0001$), suggesting that Pack-MULEs inserted in relatively 'open' sites even in the overall 'closed' pericentromeric regions. Alternatively, the insertion of Pack-MULEs may have converted a 'closed' pericentromeric region into an 'open' pericentromeric region. Collectively, these data suggest that the genome landscape has limited impact on Pack-MULE expression potential and Pack-MULEs are likely to have an inherent capacity to be expressed and are not as dependent on the local genomic region to drive expression as are protein-coding genes.

To test whether Pack-MULEs with different ages are distributed similarly, we compared the distribution of elements with different insertion times. As shown in Figure 6D, the density of old Pack-MULEs (>3 MY) was only slightly higher on euchromatic chromosomal arms than that in percentromeric regions. In contrast, the elements inserted within 0.5–3 MY were preferentially located in the euchromatic chromosomal arms, suggesting that Pack-MULEs tend to insert into euchromatic chromosomal arms but are more likely to be retained in pericentromeric regions. It is worth of mentioning that the two groups of Pack-MULEs have similar composition in terms of TIR families (Supplementary Table S7), thus the differential distribution is not an artifact of distinct target selection specificity among different TIR families. To test whether the elevated expression frequency of the elements (0.5–3 MY) is associated with their preferential insertions on euchromatic chromosomal arms, a comparison was made between elements only on euchromatic chromosomal arms and the results were very similar to that with total elements (Supplementary Figure S4F), suggesting that the insertion time does influence the expression of Pack-MULEs.

The insertion of Pack-MULEs causes sequence flow in different chromosomal domains and increases the local GC content

Duplicated Pack-MULE sequences from their parental genes mostly insert into unlinked loci (64). In Nipponbare, nearly all (>96%) of the Pack-MULE parental genes are located in high recombination regions, which is consistent with the results from a previous study (22). In contrast, Pack-MULE density in pericentromeric regions is only slightly lower than that in other regions (6 versus 8 Pack-MULEs per Mb sequence). In addition, pericentromeric regions are associated with more older Pack-MULEs (Figure 6D), suggesting Pack-MULEs may have been retained longer or less likely to be excluded in these regions. As a result, Pack-MULEs have been appending sequences from high recombination rate regions to low recombination rate regions, but not vice versa.

In maize, the *Mutator* element preferentially inserts into GC-rich sequences (65). To test whether rice Pack-MULEs have the same preference, we calculated the GC content of flanking sequences of Pack-MULEs. When MULEs insert into genome, they duplicate a small piece (8–11 bp) of flanking sequence, the target site duplication (TSD). As shown in Supplementary Figure S5, the sequences immediately flanking Pack-MULEs, largely representing the TSD, is very GC-poor, yet the flanking sequences beyond TSD have a GC content very close to the genomic average. As Pack-MULEs preferentially duplicate GC-rich sequences (16), the insertion of Pack-MULEs causes a dramatic transition in local GC content (Supplementary Figure S5). Comparison of the GC content of Pack-MULEs and their 2 kb flanking sequences (1 kbp on each side) showed that in most cases (71%) the GC content of Pack-MULEs is over 10% higher than the flanking sequences (e.g. 60% versus 50%). This number is higher for Pack-MULEs in pericentromeric regions than those in other regions (77% versus 70%, χ^2 test, $P < 0.00001$) as the flanking sequences of Pack-MULEs in pericentromeric regions are slightly less GC-rich than those in other regions (41% versus 43%, median value). As a consequence, the insertion of Pack-MULEs likely results in a more significant increase of local GC content in regions with low recombination rate.

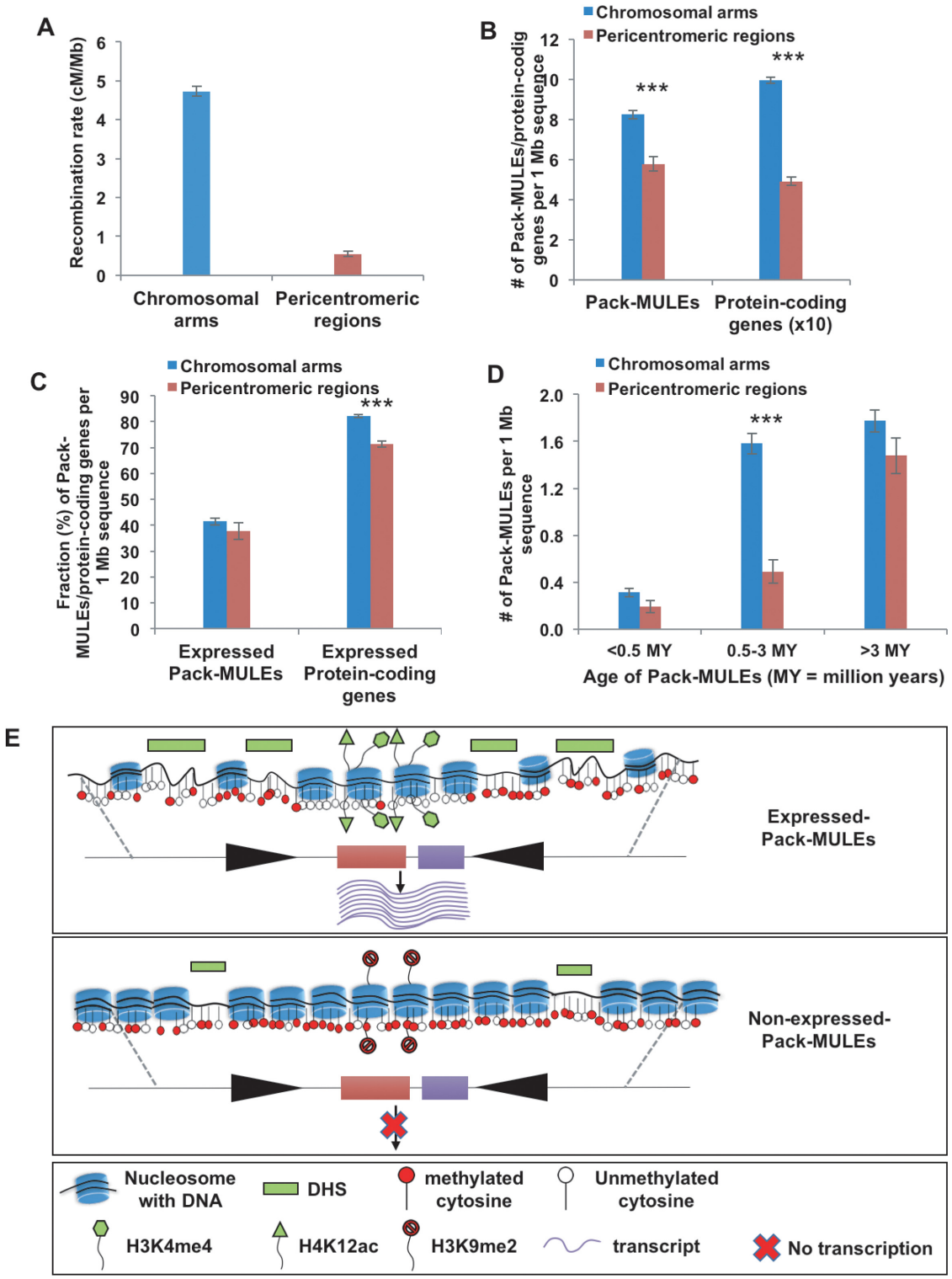

Figure 6. The relationship between recombination rate and distribution of Pack-MULEs and protein-coding genes. (**A**) The average recombination rates of euchromatic chromosomal arms and pericentromeric regions of all 12 chromosomes. (**B**) The distribution of Pack-MULEs and protein-coding genes on euchromatic chromosomal arms and pericentromeric regions. (**C**) The fraction of expressed Pack-MULEs and protein-coding genes on euchromatic chromosomal arms and pericentromeric regions. (**D**) The distribution of Pack-MULEs with different insertion times on euchromatic chromosomal arms and pericentromeric regions. (**E**) A diagram showing the epigenetic and expression status of Pack-MULEs.

DISCUSSION

Pack-MULEs represent a heterogeneous population with distinct epigenetic profiles

Gene acquisition by TEs (transduplication) is associated with the potential to generate new genes and increase the diversity of the gene pool by capturing and rearranging gene/gene fragments from existing genomic sequences. Although nearly all TE families can acquire gene fragments, the frequency varies dramatically (66). Moreover, the evolutionary impact and fate of transduplicated gene sequences remains largely unknown. Pack-MULEs are distinguished from the majority of other TEs in that they represent a mixture of *bona fide* TE sequences and non-TE sequences. The true TE sequences include terminal and sub-terminal sequences, which are often highly repetitive, whereas the non-TE sequences include gene fragments and their regulatory sequences, which are often low copy number sequences. Given the mixed structure, an intriguing question is whether Pack-MULEs represent an intermediate between TEs and protein-coding genes at the epigenetic level and as a consequence, have transcriptional and translational profiles associated with repressed vs. permissive genic regions.

Using a suite of large-scale datasets, we demonstrate a wide spectrum of variation in terms of expression, epigenetic, and chromatin status among rice Pack-MULEs (Figure 6E). For simplicity, we analyzed those with unambiguous expression evidence. Expressed Pack-MULEs (n = 1189, ∼40% of the total Pack-MULEs) have low DNA methylation, active histone marks, and abundant DHSs whereas non-expressed Pack-MULEs (n = 1070, 36% of the total) have a high level of DNA methylation, repressive histone marks, and a low density of DHSs. This integrated study, using transcriptomic, translatomic and epigenomic datasets, revealed that a majority of expressed Pack-MULEs possess features of *bona fide* protein-coding genes and thus may have the potential to be functional at the biological level.

The majority of Pack-MULEs were formed in the past six million years

In this study, we used two distinct approaches to determine the origin of Pack-MULEs. The transversion rate between Pack-MULEs and their parental genes refers to the initial time when the duplication event occurred. The transversion rate could be influenced by functional constraints on the duplicated copies as well as the epigenetic state of the element. We also estimated the insertion time of the Pack-MULEs with the assumption that polymorphic insertions were formed after the divergence of the two relevant genomes, which is not always true. Moreover, not all insertion loci are conserved in all genomes, and most genome assemblies are incomplete, thus, only a portion of the elements can be evaluated by this method. As a result, the two approaches evaluate different features, and both are associated with advantages and disadvantages. Regardless, the positive correlation between the two methods, i.e. older insertions are associated with higher transversion rate (Supplementary Table S6), suggests the credibility of the approaches.

To date, it is unclear whether the gene duplication process is associated with the transposition process. Since some Pack-MULEs are associated with multiple copies (15), it is not certain that every transposition event is coupled with a new duplication event. This implies that a new insertion or a new element could carry an old duplication through inheritance from its ancestral copy or alternatively, the duplication event occurred between two Pack-MULEs. This explains why the transversion rate is not perfectly proportional to the insertion time.

The fact that *O. punctata* shares few Pack-MULEs with Nipponbare indicates that the majority of the ∼3000 Pack-MULEs were formed after the divergence of AA and BB genomes, which was about 6 MY, suggesting that transposable elements could rapidly shape host genomes. This does not imply that there was little Pack-MULE activity prior to the divergence of the AA and BB genomes, it only indicates that Pack-MULEs would become unrecognizable after 6 MY. It is apparent that the activity of Pack-MULEs continued as the AA genomes diverged into different species, with about half of them formed in last 3 MY. Interestingly, the elements with the highest expression potential are those that inserted within 0.5 to 3 MY with expression enriched in pistil and seeds (Supplementary Figures S4D and E). In animals, new genes are initially expressed in male reproductive tissues such as testis, and an 'out of testis' hypothesis was proposed for the emergence of new genes (67). If expression is the first step for Pack-MULEs to evolve a function, it should be 'out of pistil and seeds'.

Open chromatin favors insertion and sequence acquisition by Pack-MULEs

Previous studies indicate that MULEs and Pack-MULEs preferentially insert at the 5′ end of protein-coding genes (65,68) with the highest density within 500 bp upstream of transcription start sites (64). In this study, we demonstrate that this is the region with the highest DHS density (Figure 3B). Even for Pack-MULEs located within pericentromeric regions, the DHS density in their flanking sequences is much higher than genomic average, suggesting that open chromatin favors the targeted insertion bias of MULEs. Alternatively, the insertion of Pack-MULEs may enable the surrounding sequences to become more relaxed. With respect to acquisition, Pack-MULE parental genes are flanked with more DHSs than other protein-coding genes (Figure 3B), are associated with higher GC content, have a wider expression breadth, and have lower levels of methylation (16,22). It remains unresolved whether these characteristics (expression, GC content, and methylation) directly influence sequence acquisition by Pack-MULEs (e.g. the acquisition machine interacts with transcription machinery) or as these features are associated with open chromatin they may simply provide a physical environment that is permissive for acquisition.

Although open chromatin favors both insertion and acquisition of Pack-MULEs, it appears that the acquisition process is more selective. This is revealed by the fact that nearly all Pack-MULE parental genes are located in regions with high recombination rates whereas Pack-MULE density is only slightly higher in these regions (see Results). This

seems to imply that transposition and acquisition specificity are controlled by different machineries. Alternatively, Pack-MULEs in regions with low recombination rates are less likely to be excluded, which is supported by the observation that they are in general older than their counterparts in other regions. This could be due to the more effective elimination of non-essential sequences in high recombination regions, or there is a selective advantage for the retention of Pack-MULEs in low recombination regions. The two possibilities are not mutually exclusive.

Expression of Pack-MULEs is largely independent of its genomic environment

Our analyses clearly indicate that Pack-MULE expression is correlated with epigenetic state, especially chromatin accessibility. However, little is known about the initial 'trigger' to induce the active or repressive state. For protein-coding genes, expression is often influenced by genomic context. For example, in *A. thaliana*, paralogous or duplicated genes demonstrate significant differential expression depending on their genomic location (69). In this study, we also detected a positive correlation between expression of protein-coding genes and local recombination rate (Figure 6A–C). Given these facts it is surprising to observe that the expression of Pack-MULEs is largely independent of either recombination rate or distance to protein-coding genes, suggesting that Pack-MULE expression potential is not determined by the local genomic environment.

An alternative explanation for the lack of correlation between expression status and genomic environment of Pack-MULEs is the target specificity of insertion in which Pack-MULEs are inserted within open chromatin regardless of the overall local genomic region, providing the fundamental requirement for transcription. This is consistent with the observation that for all Pack-MULEs, the DHS density in the flanking sequence is dramatically higher than the genomic average. Alternatively, Pack-MULEs themselves could promote the relaxation of local chromatin even within relatively 'closed' chromatic regions thereby creating a favorable environment for its expression. This hypothesis is supported by the higher than genomic average DHS density in the immediate flanking sequences of Pack-MULEs located in pericentromeric regions and by the location-specific enrichment of DHSs for protein-coding genes with adjacent Pack-MULEs at the 5′ and 3′ flanking sequences compared to those without adjacent Pack-MULEs (Figure 3C). However, we cannot rule out the possibility that an unusual distribution of DHSs existed prior to the insertion of the Pack-MULEs.

The insertion and expression of Pack-MULEs result in redistribution of GC-rich sequences on chromosomes

As shown by this and other studies, true transposon sequences such as TIRs only remain recognizable for a few million years (70) and likewise, an expressed Pack-MULE is largely indistinguishable from other genic sequences a few million years after their formation. This implies that the current recognizable Pack-MULEs may only represent a small subset of such elements that have been generated in the genome, and Pack-MULEs have duplicated and transposed large amount of genomic sequences over the course of evolution. Thus, we cannot rule out the possibility that some of the 'normal' GC-rich genes are ancient Pack-MULEs.

The grass (Poaceae) genomes including that of rice, are distinguished from dicot genomes (such as that of Arabidopsis) by the presence of GC-rich genes (71,72). It has been proposed that GC-biased gene conversion (gBGC) is primarily responsible for the evolution of GC content (73,74), a process that results in gene alleles with G/C at heterozygous sites being preferentially retained during recombination compared to alleles with A/T at heterozygous sites leading to high GC content at regions with high recombination rates. In *Gallus gallus*, for example, the variation of GC content is largely dependent on recombination rate (correlation coefficient = 0.89) (75); however, no significant correlation between GC content and recombination rate was observed in rice (37,38,75).

The correlation between GC content and recombination rate requires: (i) there is no significant mobility of sequences among chromosomal domains with distinct recombination rates and (ii) the local recombination rate remains stable for an extended time. Clearly, Pack-MULEs are capable of duplicating GC-rich sequences from regions with high recombination rates and inserting into those with low recombination rates. Furthermore, when Pack-MULE insertion alters the local GC content as well as the expression status, it is possible that the local recombination rate would change as well, thus further complicating the relationship between GC content and recombination. As a result, it is conceivable that the exceptional abundance of Pack-MULEs in rice may have, at least partially, masked the effect of gBGC.

The alteration of local GC content and expression status may further influence the future evolution of relevant regions. GC-rich sequences tend to augment bendability and the ability to undergo a B-Z transition of DNA helical structure, which is often associated with open chromatin (76). This is consistent with the fact that Pack-MULEs are associated with a high density of DHSs. Different types of transposons have distinct target selection in terms of chromatin structure and GC content. For example, the *Dasheng* and *RIRE2* retrotransposons in rice preferentially insert into the condensed heterochromatic regions with AT-rich sequences (77,78), and the insertion of more retrotransposons often causes further expansion of heterochromatin. In this case, an expressed Pack-MULE in an otherwise silenced chromosomal domain may influence the further insertion of transposons and mitigate the expansion of heterochromatin.

The sequence of Pack-MULEs determines their expression status

While chromatin state is critical for expression, regulatory sequences that drive transcription seem to originate from the Pack-MULE itself. Given the fact that Pack-MULEs carry gene fragments, it is intuitive that some of them may carry regulatory regions from normal protein-coding genes, and that these regulatory sequences direct the transcription of Pack-MULEs. Indeed, a small subset of Pack-MULEs (~6%) contain recognizable gene regulatory sequences and these elements are slightly enriched among elements with

expression evidence (Supplementary Table S4), suggesting that the regulatory sequences can contribute to the expression of Pack-MULEs but are not essential for expression of all Pack-MULEs.

Multiple lines of evidence suggest that the TIR region of Pack-MULEs may have an important role in expression. First, a previous study showed that TIRs of *MuDR* harbor promoters for transcription (54), and the majority of transcription start sites are either within TIRs or close to TIRs. Second, Pack-MULEs with different TIRs demonstrate differential expression frequency. Third, only a limited number of Pack-MULEs carry recognizable regulatory sequences from genes, suggesting that the majority of Pack-MULEs are not reliant on acquisition of regulatory sequences for their expression. Fourth, repressive histone marks are absent from Pack-MULE TIRs (Figure 5). Lastly, there is a distinct difference between expressed and non-expressed Pack-MULEs in terms of DHS density, i.e. DHSs are nearly absent inside non-expressed Pack-MULEs while expressed Pack-MULEs have a higher frequency of DHSs than the genome average (Figure 3B). The higher DHS density in the TIR relative to the internal region of Pack-MULEs (Figure 3B) is reminiscent of promoter and gene body regions of protein-coding genes, although protein-coding genes are only associated with high DHS density at the 5′ end whereas Pack-MULEs are associated with high DHS density at both 5′ and 3′ ends, consistent with the observation that some Pack-MULEs are expressed bi-directionally (15). If TIRs serve as promoters, it explains why the tissue specificity between Pack-MULEs and their parental genes is not correlated. On the other hand, the alteration of expression specificity conferred by Pack-MULEs would favor the generation of novel function.

Differential methylation of Pack-MULE internal regions and TIRs

A characteristic feature of expressed Pack-MULEs is the high DNA methylation of the TIRs, which immediately flank internal sequences that are lowly methylated. For non-expressed Pack-MULEs, both TIR and internal regions are associated with high methylation in the CG and CHG contexts. This is in contrast to CHH methylation of TIRs, with which expressed Pack-MULEs contained much higher CHH methylation than non-expressed Pack-MULEs. This is unexpected given the fact that non-CG methylation is linked to KRYPTONITE-dependent H3K9me2 in *A. thaliana* (79) which causes transcription suppression. However, neither high H3K9me2 deposition nor suppression of expression was observed for the Pack-MULEs with TIRs highly methylated in the CHH context. Thus, either the epigenetic regulation is distinct in grasses compared to that in dicots or the TIR (on average ∼200 bp) is too small to recruit the suppression machinery. As a result, high CHH methylation itself may not cause silencing unless it is coupled with suppressive marks and closed chromatin. For TEs in genic regions in maize, Gent *et al.* found that TEs close to cellular genes are highly methylated in the CHH context and that this *de novo* DNA methylation is a strategy for the genome to avoid transcription of TEs, yet maintain active transcription of cellular genes close to those TEs (19,56).

In *A. thaliana*, TEs are separated into two classes based on their epigenetic status (80). One class is present in large constitutive heterochromatic regions and their CHH methylation is maintained by Chromomethylase 2 (CMT2) while the other class is located near genes where CHH methylation is constantly targeted via RNA-directed DNA methylation (RdDM). Given such distinction, one explanation for the high CHH methylation of TIRs of expressed Pack-MULEs is that those TIRs are recognized as genic 'TEs' and are subjected to reinforced CHH methylation through RdDM, since TIRs are associated with a high abundance of siRNAs (81). The constant targeting by RdDM leads to increased methylation levels. In this scenario, the internal regions of expressed Pack-MULEs are sensed as active genes, and the high CHH methylation of their TIRs is the consequence of expression, not the reason for expression. This is consistent with the recent finding that transcriptional activation triggers additional RdDM mechanisms (82). This also explains why the proximity of Pack-MULEs to protein-coding genes is correlated with the CHH methylation in the TIR but not significantly with the expression of Pack-MULEs. In contrast, non-expressed Pack-MULEs are recognized as TEs in heterochromatic regions as the CHH methylation of TIR and internal region are not dramatically different. Alternatively, the transition of CHH methylation (from high to low) in the expressed Pack-MULEs may actually favor the expression of downstream sequences, consistent with the finding that highly expressed genes are enriched with CHH methylation islands upstream of promoters (83). Future experiments are required to assess the relationship between CHH methylation and expression.

Taken together, we have proposed a model for the evolutionary trajectory of Pack-MULEs. Due to their target specificity, most Pack-MULEs insert into the regions with open chromatin, despite the distance to protein-coding genes. However, Pack-MULEs in pericentromeric regions are more likely to be retained and accumulate, thus with more impact on GC content in those regions in the evolutionary scale. Not all newly formed Pack-MULEs have the same potential to be expressed. For example, young Pack-MULEs with high similarity among individual copies and high similarity to their parental genes lead to more abundant siRNAs (15,81) and thus increased silencing. Moreover, TIR family and internal sequence, also influence the expression of the elements. As such, Pack-MULEs likely evolve in two directions, with some harboring low DNA methylation, active histone marks, and open chromatin and as a consequence are often expressed; the other group having high DNA methylation, repressive histone marks, and 'closed' chromatin, and thereby silenced. As time elapses and Pack-MULEs enter 'middle age', the initial silencing of some Pack-MULEs is released and more Pack-MULEs are expressed (Supplementary Figures S4B and C). Such release predominates in reproductive tissues likely because TEs undergo developmental relaxation of silencing in these tissues (84). Elements that confer increased fitness are selected for and remain active. With increased age, Pack-MULEs that are not beneficial gradually lose their expression potential and are silenced again. Thus, the 'middle-aged' Pack-MULEs are associated with higher expression frequency than either very young or very old Pack-MULEs (Supple-

mentary Figures S4B and C). After a few million years, Pack-MULEs are no longer recognizable—they appear as normal genes, or become pseudogenes and eventually lost from the genome. From this perspective, Pack-MULEs represent natural 'cloning labs' that not only provide raw materials for natural and artificial selection, but also influence the overall genome structure and evolution.

Concluding remarks. In this study, we demonstrate that Pack-MULEs represent a heterogeneous population, and the evolutionary trajectory of individual Pack-MULEs is determined by their epigenetic status. Expressed Pack-MULEs largely resemble protein-coding genes in terms of methylation, histone mark and openness of chromatin, suggesting their functional role *in vivo*. However, there are differences between Pack-MULEs and regular protein-coding genes. Pack-MULEs are preferentially transcribed and translated in reproductive tissues, and it would be intriguing to test whether this is associated with selection of seed traits in rice. Expressed Pack-MULEs are associated with high CHH methylation levels (higher than any other genomic components) within the TIRs, which harbor regulatory sequences for Pack-MULE expression. The negative correlation between expression and CHH methylation contradicts the dogma of suppressive role of CHH methylation on expression suggesting alternative mechanisms may be at play. Unlike protein coding genes, the expression potential of Pack-MULEs is not dependent on the local genomic environment, suggesting that Pack-MULEs either preferentially insert into open, expression-permissive regions of the genome or alternatively, create expression competence following insertion. As Pack-MULEs duplicate GC-rich genes from their parental genes, which are mostly located in regions with high recombination rate, insertion of Pack-MULEs into regions with low recombination rate, results in an elevation of the local GC content thereby influencing the distribution of GC-rich sequences along the chromosome and a breakdown in the correlation between recombination rate and GC-content. As a result, the insertion and expression of Pack-MULEs alter chromosomal base composition and expression patterns, and as a consequence, the future evolution of chromosome structure.

DATA AVAILABILITY

Raw reads from BS-seq and ChIP-seq are available in the NCBI Sequence Read Archive under BioProject Number PRJNA386513.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Dr Martin Groth (University of California, Los Angeles) for valuable discussions.

FUNDING

National Science Foundation [MCB-1121650 to Y.C., J.J., C.R.B., N.J.]; United States Department of Agriculture National Institute of Food and Agriculture and AgBioResearch at Michigan State University [Hatch grant MICL02408 to N.J.]. Funding for open access charge: Michigan State University.
Conflict of interest statement. None declared.

REFERENCES

1. Kidwell,M.G. (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, **115**, 49–63.
2. Kapitonov,V.V. and Jurka,J. (2005) RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. *PLoS Biol.*, **3**, e181.
3. Hudson,M.E., Lisch,D.R. and Quail,P.H. (2003) The *FHY3* and *FAR1* genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. *Plant J.*, **34**, 453–471.
4. Lin,R. and Wang,H. (2004) Arabidopsis *FHY3/FAR1* gene family and distinct roles of its members in light control of Arabidopsis development. *Plant Physiol.*, **136**, 4010–4022.
5. Lin,R., Ding,L., Casola,C., Ripoll,D.R., Feschotte,C. and Wang,H. (2007) Transposase-derived transcription factors regulate light signaling in Arabidopsis. *Science*, **318**, 1302–1305.
6. Bundock,P. and Hooykaas,P. (2005) An Arabidopsis *hAT*-like transposase is essential for plant development. *Nature*, **436**, 282–284.
7. Lockton,S. and Gaut,B.S. (2009) The contribution of transposable elements to expressed coding sequence in *Arabidopsis thaliana*. *J. Mol. Evol.*, **68**, 80–89.
8. Sorek,R., Ast,G. and Graur,D. (2002) Alu-containing exons are alternatively spliced. *Genome Res.*, **12**, 1060–1067.
9. Jiang,N., Bao,Z., Zhang,X., Eddy,S.R. and Wessler,S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569–573.
10. Zabala,G. and Vodkin,L.O. (2005) The *wp* mutation of *Glycine max* carries a gene-fragment-rich transposon of the *CACTA* superfamily. *Plant Cell*, **17**, 2619–2632.
11. Daron,J., Glover,N., Pingault,L., Theil,S., Jamilloux,V., Paux,E., Barbe,V., Mangenot,S., Alberti,A., Wincker,P. *et al.* (2014) Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biol.*, **15**, 546.
12. Yang,L. and Bennetzen,J.L. (2009) Distribution, diversity, evolution, and survival of Helitrons in the maize genome. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 19922–19927.
13. Wang,W., Zheng,H., Fan,C., Li,J., Shi,J., Cai,Z., Zhang,G., Liu,D., Zhang,J., Vang,S. *et al.* (2006) High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell*, **18**, 1791–1802.
14. Barbaglia,A.M., Klusman,K.M., Higgins,J., Shaw,J.R., Hannah,L.C. and Lal,S.K. (2012) Gene capture by Helitron transposons reshuffles the transcriptome of maize. *Genetics*, **190**, 965–975.
15. Hanada,K., Vallejo,V., Nobuta,K., Slotkin,R.K., Lisch,D., Meyers,B.C., Shiu,S.-H. and Jiang,N. (2009) The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile. *Plant Cell*, **21**, 25–38.
16. Ferguson,A.A., Zhao,D.Y. and Jiang,N. (2013) Selective acquisition and retention of genomic sequences by Pack-*Mutator*-like elements based on Guanine-Cytosine content and the breadth of expression. *Plant Physiol.*, **163**, 1419–1432.
17. West,P.T., Li,Q., Ji,L., Eichten,S.R., Song,J., Vaughn,M.W., Schmitz,R.J. and Springer,N.M. (2014) Genomic distribution of H3K9me2 and DNA methylation in a maize genome. *PLoS One*, **9**, 1–10.
18. Zhang,X. (2008) The epigenetic landscape of plants. *Science*, **320**, 489–492.
19. Hollister,J.D. and Gaut,B.S. (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.*, **19**, 1419–1428.
20. Du,Z., Li,H., Wei,Q., Zhao,X., Wang,C., Zhu,Q., Yi,X., Xu,W., Liu,X.S., Jin,W. *et al.* (2013) Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in *Oryza sativa* L.*J. Mol. Plant*, **6**, 1463–1472.
21. Zhang,X., Yazaki,J., Sundaresan,A., Cokus,S., Chan,S.W.L., Chen,H., Henderson,I.R., Shinn,P., Pellegrini,M., Jacobsen,S.E. *et al.*

(2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, **126**, 1189–1201.
22. Wang,J., Yu,Y., Tao,F., Zhang,J., Copetti,D., Kudrna,D., Talag,J., Lee,S., Wing,R.A. and Fan,C. (2016) DNA methylation changes facilitated evolution of genes derived from *Mutator*-like transposable elements. *Genome Biol.*, **17**, 92.
23. Feng,S. and Jacobsen,S.E. (2011) Epigenetic modifications in plants: An evolutionary perspective. *Curr. Opin. Plant Biol.*, **14**, 179–186.
24. Kawahara,Y., de la Bastide,M., Hamilton,J.P., Kanamori,H., McCombie,W.R., Ouyang,S., Schwartz,D.C., Tanaka,T., Wu,J., Zhou,S. *et al.* (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.
25. Zhao,D. and Jiang,N. (2014) Nested insertions and accumulation of indels are negatively correlated with abundance of *Mutator*-like transposable elements in maize and rice. *PLoS One*, **9**, e87069.
26. Ou,S. and Jiang,N. (2018) LTR_retriever: a highly accurate and sensitive program for identification of long terminal-repeat retrotransposons. *Plant Physiol.*, **176**, doi:10.1104/pp.17.01310.
27. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
28. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
29. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Cufflinks manual. *Nat. Biotechnol.*, **28**, 511–515.
30. Wu,Y., Zhang,W. and Jiang,J. (2014) Genome-wide nucleosome positioning is orchestrated by genomic regions associated with DNase I hypersensitivity in rice. *PLoS Genet.*, **10**, e1004378.
31. Counce,P.A., Keisling,T.C. and Mitchell,A.J. (2000) A uniform, objective, and adaptive system for expressing rice development. *Crop Sci.*, **40**, 436–443.
32. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
33. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
34. Zhang,W., Wu,Y., Schnable,J.C., Zeng,Z., Freeling,M., Crawford,G.E. and Jiang,J. (2012) High-resolution mapping of open chromatin in the rice genome. *Genome Res.*, **22**, 151–162.
35. Langmead,B. (2010) Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinforma*, **11**, doi:10.1002/0471250953.bi1107s32.
36. Zang,C., Schones,D.E., Zeng,C., Cui,K., Zhao,K. and Peng,W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.
37. Zhao,D., Hamilton,J.P., Hardigan,M., Yin,D., He,T., Vaillancourt,B., Reynoso,M., Pauluzzi,G., Funkhouser,S., Cui,Y. *et al.* (2017) Analysis of ribosome-associated mRNAs in rice reveals the importance of transcript size and GC content in translation. *G3 (Bethesda).*, **7**, 203–219.
38. Tian,Z., Rizzon,C., Du,J., Zhu,L., Bennetzen,J.L., Jackson,S.A., Gaut,B.S. and Ma,J. (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.*, **19**, 2221–2230.
39. Wu,Y., Kikuchi,S., Yan,H., Zhang,W., Rosenbaum,H., Iniguez,A.L. and Jiang,J. (2011) Euchromatic subdomains in rice centromeres are associated with genes and transcription. *Plant Cell*, **23**, 4054–4064.
40. Wang,L., Czedik-Eysenberg,A., Mertz,R.A., Si,Y., Tohge,T., Nunes-Nesi,A., Arrivault,S., Dedow,L.K., Bryant,D.W., Zhou,W. *et al.* (2014) Comparative analyses of C4 and C3 photosynthesis in developing leaves of maize and rice. *Nat. Biotechnol.*, **32**, 1158–1165.
41. Watanabe,K.A., Ringler,P., Gu,L. and Shen,Q.J. (2014) RNA-sequencing reveals previously unannotated protein- and microRNA-coding genes expressed in aleurone cells of rice seeds. *Genomics*, **103**, 122–134.
42. Anderson,S.N., Johnson,C.S., Jones,D.S., Conrad,L.J., Gou,X., Russell,S.D. and Sundaresan,V. (2013) Transcriptomes of isolated *Oryza sativa* gametes characterized by deep sequencing: Evidence for distinct sex-dependent chromatin and epigenetic states before fertilization. *Plant J.*, **76**, 729–741.
43. Wilkins,K.E., Booher,N.J., Wang,L. and Bogdanove,A.J. (2015) TAL effectors and activation of predicted host targets distinguish Asian from African strains of the rice pathogen Xanthomonas oryzae pv. oryzicola while strict conservation suggests universal importance of five TAL effectors. *Front. Plant Sci.*, **6**, 536.
44. Davidson,R.M., Gowda,M., Moghe,G., Lin,H., Vaillancourt,B., Shiu,S.-H., Jiang,N. and Robin Buell,C. (2012) Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J.*, **71**, 492–502.
45. Sakai,H., Mizuno,H., Kawahara,Y., Wakimoto,H., Ikawa,H., Kawahigashi,H., Kanamori,H., Matsumoto,T., Itoh,T. and Gaut,B.S. (2011) Retrogenes in rice (Oryza sativa L. ssp. japonica) exhibit correlated expression with their source genes. *Genome Biol. Evol.*, **3**, 1357–1368.
46. Chitteti,B.R. and Peng,Z. (2007) Proteome and phosphoproteome differential expression under salinity stress in rice (*Oryza sativa*) roots. *J. Proteome Res.*, **6**, 1718–1727.
47. Koller,A., Washburn,M.P., Lange,B.M., Andon,N.L., Deciu,C., Haynes,P.A., Hays,L., Schieltz,D., Ulaszek,R., Wei,J. *et al.* (2002) Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11969–11974.
48. Tan,F., Li,G., Chitteti,B.R. and Peng,Z. (2007) Proteome and phosphoproteome analysis of chromatin associated proteins in rice (*Oryza sativa*). *Proteomics*, **7**, 4511–4527.
49. Helmy,M., Sugiyama,N., Tomita,M. and Ishihama,Y. (2012) The Rice Proteogenomics Database OryzaPG-DB: Development, expansion, and new features. *Front. Plant Sci.*, **3**, 65.
50. Lee,D.-G., Ahsan,N., Lee,S.-H., Kang,K.Y., Lee,J.J. and Lee,B.-H. (2007) An approach to identify cold-induced low-abundant proteins in rice leaf. *C. R. Biol.*, **330**, 215–225.
51. Li,G., Nallamilli,B.R.R., Tan,F. and Peng,Z. (2008) Removal of high-abundance proteins for nuclear subproteome studies in rice (Oryza sativa) endosperm. *Electrophoresis*, **29**, 604–617.
52. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
53. Marand,A.P., Zhang,T., Zhu,B. and Jiang,J. (2017) Towards genome-wide prediction and characterization of enhancers in plants. *Biochim. Biophys. Acta - Gene Regul. Mech.*, **1860**, 131–139.
54. Raizada,M.N., Benito,M.-I. and Walbot,V. (2008) The *MuDR* transposon terminal inverted repeat contains a complex plant promoter directing distinct somatic and germinal programs. *Plant J.*, **25**, 79–91.
55. Eichten,S.R., Ellis,N.A., Makarevitch,I., Yeh,C.-T., Gent,J.I., Guo,L., McGinnis,K.M., Zhang,X., Schnable,P.S., Vaughn,M.W. *et al.* (2012) Spreading of heterochromatin is limited to specific families of maize retrotransposons. *PLoS Genet.*, **8**, e1003127.
56. Gent,J.I., Ellis,N.A., Guo,L., Harkess,A.E., Yao,Y., Zhang,X. and Dawe,R.K. (2013) CHH islands: De novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res.*, **23**, 628–637.
57. Lennartsson,A. and Ekwall,K. (2009) Histone modification patterns and epigenetic codes. *Biochim. Biophys. Acta - Gen. Subj.*, **1790**, 863–868.
58. Zhang,J., Chen,L.-L., Xing,F., Kudrna,D.A., Yao,W., Copetti,D., Mu,T., Li,W., Song,J.-M., Xie,W. *et al.* (2016) Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E5163–5171.
59. Du,H., Yu,Y., Ma,Y., Gao,Q., Cao,Y., Chen,Z., Ma,B., Qi,M., Li,Y., Zhao,X. *et al.* (2017) Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.*, **8**, 15324.
60. Ma,J. and Bennetzen,J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 12404–12410.
61. Brozynska,M., Copetti,D., Furtado,A., Wing,R.A., Crayn,D., Fox,G., Ishikawa,R. and Henry,R.J. (2017) Sequencing of Australian wild rice genomes reveals ancestral relationships with domesticated rice. *Plant Biotechnol. J.*, **15**, 765–774.
62. Ge,S., Sang,T., Lu,B.R. and Hong,D.Y. (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 14400–14405.
63. Ammiraju,J.S.S., Lu,F., Sanyal,A., Yu,Y., Song,X., Jiang,N., Pontaroli,A.C., Rambo,T., Currie,J., Collura,K. *et al.* (2008) Dynamic evolution of oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell*, **20**, 3191–3209.

64. Jiang,N., Ferguson,A.A., Slotkin,R.K. and Lisch,D. (2011) Pack-*Mutator*-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 1537–1542.
65. Liu,S., Yeh,C.-T., Ji,T., Ying,K., Wu,H., Tang,H.M., Fu,Y., Nettleton,D. and Schnable,P.S. (2009) *Mu* transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet.*, **5**, e1000733.
66. Zhao,D., Ferguson,A.A. and Jiang,N. (2016) What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochim. Biophys. Acta - Gene Regul. Mech.*, **1859**, 366–380.
67. Kaessmann,H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.*, **20**, 1313–1326.
68. Dietrich,C.R., Cui,F., Packila,M.L., Li,J., Ashlock,D.A., Nikolau,B.J. and Schnable,P.S. (2002) Maize *Mu* transposons are targeted to the 5′ untranslated region of the *gl8* gene and sequences flanking *Mu* target-site duplications exhibit nonrandom nucleotide composition throughout the genome. *Genetics*, **160**, 697–716.
69. Panchy,N., Lehti-Shiu,M. and Shiu,S.-H. (2016) Evolution of gene duplication in plants. *Plant Physiol.*, **171**, 2294–2316.
70. Bennetzen,J.L., Ma,J. and Devos,K.M. (2005) Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.*, **95**, 127–132.
71. Carels,N. and Bernardi,G. (2000) Two classes of genes in plants. *Genetics*, **154**, 1819–1825.
72. Wong,G.K.-S., Wang,J., Tao,L., Tan,J., Zhang,J., Passey,D.A. and Yu,J. (2002) Compositional gradients in Gramineae genes. *Genome Res.*, **12**, 851–856.
73. Clement,Y., Fustier,M.-A., Nabholz,B. and Glemin,S. (2015) The bimodal distribution of genic GC content is ancestral to monocot species. *Genome Biol. Evol.*, **7**, 336–348.
74. Glémin,S., Arndt,P.F., Messer,P.W., Petrov,D., Galtier,N. and Duret,L. (2015) Quantification of GC-biased gene conversion in the human genome. *Genome Res.*, **25**, 1215–1228.
75. Pessia,E., Popa,A., Mousset,S., Rezvoy,C., Duret,L. and Marais,G.A.B. (2012) Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol. Evol.*, **4**, 675–682.
76. Vinogradov,A.E. (2003) DNA helix: the importance of being GC-rich. *Nucleic Acids Res.*, **31**, 1838–1844.
77. Jiang,N., Bao,Z., Temnykh,S., Cheng,Z., Jiang,J., Wing,R.A., McCouch,S.R. and Wessler,S.R. (2002) Dasheng: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. *Genetics*, **161**, 1293–1305.
78. Jiang,N., Jordan,I.K. and Wessler,S.R. (2002) *Dasheng* and *RIRE2*. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. *Plant Physiol.*, **130**, 1697–1705.
79. Johnson,L.M., Bostick,M., Zhang,X., Kraft,E., Henderson,I., Callis,J. and Jacobsen,S.E. (2007) The SRA methyl-cytosine-binding domain links DNA and histone methylation. *Curr. Biol.*, **17**, 379–384.
80. Sigman,M.J. and Slotkin,R.K. (2016) The first rule of plant transposable element silencing: location, location, location. *Plant Cell*, **28**, 304–313.
81. Nobuta,K., Venu,R.C., Lu,C., Beló,A., Vemaraju,K., Kulkarni,K., Wang,W., Pillay,M., Green,P.J., Wang,G. *et al.* (2007) An expression atlas of rice mRNAs and small RNAs. *Nat. Biotechnol.*, **25**, 473–477.
82. Panda,K., Ji,L., Neumann,D.A., Daron,J., Schmitz,R.J. and Slotkin,R.K. (2016) Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. *Genome Biol.*, **17**, 170.
83. Li,Q., Gent,J.I., Zynda,G., Song,J., Makarevitch,I., Hirsch,C.D., Hirsch,C.N., Dawe,R.K., Madzima,T.F., McGinnis,K.M. *et al.* (2015) RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 14728–14733.
84. Martínez,G. and Keith Slotkin,R. (2012) Developmental relaxation of transposable element silencing in plants: functional or byproduct? *Curr. Opin. Plant Biol.*, **15**, 496–502.