



**HAL**  
open science

## Genome-wide characterization of i-motifs and their potential roles in the stability and evolution of transposable elements in rice

Xing Ma, Yilong Feng, Ying Yang, Xin Li, Yining Shi, Shentong Tao, Xuejiao Cheng, Jian Huang, Xiu-E Wang, Caiyan Chen, et al.

► **To cite this version:**

Xing Ma, Yilong Feng, Ying Yang, Xin Li, Yining Shi, et al.. Genome-wide characterization of i-motifs and their potential roles in the stability and evolution of transposable elements in rice. *Nucleic Acids Research*, 2022, 50 (6), pp.3226-3238. 10.1093/nar/gkac121 . hal-03586421

**HAL Id: hal-03586421**

**<https://hal.science/hal-03586421>**

Submitted on 23 Feb 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# **Genome-wide characterization of i-motifs and their potential roles in the stability and evolution of transposable elements in rice**

Xing Ma<sup>1,#</sup>, Yilong Feng<sup>1,#</sup>, Ying Yang<sup>1</sup>, Xin Li<sup>2</sup>, Yining Shi<sup>1</sup>, Shentong Tao<sup>1</sup>, Xuejiao Cheng<sup>1</sup>, Jian Huang<sup>3</sup>, Xiu-e Wang<sup>1</sup>, Caiyan Chen<sup>2</sup>, David Monchaud<sup>4</sup>, Wenli Zhang<sup>1\*</sup>

<sup>1</sup>State Key Laboratory for Crop Genetics and Germplasm Enhancement, Collaborative Innovation Center for Modern Crop Production co-sponsored by Province and Ministry (CIC-MCP), Nanjing Agricultural University, No.1 Weigang, Nanjing, Jiangsu 210095, P. R. China. <sup>2</sup>Institute of Subtropical Agriculture, Chinese Academy of Sciences, Changsha, Hunan, 410125, P. R. China. <sup>3</sup>School of Biology & Basic Medical Science, Soochow University, Suzhou, Jiangsu 215123, P. R. China. <sup>4</sup>Institut de Chimie Moléculaire, ICMUB CNRS UMR 6302, UBFC Dijon, France.

**Running Title: i-motifs characterization and their impacts on TE dynamics in rice**

# These authors contributed equally to this work

\*Corresponding authors Zhang, W.

E-mail: [wzhang25@njau.edu.cn](mailto:wzhang25@njau.edu.cn)

ORCID:0000-0003-0710-1966

## **ABSTRACT**

I-motifs (iMs) are non-canonical DNA secondary structures that fold from cytosine (C)-rich genomic DNA regions termed putative i-motif forming sequences (PiMFSs). The structure of iMs is stabilized by hemiprotonated C–C base pairs and their functions are now suspected in key cellular processes in human cells such as genome stability and regulation of gene transcription. In plants, their biological relevance is still largely unknown. Here, we characterized PiMFSs with high potential for i-motif formation in the rice genome by developing and applying a protocol hinging on an iMab antibody-based immunoprecipitation (IP) coupled with high throughput sequencing (seq), consequently termed iM-IP-seq. We found that PiMFSs had intrinsic subgenomic distributions, *cis*-regulatory functions and an intricate relationship with DNA methylation. We indeed found that the coordination of PiMFSs with DNA methylation may affect dynamics of transposable elements (TEs) among different cultivated *Oryza* subpopulations or during evolution of wild rice species. Collectively, our study provides first and unique insights into the biology of iMs in plants, with potential applications in plant biotechnology for improving important agronomic rice traits.

**KEY WORDS:** i-motif, PiMFSs, iM-IP-seq, DNA methylation, TE dynamics and evolution, rice

## INTRODUCTION

I-motifs (iMs) are non-canonical DNA quadruplexes held by hemiprotonated cytosine–cytosine (C) base pairs. iMs fold from C-rich genomic DNA regions harboring putative i-motif forming sequences (PiMFSs) under slightly acidic environment (1,2). Hemiprotonated C-C base pairs were first identified in 1962 (3) and proposed for the stabilization of hairpins (4). The structure of the first intercalated iM tetramers was elucidated by NMR in 1993 (1) and crystallographic studies in 1994 (5). Because of the acidic condition requirement, iMs were long considered as structural oddities, as their folding was supposed not to be compatible with physiological conditions (pH ~7.4). Their biological relevance was therefore largely overlooked, and the pace of progress towards its elucidation was very slow as compared to other secondary DNA structures, notably G-quadruplexes (G4s) that fold from guanine (G)-rich sequences (6). This is, however, of utmost importance as genomic G4- and iM-forming sequences are, by definition, complementary.

Therefore, specific conditions have been discovered to favor or facilitate iM formation at physiological pH, such as DNA supercoiling (7), molecular crowding (8,9) and the presence of small molecules (ligands) (10,11), reinvigorated research on iMs (12). Recent immunodetection studies performed with the iM-specific iMab antibody have demonstrated that iM folding is cell cycle-dependent in HeLa cells, pH-dependent in MCF7 cells and occurs preferentially at the G1/S phase in U2OS cells (13). The demonstration of the relevance of iMs in a normal cellular context thus provides a key turning point in iM research.

iMs were found to be markedly overrepresented in promoters, telomeres and centromeres, which is indicative of their critical cellular roles (6,14), including maintenance of genome stability (15) and regulation of gene expression (16-24). In humans, iMs can be found in several oncogenes such as *c-MYC*, *c-MYB*, *BCL-2*, *ILPR*, *Rb*, *RET* and *VEGF* (18,25-29). iMs have also been involved in the regulation of the gene transcription during the metamorphosis of *Bombyx mori* (23). iMs frequently localize at centromeres of both humans and *Drosophila* (30-33), and at telomeres of both humans and *Arabidopsis* (34,35).

These results substantiate the putative key roles that the iMs might play in human cells (36) or other eukaryotic genomes, but much less is known about their possible roles in plants so far. Fragmentary

information has been gathered over the years through low throughput-screening techniques such as circular dichroism (CD) (37), nuclear magnetic resonance (NMR) (38), native PAGE (34) and chemical ( $\text{Br}_2$ ) footprinting (27). Unlike G4s, high throughput techniques, such as sequencing-based approaches, -related characterization of iMs are still missing in both mammals and plants. Here, we designed and performed iMab-based immunoprecipitation followed by sequencing (iM-IP-seq) to identify PiMFSs with high potential for i-motif formation in the rice genome. We showed the relationships between DNA methylation and iM formation, and focused on the potential impacts of iMs on TEs (transposable elements) stability and its evolution across *Oryza* populations.

## **MATERIALS AND METHODS**

### **Plant materials**

Nipponbare (Japonica) rice seeds (*Oryza. sativa* L.) were pre-germinated at room temperature (RT) for three days. Germinated seeds were grown in a greenhouse at 28-30°C and a 14h/10h light-dark cycle. Two-week-old rice seedlings were cut into 1-1.5 cm slices and cross-linked in HEPES buffer pH=8.0 (20mM HEPES, 1mM EDTA, 100mM NaCl and 1mM PMSF) with a final 1% of formaldehyde under vacuum for 10 min at RT. A final concentration of 0.125 M glycine was added for an additional 5min vacuum for quenching excess of formaldehyde. The cross-linked seedlings were ground to a fine powder using liquid nitrogen.

### **iM-IP-seq**

The fine powder was used for nuclei preparation and genomic DNA extraction following the procedures as described previously (39). The genomic DNA extracted from wild-type or Zebularine treated seedlings was fragmented into sizes, ranging from 100-500 bp, using the water-based Biorupter (Diagnode), followed by DNA extraction and purification. Total 5 $\mu\text{g}$  fragmented genomic DNA was diluted in iM stabilization buffer (50mM Tris-AcOH, pH=5.5) with (a crowding condition for mimicking the cellular and/or nucleus condition) or without (CK) 40% PEG200, then was denatured and re-associated using a PCR program as below: 95°C for 8 min, 95°C for 30s (-0.5 °C /cycles, 129 cycles), 35°C hold on. The re-associated DNA was diluted with iM-IP incubation buffer (50mM Tris-

AcOH, 1mM MgCl<sub>2</sub>, 130nM CaCl<sub>2</sub>, 1%BSA and Complete mini, pH=5.5), then incubated with 3μg iMab antibody (Ab01462-23.0, Kappa) for 4 h at 4°C. The antibody incubation reaction was incubated with 30 μl of washed protein G Dynalbeads (10004D, Invitrogen) for another 4 h at 4°C. The iMab-bound DNA was finally eluted with 200 μl elution buffer (0.1M NaHCO<sub>3</sub> and 1% SDS) at 65°C for two times with 15 min each. Three biologically replicated PiMFS-IPed CK DNA and two biologically replicated PiMFS-IPed demethylated DNA, two biologically replicated PiMFS-IPed DNA in a crowding condition, and the corresponding Input/IgG-IPed control DNA were used for library preparation. All libraries were prepared using the NEBNext<sup>®</sup> Ultra<sup>™</sup> II DNA Library Prep Kit (NEB, E7645S). Libraries were finally quality controlled and sequenced using the paired-end mode on Illumina NovaSeq platform.

### **Protoplast transient transfection**

Preparation of a modified pJIT163-hGFP vector containing PiMFSs and protoplast transient transfection assay were conducted following the procedures as previously described (40). Briefly, PiMFS DNA fragments were amplified using PCR (**Supplementary Table S4**) and cloned into a modified pJIT163-hGFP vector. The pJIT163-hGFP vector containing mini35S promoter only (negative control), an intact 35S promoter (positive control) and a mini 35S promoter + amplified PiMFS DNA fragment were transfected into protoplasts mediated by PEG. GFP signals were captured using fluorescent microscopy.

### **iM-IP-qPCR assay**

For iM-IP-qPCR assay, 1μl of input and IPed DNA (2ng/ μl each) was used as DNA templates. The enrichment of IPed DNA was calculated using the  $2^{(\Delta\Delta Ct)}$  method and expressed as fold change over the corresponding input. Each primer set was repeated three times in each qPCR. All primer sequences are listed in Supplementary Table S4.xlsx. Significance test was performed using One-way ANOVA analysis. \*\*\*p < 0.001, \*\*p < 0.01 and \*p < 0.05.

## **Dot blotting assays**

For the dot blotting assay, genomic DNA, synthesized oligonucleotides and M.SssI-treated genomic DNA were denatured and re-associated in iM reconstruction buffer (50mM Tris-AcOH, pH=5.5) at 95°C for 8 min. Denatured and re-associated DNA or sonicated chromatin prepared from the purified nuclei was loaded on Amersham Hybond-N+-nylon membrane followed by pre-blocking in 5% milk for 45min at RT. The pre-blocked membrane was incubated with the iMab antibody overnight at 4°C, then with anti-IgG (HRP) antibody for an additional 1.5h. The procedures for immune-signal development were the same as described before (41). Each blot was repeated at least two times for quantification signal intensity.

## **Native Polyacrylamide Gel Electrophoresis**

Synthesized DNA oligonucleotides were denatured at 95 °C for approximately 8 min in iM reconstruction buffer (50mM Tris-AcOH, 30mM KCl, pH=5.5), then slowly cooled down to room temperature overnight for re-association, the counterpart for each sample with denaturation but without re-association was used as control. All samples were incubated at 4 °C for 10min before loading onto 15% polyacrylamide gel. The gels were run at a constant voltage of 85 V at RT for 1.5 h. Gels were stained using Stains GeneRed (Tiangen, RT211) solution and visualized under UV light.

## **CD measurements**

500 µl of 5µM PiMFS oligonucleotide in iM stabilization buffer (50mM Tris-AcOH, pH=5.5) was measured in an optical chamber (1mm path length) with a JASCO J-815 spectropolarimeter (Tokyo, Japan). Dry purified nitrogen gas was used to maintain a deoxygenation atmosphere. The solution background was subtracted from the CD signal.

## **iM-IP-seq data analysis**

Raw iM-IP-seq data were trimmed by using Fastp for removal of adapter sequence. All clean reads in three biological replicates were aligned to the MSU v7.0 reference genome([http://rice.plantbiology.msu.edu/pub/data/Eukaryotic\\_Projects/o\\_sativa/annotation\\_dbs/pseud](http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseud)

[omolecules/version\\_7.0/all.dir/](#)) using BWA (Burrows-Wheeler Aligner) (mem algorithm, version 0.7.17) with default parameters. SAMtools was used to exclude reads with mapping quality below 10. PCR duplicates were removed by using Picard. The Spearman rank correlation coefficient between replicates was calculated using multiBamSummary and plotCorrelation function of deepTools. Aligned reads with at least length 50 were used for calling PiMFS-IPs<sup>+</sup> peaks, which were identified by iM-IP-seq, by using MACS2 (version 2.1.1) (42) with parameters as below: callpeak -g 3.8e+8 -p 1e-4 -nomodel -f BAMPE. Input and IgG data were used as controls.

### ***In silico* identification of putative i-motif forming sequences**

Putative i-motif forming sequences (PiMFSs) were identified through screening the MSU v7.0 reference genome using G4iMGrinder from the R package (43). Parameters were tested for obtaining regular PiMFSs with sequences as [CC<sub>2-5</sub>L<sub>1-15</sub>]<sub>3-8</sub>C<sub>2-5</sub> followed by further selection of several subtypes of PiMFSs. To assess folding potential of each sequence, score of each individual putative sequence was calculated by considering the run size, bulges between runs and loop size. PiMFSs with score below -40 were chosen for the further analyses.

To distinguish PiMFSs with long and short loop sizes, the entire regular PiMFS sequences were computationally scanned using [fastaRegexFinder.py](https://github.com/dariober/bioinformaticscafe/blob/master/fastaRegexFinder.py) (44). Short loops contain sequences as (loop 1 (2-nt): loop 2 (3 to 4-nt): loop 3 (2-nt), i.e. 2:3-4:2), while long loops contain sequences as (loop1 (6-8):loop2 (2-5):loop 3 (6-7)).

### **Genomic distribution and calculation of GC content, GC/AT skew**

ChIPseeker from the R package (45) was used to investigate genomic distribution of PiMFS-IPs<sup>+</sup> and PiMFS-IPs<sup>-</sup>, representing PiMFSs detected and undetected by iM-IP-seq, respectively. For calculating GC content and GC/AT skew, ±1 kb regions around the center of PiMFS-IPs<sup>+/-</sup> were divided into 50 bp windows, then were calculated using the following formulas: GC content = (C+G)/(C+G+A+T); GC skew = (G-C)/(G+C) and AT skew = (A-T)/(A+T). The genomic regions lacking PiMFSs were randomly selected as controls by using bedtools shuffle.



### **Motif prediction**

$\pm 100$  bp DNA sequences around the center of PiMFS-IPs<sup>+</sup> peaks were used for motif identification using MEME-ChIP (<http://memesuite.org/tools/meme-chip>) (46) with parameters as options minimum width 5 and maximum width 20. TF database from *Arabidopsis* was used to match putative TF-binding sites (Tomtom tool) corresponding to all identified motifs. The motifs listed in the result represent the top three significantly enriched motifs with the highest E-values.

### **DNase-seq data analysis**

To integrate DNase-sequencing (DNase-seq) with iM-IP-seq, the published DNase-seq data (GSE26734) (39) were used for identification of DNase I hypersensitive sites (DHSs) using F-seq (47) with 200 bp bandwidth and the FDR (false discovery rate) cutoff  $<0.05$ . The FDR represents the ratio of DHSs from DNase-seq relative to DHSs identified from 10 random data sets.

### **BS-seq data analysis**

To examine the DNA methylation levels within PiMFS-IPs<sup>+</sup> regions, Bisulfite-sequencing (BS-seq) data (48) were analyzed by mapping clean data to the rice reference genome (MSU7.0) using Bismark (49). The methylated cytosines were counted from total uniquely mapped reads using the bismark methylation\_extractor program. The DNA methylation levels were calculated using the total number of all (C+T) greater than or equal to 5 in each position. The DNA methylation levels in different regions were calculated from 50 bp windows within each region.

### **Read count normalization**

The  $\pm 1$  kb upstream/downstream of center of PiMFS-IPs<sup>+/-</sup> were divided into 50 bp windows. The  $\pm 500$  bp upstream/downstream of TEs and TE regions were equally divided into 25 bins for normalization. The number of reads per sliding window was divided by the window length, then by the count of all uniquely mapped reads within the genome (Mb). The midpoint of the fragment was used to determine the position in rice genome.

## RESULTS

### Global identification of PiMFSs with high potential for i-motif formation in the rice genome

iMab is an antibody that binds iMs with a high affinity. It was obtained by screening the Garvan-2 human single-chain variable fragment (scFv) library (13). It was reported to specifically bind to iMs present in human telomere *in vitro*, termed hTelo iM, in human cells (13). The use of this antibody has greatly helped gain insights into the biological relevance of iMs in eukaryotes. However, iMab-based high throughput sequencing for characterization of iMs is still missing.

We decided to investigate this in the rice genome, using a protocol referred to as iM-IP-seq. We firstly validated the use of iMab by performing dot blotting assays: as seen in **Figure 1A**, clear dot signals were obtained using the synthetic MYC iM, the rice genomic DNA and sonicated chromatin, but almost nothing with the controls, *i.e.*, the synthetic MYC G4 (13), an AT-rich oligonucleotide and ddH<sub>2</sub>O (**Figure 1A**). This first series of results confirmed the efficiency of iMab as well as its specificity for iMs due to almost non-detectable cross-reactivity of iMab with G4s *in vitro*.

Next, we developed an iM-IP-seq protocol for the global identification of PiMFSs with high potential for i-motif formation, termed PiMFS-IPs<sup>+</sup>, representing the maximum possibility of i-motif formation for the rice genomic DNA. As seen in Figure 1B, the main procedures include genomic DNA purification and fragmentation, reconstruction of iM structures using iM-stabilizing conditions (50mM Tris-AcOH, pH=5.5), iMab incubation and recovery of iMab-captured DNA fragments for library preparation and sequencing. We sequenced three biological iM-IP-seq libraries along with one input and one IgG library as controls (**Supplementary Table S1**). As the three biological replicates were highly correlated ( $r = 0.98$ ) (**Supplementary Figure S1**), we merged them, which led to the identification of 45,851 and 32,897 PiMFS-IPs<sup>+</sup> peaks relative to input and IgG control, respectively, and 25,306 common PiMFS-IPs<sup>+</sup> peaks (**Figure 1C**). To mimic the cellular and/or nucleus condition, we reconstructed the i-motif structures in a crowding condition (iM stabilization buffer with 40% PEG200), then conducted iM-IP-seq with two biological replicates. After comparing PiMFS-IPs<sup>+</sup> formed in the crowding condition relative to CK by using MA<sub>norm</sub> (50), we identified 3,126 crowding condition biased, 4,153 CK biased and 31,550 unbiased PiMFS-IPs<sup>+</sup> peaks (**Figure 1D**). This result

indicated that a portion of iMs can be differentially formed under the normal cellular and/or nucleus condition, which is necessary to be further investigated. A representative Integrative Genomics Viewer (IGV) spanning a 22 kb region in the rice Chromosome 2 show the reproducibility of PiMFS-IPs<sup>+</sup> detected among the replicates in normal (blue) and a crowding (red) condition (**Figure 1E**). We finally determined to use 25,306 common PiMFS-IPs<sup>+</sup> peaks for the downstream analyses.

### Validation of PiMFS-IPs<sup>+</sup>

To further demonstrate that the detected C-rich sequences are iM-prone, we conducted *in silico* investigations with the 25,306 common PiMFS-IPs<sup>+</sup> peaks using G4iMGrinder (43). We searched for all PiMFSs bearing at least four runs of two consecutive Cs, with loop length between 1 to 15 nt (threshold score < -40). On a genome-wide scale, we identified 1,475,053 PiMFSs of formula  $[CC_{2-5}L_{1-15}]_3C_{2-5}$  (**Supplementary Table S2**): 99.9% of these PiMFSs were found to overlap putative G4-forming sequences (PQS) with runs of two consecutive Gs, confirming the genomic connection between G4s and iMs (**Figure 2A, right**). 88.6 % of them belonged to PiMFS-IPs<sup>+</sup> peaks (**Figure 2A left, 2B**), implying that the remaining 2,881 PiMFS-IPs<sup>+</sup> peaks actually correspond to non-regular PiMFSs, that is, with sequences variations. This was further investigated by lowering the threshold score to < -30, which led to the identification of 25,114 (*ca.* 99.4%) PiMFS-IPs<sup>+</sup> peaks overlapping all PiMFSs. It confirmed that the 2,881 candidates correspond to non-classical iMs: a closer look to the sequences confirmed occurrence of sequence variations as compared to the regular ones (**Supplementary Table S3**); they included iMs in which C-runs disrupted by at least one additional nucleotide (A, T, G) and iMs with four to nine C-runs ( $[CC_{2-5}L_{1-15}]_{4-9}C_{2-5}$ ). All PiMFSs were thus classified into two subtypes with PiMFS-IPs<sup>+</sup> ( $n = 180,608$ ), and PiMFS-IPs<sup>-</sup> ( $n = 1,294,445$ ), representing PiMFSs detected and undetected by iM-IP-seq, respectively (**Supplementary Figure S2**).

The iM-folding ability of randomly selected 6 PiMFS-IPs<sup>+</sup> peaks was confirmed by dot blotting assay (**Figure 2C, right panel**) and native polyacrylamide gel electrophoresis (PAGE), using C7 and C8 as positive controls, and a synthetic AT-rich oligonucleotide and ddH<sub>2</sub>O as negative controls (**Supplementary Table S4**). We observed 5 (*ca.* 83%) loci with detectable immune-signals (**Figure 2C, right panel**). For the native PAGE, the formation of iMs was promoted by a denaturation/renaturation

cycle (lanes 2), which leads to distinct DNA bands (yellow star, as compared to lane 1) that correspond to iMs formation (**Figure 2C, left panel**). Moreover, eight positive loci were also significantly enriched during iM-IP-qPCR experiments (**Figure 2D**). In addition, for CD (circular dichroism) spectroscopy assay, we observed that randomly selected other 6 oligonucleotides exhibited a clear absorbance peak increased at around 290 nm but decreased at around 260 nm, confirming their ability to adopt iM structures (**Figure 2E**).

Collectively, all above analyses show that those PiMFS-IPs<sup>+</sup> peaks are reliable for the downstream assays.

### **Genomic distribution and sequence features of PiMFS-IPs<sup>+</sup>**

It is now established that iMs or G4s are abundantly present in promoters and untranslated regions (UTRs) of human oncogenes and tumor suppressor genes (51,52). To assess genomic distributions of iMs in the rice genome, we partitioned the whole genome into 7 functionally annotated subregions, including promoters, 5'UTRs, 3'UTRs, exons, introns, downstream and distal intergenic regions. We observed that PiMFS-IPs<sup>+</sup> were enriched in some subgenomic regions (1.2-fold in promoters, 1.3-fold in 5'UTRs) but depleted in the others (*e.g.*, 0.8-fold in exons, 0.7-fold in distal intergenic regions) as compared to PiMFS-IPs<sup>-</sup> (**Figure 3A**). As shown in Figure 3B and 3C, we found no significant differences in the mean length of PiMFSs between PiMFS-IPs<sup>+</sup> and PiMFS-IPs<sup>-</sup>, while the distance between PiMFSs is longer for PiMFS-IPs<sup>+</sup> as compared to PiMFS-IPs<sup>-</sup>, indicating that the genomic density of PiMFSs may affect iM formation *in vitro* (**Figure 3B, 3C**). methylation represses iM formation, in contradiction with results seen in **Figure 6C**, a discrepancy that might originate in the extent to which the DNA is methylated. To confirm this, we used synthetic oligonucleotides without methylated C (C0), 1 methylated C (C1) and 3 consecutive methylated Cs (C3) and 5 methylated Cs (C5). As seen by CD spectroscopy, all oligonucleotides displayed a clear iM-typical CD signature, but we found that C3 and especially C5 exhibited a weaker CD fingerprint than C1, also indicating that a higher level of methylated Cs tends to disfavor iM formation (**Figure 6E**). We further conducted iM-IP-seq using globally demethylated genomic DNA induced by Zebularine treatment, a chemical functioning as an inhibitor of DNA methyltransferase, as we did before (68). As shown in

Supplementary Figure S11, after analyzing two biological replicates, we identified 5,841 demethylated DNA biased PiMFS-IPs<sup>+</sup> peaks with 26.5% of mean C content, 1,217 CK biased PiMFS-IPs<sup>+</sup> peaks and 67,339 unbiased PiMFS-IPs<sup>+</sup> peaks with 31.3% and 29.1% of mean C content, respectively (**Supplementary Figure S12**). After plotting the DNA methylation levels at each cytosine context around  $\pm 1$  kb of the summit of indicated PiMFS-IPs<sup>+</sup> peaks, we observed that demethylated DNA biased PiMFS-IPs<sup>+</sup> peaks had the highest CG and CHG methylation across all region examined, and the second highest CHH methylation at around  $\pm 250$  bp of the summit as compared to the CK and unbiased ones (**Figure 6F**). This result indicated that DNA demethylation may facilitate formation of highly methylated iMs. How DNA demethylation affect iM formation *in vivo* is necessary to be further investigated.

Thus, all above analyses indicate that DNA methylation exhibits a complex relationship with i-motif formation *in vitro*, may depend on methylation extent or combined actions of all cytosine methylations.

Finally, we decided to assess whether PiMFS-IPs<sup>+</sup> read density is related to changes in DNA methylation levels in the different TE families. To this end, we plotted DNA methylation levels of CG, CHG and CHH across  $\pm 500$  bp around the whole TE region, and found that common TEs with PiMF-IPs<sup>+/</sup> had higher DNA methylation levels in all cytosine contexts relative to the corresponding non-common ones (**Figure 6G**). We also found that conserved TEs with PiMF-IPs<sup>+/</sup> exhibited higher CG and CHG methylation but lower CHH methylation levels relative to the corresponding non-conserved ones (**Figure 6H**). An IGV showing DNA methylation distributed across a conserved/common TEs and non-conserved/non-common TEs is illustrated in Supplementary Figure S13.

Taken together, all above analyses show that DNA methylation levels are globally inversely correlated with iM formation, supporting the hypothesis that PiMFSs may coordinate with DNA methylation to control TE activity in the genome.

## DISCUSSION

Our study aimed at investigating the iM landscape on a genome-wide scale in rice. To this end, we developed and applied a novel protocol, termed iM-IP-seq (**Figure 1B**), which allowed for the

identification of rice genomic DNA capable of forming PiMFS-IPs<sup>+</sup> *in vitro*. The C-rich sequences prone to fold into iMs are substrates for DNA methylation. Our study showed that PiMFS-IPs<sup>+</sup> were globally hypomethylated as compared to sequences devoid of PiMFS-IPs<sup>+</sup>-forming capability (**Figure 6A, 6B**). However, while CHG and CHH methylation exhibited an inverse correlation with PiMFS-IPs<sup>+</sup> formation, the impact of CG methylation on PiMFS-IPs<sup>+</sup> folding was more paradoxical (being associated with both iM formation and destabilization) (**Figure 6C, 6D**). Our study thus revealed a complex relationship between DNA methylation and iM formation *in vitro*, which strongly depends on methylation level and position. Such a potential impact of methylation on the stability of telomeric iM was reported in *Arabidopsis* (34), in which a differential C-methylation frequency and a methylation position-dependent effect on iM stability were also detected (69,70). The mechanism behind these observations is not trivial: in line with the results presented here, it has been shown that one or two methylated cytosines could help stabilize telomeric iM structures, while hypermethylation results in iM destabilization in humans (67). Therefore, we can postulate that moderate levels of C-methylation could help increase C-C base-pair thermo-stability (71,72), while C-hypermethylation will increase more frequency of mC-mC base pairs, which could create steric hindrance, resulting in destabilization of C-C base pair. This was further evidenced in our study: iMs were overall much less enriched with repressive histone marks like H3K27me3 and H3K9me2/3 (**Figure 4D**), indicating that iMs are preferably present in euchromatic regions with less methylated DNA and repressive histone marks relative to heterochromatic regions; demethylated DNA biased PiMFS-IPs<sup>+</sup> exhibited higher DNA methylation levels in CK as compared to the CK biased or unbiased PiMFS-IPs<sup>+</sup> peaks (**Figure 6F, Supplementary Figure S12**). This indicates that DNA demethylation and/or the presence of euchromatic marks help(s) to open up chromatin, thereby creating more space to facilitate iM formation. Also, the opposite effects of 5mC and 5hmC on iM stability are indicative of their influences on DNA flexibility that affects iM folding, the former reducing it, the latter increasing it (73). Finally, these modifications may have distinct impacts on binding or recruitment of certain factors: as an example, 5hmC was found to reduce binding ability of proteins (nucleolin) and small-molecules (TMPyP4) to iMs (74); also, in line with what was described for G4s (75), iM formation may hamper access of DNA methyltransferase to DNA, which results in DNA hypomethylation, which in turn reduces the folding

ability of PiMFSs. Moreover, we cannot exclude the possibility that effects of mC on iM stability may be sequence context-dependent. Since effects of cytosine methylation or other modifications on iM stability *in vitro* were mainly examined in special DNA sequences in humans like telomeric DNA and (CCG)(n)(CCG)(n) trinucleotide repeats in humans (67,71,72,76). Massive efforts must now be invested to provide evidence for these different hypotheses. In particular, it is necessary to investigate the relationship between mC and iM stability on a genome-wide scale in humans or other species. Our study casts also a new light on the intricate relationship between PiMFS-IPs<sup>+</sup> and transposable elements (TEs). TEs, which represent 35% of the genome in rice (77), 85% in maize (78) and wheat (79), serve as a major driving force contributing to dynamics of genome size and structure, and gene/genome evolution (80). We showed here that PiMFSs were enriched in certain subtypes of TEs in rice, and that the amount of PiMFSs, rather than folding potential of PiMFSs, contributed to TE dynamics, notably modulating their methylation status. This wealth of data also highlights the possible roles of PiMFS-IPs<sup>+</sup> in TE evolution: we showed that non-common or non-conserved TEs were highly active in different rice species likely *via* PiMFS-IPs<sup>+</sup>-mediated DNA hypomethylation. Collectively, this study indicates that iMs alone and in combination with epigenetic regulations play active roles in TE life cycle, thereby functioning as drivers of genome evolution in rice.

## **DATA AVAILABILITY**

The iM-IP-seq and control data used in this study have been deposited in the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE184783.

## **SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

## **FUNDING**

This research was supported by grants from the National Natural Science Foundation of China (32070561 and U20A2030).

*Conflict of interest statement.* None declared.

## ACKNOWLEDGMENTS

We thank the Bioinformatics Center, Nanjing Agricultural University for providing computing facilities for data processing and analyses.

*Author contributions:* W.L.Z. conceived and designed the study. X.M. analyzed the data. Y.L.F. performed iM-DNA-IP, qPCR and dot blotting experiments. X.J.C. helped with material preparation. Y.N.S. helped with the genomic DNA preparation. Y. Y. prepared DNA plasmids for transient transformation. Y.L. and C.Y.C. assisted with rice protoplast transformation. X.E.W. supervised the experiments. S.T.T. helped with data analyses. J.H. assisted with CD experiment. X.M., Y.L.F., D.M. and W.L.Z. interpreted the results. D.M. and W.L.Z. wrote the manuscript with contributions from all other authors.

## REFERENCES

1. Gehring, K., Leroy, J.L. and Gueron, M. (1993) A tetrameric DNA structure with protonated cytosine-cytosine base pairs. *Nature*, **363**, 561-565.
2. Leroy, J.L., Gehring, K., Kettani, A. and Gueron, M. (1993) Acid multimers of oligodeoxycytidine strands: stoichiometry, base-pair characterization, and proton exchange properties. *Biochemistry*, **32**, 6019-6031.
3. Marsh, R.E., Bierstedt, R. and Eichhorn, E.L. (1962) The crystal structure of cytosineacetic acid. *Acta Cryst.*, **15**, 310.
4. Ahmed, S. and Henderson, E. (1992) Formation of novel hairpin structures by telomeric C-strand oligonucleotides. *Nucleic Acids Res.*, **20**, 507-511.
5. Kang, C.H., Berger, I., Lockshin, C., Ratliff, R., Moyzis, R. and Rich, A. (1994) Crystal structure of intercalated four-stranded d(C3T) at 1.4 Å resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 11636-11640.
6. Benabou, S., Aviñó, A., Eritja, R., González, C. and Gargallo, R. (2014) Fundamental aspects of the nucleic acid i-motif structures. *RSC Adv.*, **4**, 26956-26980.



7. Sun, D. and Hurley, L.H. (2009) The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression. *J. Med. Chem.*, **52**, 2863-2874.
8. Bhavsar-Jog, Y.P., Van Dornshuld, E., Brooks, T.A., Tschumper, G.S. and Wadkins, R.M. (2014) Epigenetic modification, dehydration, and molecular crowding effects on the thermodynamics of i-motif structure formation from C-rich DNA. *Biochemistry*, **53**, 1586-1594.
9. Cui, J., Waltman, P., Le, V.H. and Lewis, E.A. (2013) The effect of molecular crowding on the stability of human c-MYC promoter sequence I-motif at neutral pH. *Molecules*, **18**, 12751-12767.
10. Li, X., Peng, Y., Ren, J. and Qu, X. (2006) Carboxyl-modified single-walled carbon nanotubes selectively induce human telomeric i-motif formation. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 19658-19663.
11. Chen, X., Zhou, X., Han, T., Wu, J., Zhang, J. and Guo, S. (2013) Stabilization and induction of oligonucleotide i-motif structure via graphene quantum dots. *ACS Nano.*, **7**, 531-537.
12. Wright, E.P., Huppert, J.L. and Waller, Z.A.E. (2017) Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH. *Nucleic Acids Res.*, **45**, 2951-2959.
13. Zeraati, M., Langley, D.B., Schofield, P., Moye, A.L., Rouet, R., Hughes, W.E., Bryan, T.M., Dinger, M.E. and Christ, D. (2018) I-motif DNA structures are formed in the nuclei of human cells. *Nat. Chem.*, **10**, 631-637.
14. Abou Assi, H., Garavis, M., Gonzalez, C. and Damha, M.J. (2018) i-Motif DNA: structural features and significance to cell biology. *Nucleic Acids Res*, **46**, 8038-8056.
15. Bacolla, A. and Wells, R.D. (2004) Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.*, **279**, 47411-47414.
16. Kendrick, S. and Hurley, L.H. (2010) The role of G-quadruplex/i-motif secondary structures as cis-acting regulatory elements. *Pure Appl. Chem.*, **82**, 1609-1621.
17. Kang, H.J., Kendrick, S., Hecht, S.M. and Hurley, L.H. (2014) The transcriptional complex between the BCL2 i-motif and hnRNP LL is a molecular switch for control of gene expression that can be modulated by small molecules. *J. Am. Chem. Soc.*, **136**, 4172-4185.

18. Brooks, T.A., Kendrick, S. and Hurley, L. (2010) Making sense of G-quadruplex and i-motif functions in oncogene promoters. *FEBS J.*, **277**, 3459-3469.
19. Ruggiero, E., Lago, S., Sket, P., Nadai, M., Frasson, I., Plavec, J. and Richter, S.N. (2019) A dynamic i-motif with a duplex stem-loop in the long terminal repeat promoter of the HIV-1 proviral genome modulates viral transcription. *Nucleic Acids Res.*, **47**, 11057-11068.
20. Saha, P., Panda, D., Muller, D., Maity, A., Schwalbe, H. and Dash, J. (2020) In situ formation of transcriptional modulators using non-canonical DNA i-motifs. *Chem. Sci.*, **11**, 2058-2067.
21. Kaiser, C.E., Van Ert, N.A., Agrawal, P., Chawla, R., Yang, D. and Hurley, L.H. (2017) Insight into the Complexity of the i-Motif and G-Quadruplex DNA Structures Formed in the KRAS Promoter and Subsequent Drug-Induced Gene Repression. *J. Am. Chem. Soc.*, **139**, 8522-8536.
22. Kendrick, S., Kang, H.J., Alam, M.P., Madathil, M.M., Agrawal, P., Gokhale, V., Yang, D., Hecht, S.M. and Hurley, L.H. (2014) The dynamic character of the BCL2 promoter i-motif provides a mechanism for modulation of gene expression by compounds that bind selectively to the alternative DNA hairpin structure. *J. Am. Chem. Soc.*, **136**, 4161-4171.
23. Niu, K., Zhang, X., Deng, H., Wu, F., Ren, Y., Xiang, H., Zheng, S., Liu, L., Huang, L., Zeng, B. *et al.* (2018) BmILF and i-motif structure are involved in transcriptional regulation of BmPOUM2 in *Bombyx mori*. *Nucleic Acids Res.*, **46**, 1710-1723.
24. Dzatko, S., Krafcikova, M., Hansel-Hertsch, R., Fessler, T., Fiala, R., Loja, T., Krafcik, D., Mergny, J.L., Foldynova-Trantirkova, S. and Trantirek, L. (2018) Evaluation of the stability of DNA i-Motifs in the nuclei of living mammalian cells. *Angew. Chem. Int. Ed. Engl.*, **57**, 2165-2169.
25. Simonsson, T., Pribylova, M. and Vorlickova, M. (2000) A nuclease hypersensitive element in the human c-myc promoter adopts several distinct i-tetraplex structures. *Biochem. Biophys. Res. Commun.*, **278**, 158-166.
26. Xu, Y. and Sugiyama, H. (2006) Formation of the G-quadruplex and i-motif structures in retinoblastoma susceptibility genes (Rb). *Nucleic Acids Res.*, **34**, 949-954.
27. Guo, K., Gokhale, V., Hurley, L.H. and Sun, D. (2008) Intramolecularly folded G-quadruplex and i-motif structures in the proximal promoter of the vascular endothelial growth factor gene. *Nucleic Acids Res.*, **36**, 4598-4608.

28. Cogoi, S. and Xodo, L.E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res.*, **34**, 2536-2549.
29. Li, H., Hai, J., Zhou, J. and Yuan, G. (2016) The formation and characteristics of the i-motif structure within the promoter of the c-myc proto-oncogene. *J. Photochem. Photobiol. B*, **162**, 625-632.
30. Garavis, M., Mendez-Lago, M., Gabelica, V., Whitehead, S.L., Gonzalez, C. and Villasante, A. (2015) The structure of an endogenous Drosophila centromere reveals the prevalence of tandemly repeated sequences able to form i-motifs. *Sci. Rep.*, **5**, 13307.
31. Garavis, M., Escaja, N., Gabelica, V., Villasante, A. and Gonzalez, C. (2015) Centromeric Alpha-Satellite DNA Adopts Dimeric i-Motif Structures Capped by AT Hoogsteen Base Pairs. *Chemistry*, **21**, 9816-9824.
32. Gallego, J., Chou, S.H. and Reid, B.R. (1997) Centromeric pyrimidine strands fold into an intercalated motif by forming a double hairpin with a novel T:G:G:T tetrad: solution structure of the d(TCCCGTTTCCA) dimer. *J. Mol. Biol.*, **273**, 840-856.
33. Nonin-Lecomte, S. and Leroy, J.L. (2001) Structure of a C-rich strand fragment of the human centromeric satellite III: a pH-dependent intercalation topology. *J. Mol. Biol.*, **309**, 491-506.
34. Skolakova, P., Badri, Z., Foldynova-Trantirkova, S., Rynes, J., Sponer, J., Fojtova, M., Fajkus, J., Marek, R., Vorlickova, M., Mergny, J.L. *et al.* (2020) Composite 5-methylations of cytosines modulate i-motif stability in a sequence-specific manner: Implications for DNA nanotechnology and epigenetic regulation of plant telomeric DNA. *Biochim. Biophys. Acta Gen. Subj.*, **1864**, 129651.
35. Megalathan, A., Cox, B.D., Wilkerson, P.D., Kaur, A., Sapkota, K., Reiner, J.E. and Dhakal, S. (2019) Single-molecule analysis of i-motif within self-assembled DNA duplexes and nanocircles. *Nucleic Acids Res.*, **47**, 7199-7212.
36. Amato, J., Iaccarino, N., Randazzo, A., Novellino, E. and Pagano, B. (2014) Noncanonical DNA secondary structures as drug targets: the prospect of the i-motif. *ChemMedChem.*, **9**, 2026-2030.
37. Manzini, G., Yathindra, N. and Xodo, L.E. (1994) Evidence for intramolecularly folded i-DNA structures in biologically relevant CCC-repeat sequences. *Nucleic Acids Res.*, **22**, 4634-4640.

38. Dai, J., Ambrus, A., Hurley, L.H. and Yang, D. (2009) A direct and nondestructive approach to determine the folding structure of the I-motif DNA secondary structure by NMR. *J. Am. Chem. Soc.*, **131**, 6102-6104.
39. 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.
40. Fang, Y., Wang, X., Wang, L., Pan, X., Xiao, J., Wang, X.E., Wu, Y. and Zhang, W. (2016) Functional characterization of open chromatin in bidirectional promoters of rice. *Sci. Rep.*, **6**, 32088.
41. Fang, Y., Chen, L., Lin, K., Feng, Y., Zhang, P., Pan, X., Sanders, J., Wu, Y., Wang, X.E., Su, Z. *et al.* (2019) Characterization of functional relationships of R-loops with gene transcription and epigenetic modifications in rice. *Genome Res.*, **29**, 1287-1297.
42. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
43. Belmonte-Reche, E. and Morales, J.C. (2020) G4-iM Grinder: when size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool. *NAR. Genome Bioinform.*, **2**, lqz005.
44. Fujimoto, A., Fujita, M., Hasegawa, T., Wong, J.H., Maejima, K., Oku-Sasaki, A., Nakano, K., Shiraishi, Y., Miyano, S., Yamamoto, G. *et al.* (2020) Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res.*, **30**, 334-346.
45. Yu, G., Wang, L.G. and He, Q.Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382-2383.
46. Machanick, P. and Bailey, T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696-1697.
47. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311-322.

48. Hu, L., Li, N., Xu, C., Zhong, S., Lin, X., Yang, J., Zhou, T., Yuliang, A., Wu, Y., Chen, Y.R. *et al.* (2014) Mutation of a major CG methylase in rice causes genome-wide hypomethylation, dysregulated genome expression, and seedling lethality. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 10642-10647.
49. Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571-1572.
50. Shao, Z., Zhang, Y., Yuan, G.C., Orkin, S.H. and Waxman, D.J. (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.*, **13**, R16.
51. Verma, A., Halder, K., Halder, R., Yadav, V.K., Rawal, P., Thakur, R.K., Mohd, F., Sharma, A. and Chowdhury, S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. *J. Med. Chem.*, **51**, 5641-5649.
52. Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P. and Balasubramanian, S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877-881.
53. Kendrick, S., Akiyama, Y., Hecht, S.M. and Hurley, L.H. (2009) The i-motif in the bcl-2 P1 promoter forms an unexpectedly stable structure with a unique 8:5:7 loop folding pattern. *J. Am. Chem. Soc.*, **131**, 17667-17676.
54. Eddy, J. and Maizels, N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, **36**, 1321-1333.
55. Balasubramanian, S., Hurley, L.H. and Neidle, S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.*, **10**, 261-275.
56. Lu, L., Chen, X., Sanders, D., Qian, S. and Zhong, X. (2015) High-resolution mapping of H4K16 and H3K23 acetylation reveals conserved and unique distribution patterns in Arabidopsis and rice. *Epigenetics*, **10**, 1044-1053.
57. Tan, F., Zhou, C., Zhou, Q., Zhou, S., Yang, W., Zhao, Y., Li, G. and Zhou, D.X. (2016) Analysis of chromatin regulators reveals specific features of rice DNA methylation pathways. *Plant Physiol.*, **171**, 2041-2054.

58. He, G., Zhu, X., Elling, A.A., Chen, L., Wang, X., Guo, L., Liang, M., He, H., Zhang, H., Chen, F. *et al.* (2010) Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell*, **22**, 17-33.
59. Kejnovsky, E. and Lexa, M. (2014) Quadruplex-forming DNA sequences spread by retrotransposons may serve as genome regulators. *Mob. Genet. Elements*, **4**, e28084.
60. Kejnovsky, E., Tokan, V. and Lexa, M. (2015) Transposable elements and G-quadruplexes. *Chromosome Res.*, **23**, 615-623.
61. Tokan, V., Puterova, J., Lexa, M. and Kejnovsky, E. (2018) Quadruplex DNA in long terminal repeats in maize LTR retrotransposons inhibits the expression of a reporter gene in yeast. *BMC Genomics*, **19**, 184.
62. Le, N.T., Harukawa, Y., Miura, S., Boer, D., Kawabe, A. and Saze, H. (2020) Epigenetic regulation of spurious transcription initiation in Arabidopsis. *Nat. Commun.*, **11**, 3224.
63. Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X. *et al.* (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**, 3542-3558 e3516.
64. Zhao, Q. and Feng, Q. (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.*, **50**, 278-284.
65. Deniz, O., Frost, J.M. and Branco, M.R. (2019) Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.*, **20**, 417-431.
66. Brautigam, K. and Cronk, Q. (2018) DNA Methylation and the Evolution of Developmental Complexity in Plants. *Front. Plant Sci.*, **9**, 1447.
67. Xu, B., Devi, G. and Shao, F. (2015) Regulation of telomeric i-motif stability by 5-methylcytosine and 5-hydroxymethylcytosine modification. *Org. Biomol. Chem.*, **13**, 5646-5651.
68. Feng, Y., Tao, S., Zhang, P., Sperti, F.R., Liu, G., Cheng, X., Zhang, T., Yu, H., Wang, X.E., Chen, C. *et al.* (2021) Epigenomic features of DNA G-quadruplexes and their roles in regulating rice gene transcription. *Plant Physiol.*, DOI: 10.1093/plphys/kiab566.

69. Ogrocka, A., Polanska, P., Majerova, E., Janeba, Z., Fajkus, J. and Fojtova, M. (2014) Compromised telomere maintenance in hypomethylated Arabidopsis thaliana plants. *Nucleic Acids Res.*, **42**, 2919-2931.
70. Vrbsky, J., Akimcheva, S., Watson, J.M., Turner, T.L., Daxinger, L., Vyskot, B., Aufsatz, W. and Riha, K. (2010) siRNA-mediated methylation of Arabidopsis telomeres. *PLoS Genet.*, **6**, e1000986.
71. Yang, B., Wu, R.R. and Rodgers, M.T. (2013) Base-pairing energies of proton-bound homodimers determined by guided ion beam tandem mass spectrometry: application to cytosine and 5-substituted cytosines. *Anal. Chem.*, **85**, 11000-11006.
72. Yang, B. and Rodgers, M.T. (2014) Base-pairing energies of proton-bound heterodimers of cytosine and modified cytosines: implications for the stability of DNA i-motif conformations. *J. Am. Chem. Soc.*, **136**, 282-290.
73. Ngo, T.T., Yoo, J., Dai, Q., Zhang, Q., He, C., Aksimentiev, A. and Ha, T. (2016) Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.*, **7**, 10813.
74. Morgan, R.K., Molnar, M.M., Batra, H., Summerford, B., Wadkins, R.M. and Brooks, T.A. (2018) Effects of 5-Hydroxymethylcytosine Epigenetic Modification on the Stability and Molecular Recognition of VEGF i-Motif and G-Quadruplex Structures. *J. Nucleic Acids*, **2018**, 9281286.
75. Mao, S.Q., Ghanbarian, A.T., Spiegel, J., Martinez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hansel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2018) DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol.*, **25**, 951-957.
76. Yang, B., Moehlig, A.R., Frieler, C.E. and Rodgers, M.T. (2015) Base-pairing energies of protonated nucleobase pairs and proton affinities of 1-methylated cytosines: model systems for the effects of the sugar moiety on the stability of DNA i-motif conformations. *J. Phys. Chem. B.*, **119**, 1857-1868.
77. Takata, M., Kiyohara, A., Takasu, A., Kishima, Y., Ohtsubo, H. and Sano, Y. (2007) Rice transposable elements are characterized by various methylation environments in the genome. *BMC Genomics*, **8**, 469.

78. Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112-1115.
79. Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirez-Gonzalez, R.H., De Oliveira, R., International Wheat Genome Sequencing, C., Mayer, K.F.X., Paux, E. *et al.* (2018) Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.*, **19**, 103.
80. Bennetzen, J.L. and Wang, H. (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.*, **65**, 505-530.



## FIGURE

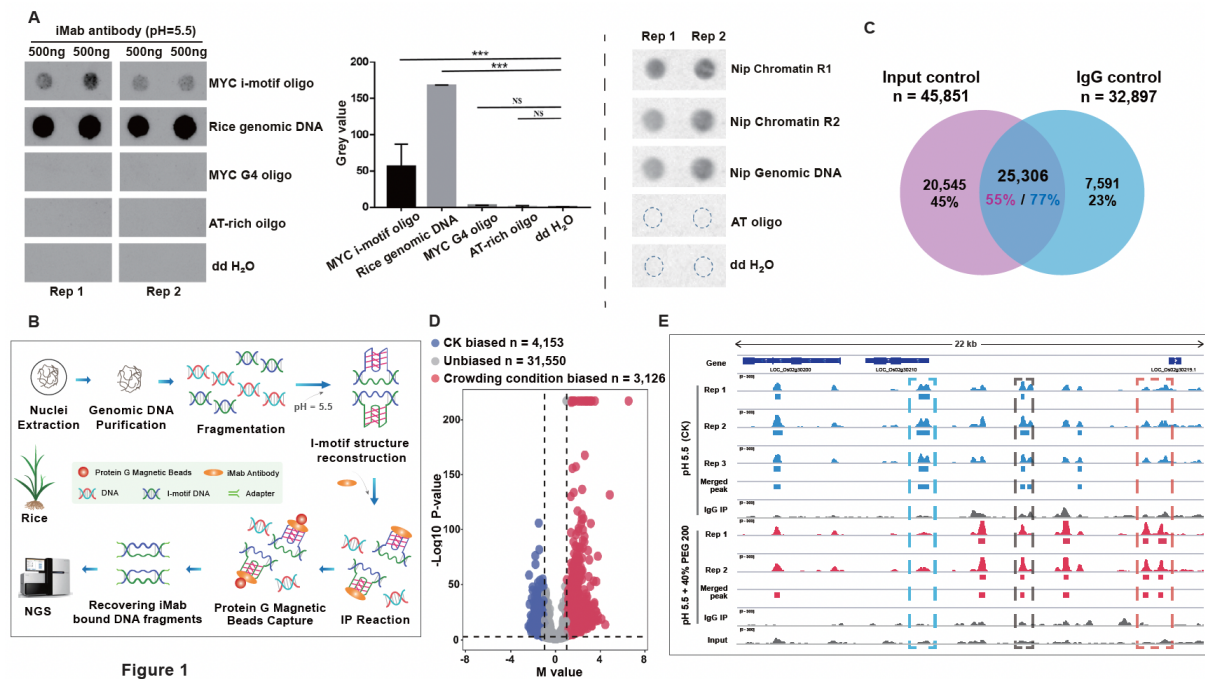


Figure 1

**Figure 1 Global identification of PiMFS-IPs<sup>+</sup> in the rice genome.** (A) Dot blotting assays illustrating immuno-signals clearly detected in the synthesized MYC i-motif oligonucleotide and the rice genomic DNA, but almost undetectable in the synthesized MYC G4 oligonucleotide, the synthesized AT-rich oligonucleotide and ddH<sub>2</sub>O (left). The signal intensity was quantified from at least two replicates (middle). Immuno-signals can also be clearly detected in the sonicated chromatin (right). (B) Schematic diagram illustrating the main procedures for iM-IP-seq methodology. (C) Venn plot illustrating 25,306 common PiMFS-IPs<sup>+</sup> peaks relative to the Input ( $n = 45,851$ ) and IgG ( $n = 32,897$ ) as controls. (D) Effects of the crowding condition on iM formation. MA plot illustrating the crowding condition biased PiMFS-IPs<sup>+</sup> peaks ( $n = 3,126$ ), CK biased PiMFS-IPs<sup>+</sup> peaks ( $n = 4,153$ ) and unbiased PiMFS-IPs<sup>+</sup> peaks ( $n = 31,550$ ) were identified using MANorm with the cutoff indicated by the dotted line ( $x = \pm 1$  and  $y = 3$ ),  $p\text{-value} < 0.001$  and  $|M\text{ value}| > 1$ .  $y$ -axis represents  $-\log_{10}(p\text{-value})$  and  $x$ -axis represents M value defined by MANorm. (E) A representative Integrative Genomics Viewer (IGV) snapshot across a 22 kb window from the rice chromosome 2 illustrating reproducibility of PiMFS-IPs<sup>+</sup> among biological replicates in normal (blue) and a crowding (red) condition. Blue, grey and red box indicating CK biased, unbiased and the crowding condition biased PiMFS-IPs<sup>+</sup> peaks, respectively.

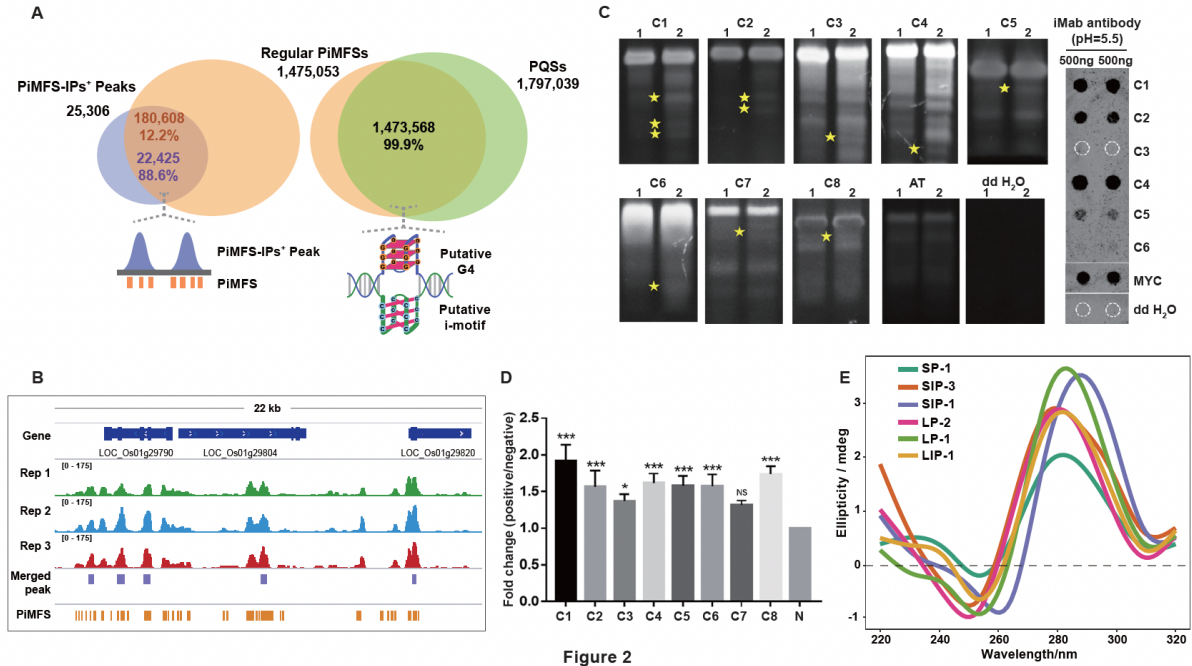


Figure 2

**Figure 2 Validation of PiMFS-IPs<sup>+</sup>.** (A) Venn plot illustrating PiMFS-IPs<sup>+</sup> peaks overlapping regular PiMFSs (left), and regular PiMFSs overlapping G2 PQSs (right). Schematic diagram right below the venn plot illustrating PiMFSs overlapping PiMFS-IPs<sup>+</sup> peaks (left) or the complementary G2 PQSs (right). (B) A representative Integrative Genomics Viewer (IGV) snapshot across a 22 kb window from the rice chromosome 1 illustrating PiMFS-IPs<sup>+</sup> peaks overlapping PiMFSs. (C) iMab based dot blotting (right) for C1-C6 loci and Native PAGE (left) assay for C1-C8 loci corresponding to PiMFS-IPs<sup>+</sup> peaks, synthesized MYC oligonucleotide and ddH<sub>2</sub>O only. Yellow star indicates distinct DNA bands formed in denaturing and re-associated DNA relative to denaturing DNA only, reflecting the formation of iMs after DNA re-association. (D) iM-IP-qPCR assay for 8 positive PiMFS-IPs<sup>+</sup> peaks loci (C1-C8) and 1 non- PiMFS-IPs<sup>+</sup> negative control (N). Fold change (P/N) indicates the enrichment levels between each positive locus over the negative locus. Significance test was determined by One-way ANOVA analysis, \*\*\* p-value < 0.001, \* p-value < 0.05, NS: non-significant. (E) CD confirmation of other 6 positive loci corresponding to PiMFS-IPs<sup>+</sup> peaks.

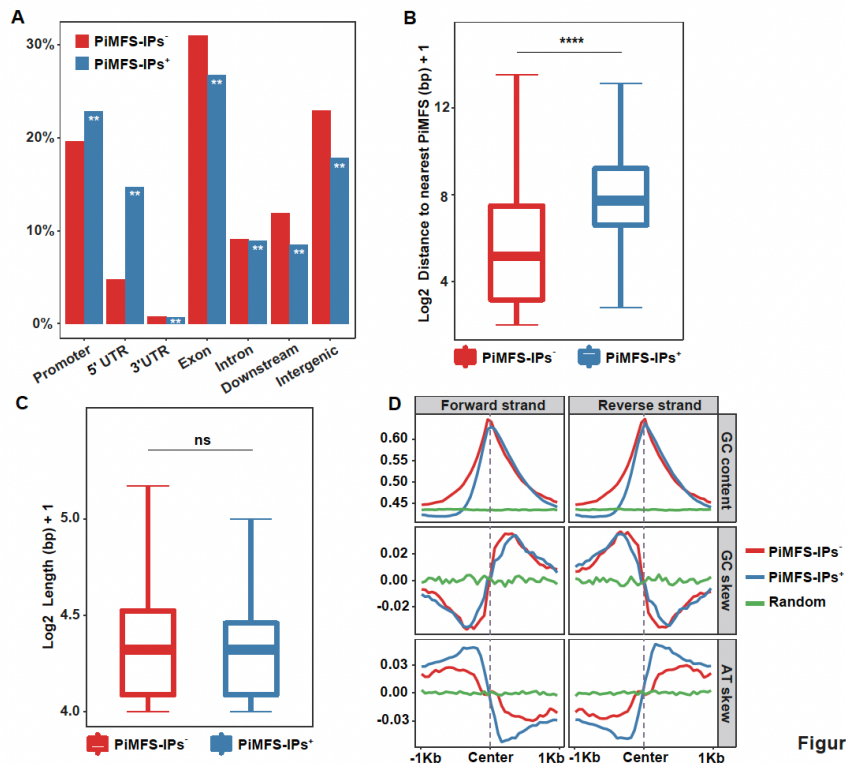


Figure 3

**Figure 3 Genomic distributions and sequence features of PiMFS-IPs<sup>+</sup>.** (A) Subgenomic distributions of PiMFS-IPs<sup>+/-</sup>, including promoters, 5'UTRs, 3'UTRs, exons, introns, downstream and intergenic regions. Significance test was determined by hypergeometric test. \*\* p-value < 0.01. (B) Distance from PiMFS-IPs<sup>+</sup> (left) or PiMFS-IPs<sup>-</sup> (right) to the nearest neighboring PiMFSs. (C) Mean length of PiMFS-IPs<sup>+</sup> (left) and PiMFS-IPs<sup>-</sup> (right). Significance tests in B and C were determined by Wilcoxon rank-sum test. \*\* p-value < 0.01, \*\*\*\* p-value < 0.0001, NS: non-significance. (D) Strand-specific GC content, GC and AT skews calculated around  $\pm$  1 kb from the center of PiMFS-IPs<sup>+</sup> (blue), PiMFS-IPs<sup>-</sup> (red) and random sequences (green).

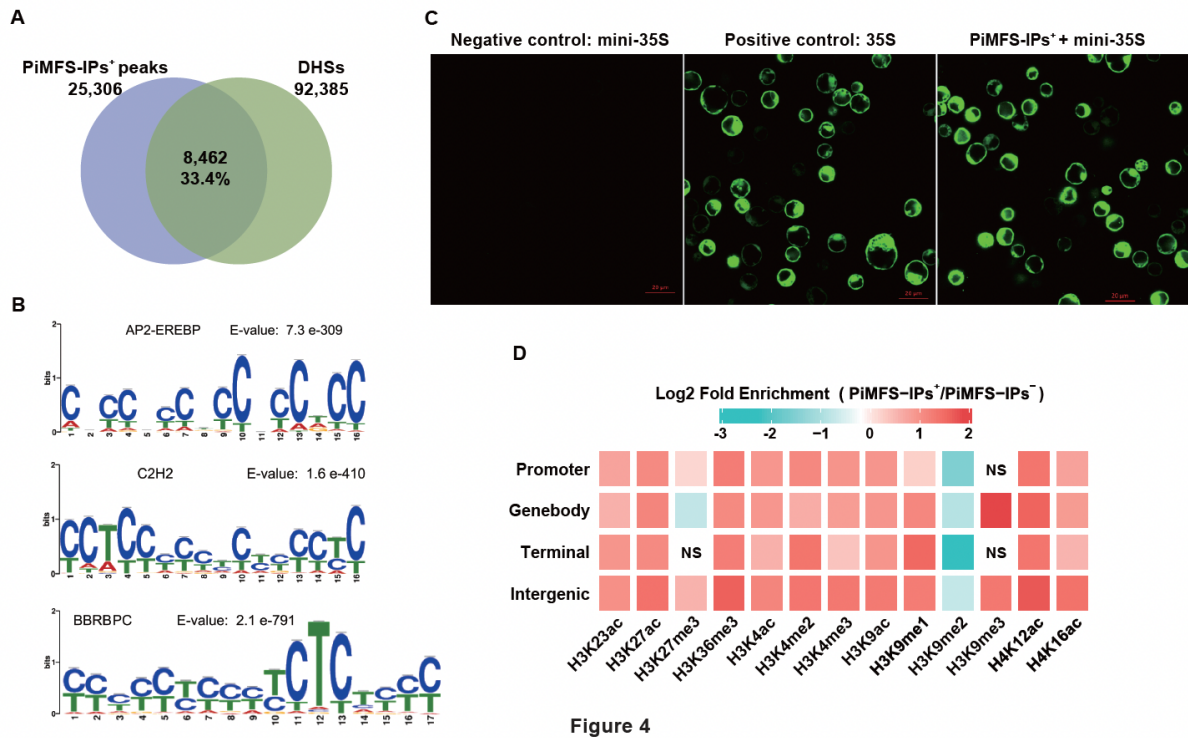


Figure 4

**Figure 4 Functional characterization of PiMFS-IPs<sup>+</sup>.** (A) Venn plot illustrating PiMFS-IPs<sup>+</sup> peaks overlapping DHSs, which were identified using published DNase-seq (39). (B) Motif discovery around sequences covering PiMFS-IPs<sup>+</sup> peaks using MEME. The top 3 significantly enriched motifs with the highest E-values are listed. (C) Transiently expressed GFP signal detected from protoplasts transfected with a modified pJIT163-hGFP vector containing mini35S promoter only (negative control, left), an intact 35S promoter (positive control, middle) and a mini35S promoter + amplified PiMFS-IPs<sup>+</sup> DNA fragment (right). GFP signals were captured using fluorescent microscopy. (D) Heat map showing enrichment of each mark for PiMFS-IPs<sup>+</sup> relative to PiMFS-IPs<sup>-</sup> in each subgenomic region as indicated. The color gradient representing the fold enrichment of PiMFS-IPs<sup>+</sup> relative to PiMFS-IPs<sup>-</sup>. Chi-square test was conducted to determine the significance of the differences. NS representing no significance and panels without NS representing significant enrichment (p-value < 0.05).

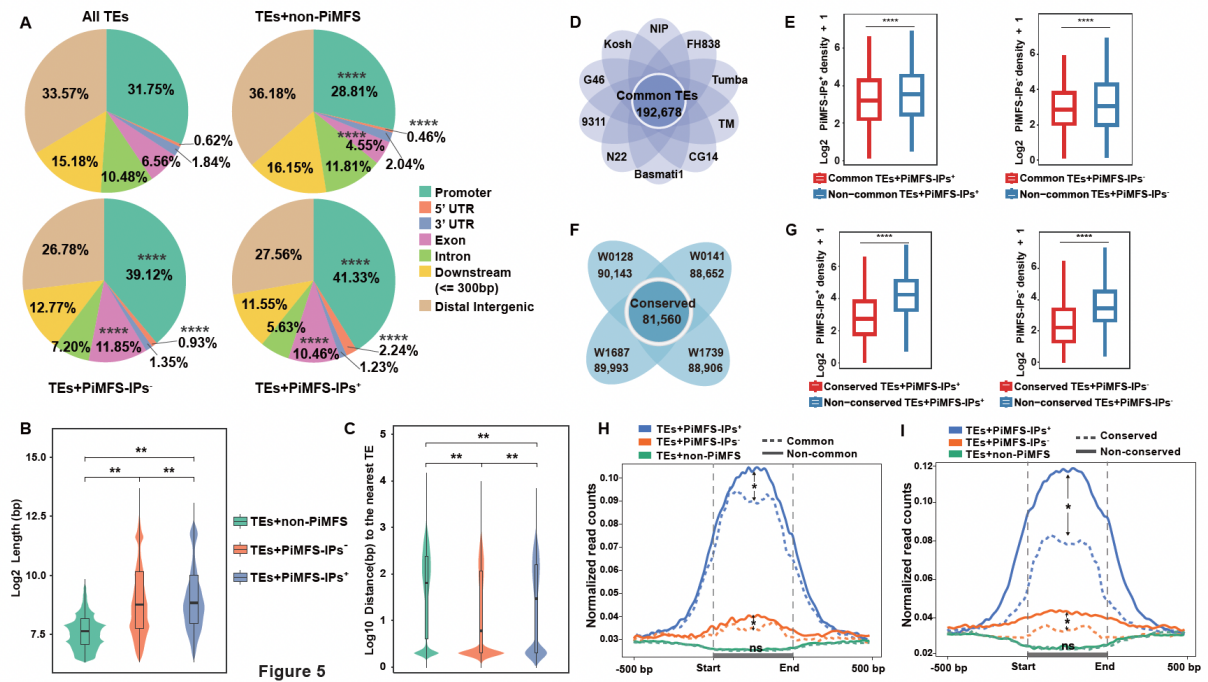
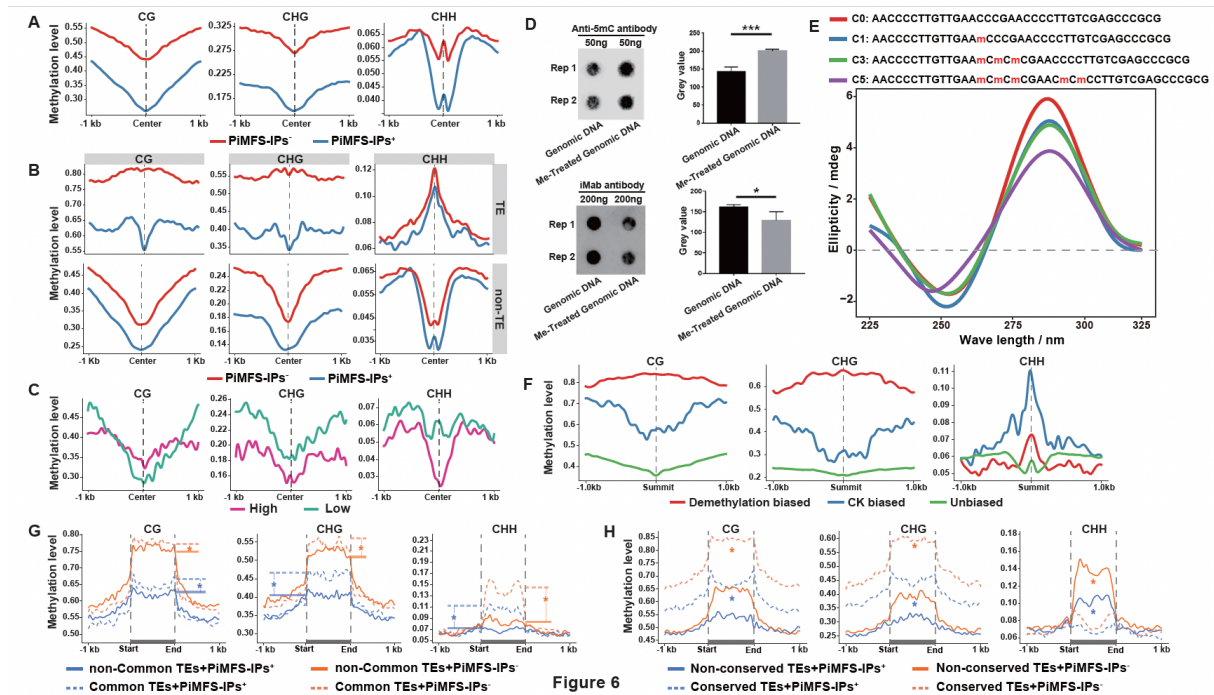


Figure 5

**Figure 5 Differential enrichment and impacts of PiMFS-IPs<sup>+</sup> on TE activity.** (A) Subgenomic distributions of all TEs, TEs containing PiMFS-IPs<sup>+/-</sup> or without PiMFSs, including promoters, 5'UTRs, 3'UTRs, exons, introns, downstream and intergenic regions. Significance test was determined by hypergeometric test. \*\*\*\* p-value < 0.0001. (B) Mean length of TEs containing PiMFS-IPs<sup>+/-</sup> or without PiMFSs. (C) Distance from TEs containing PiMFS-IPs<sup>+/-</sup> or without PiMFSs to the nearest neighboring TEs. Significance tests in B and C were determined by Wilcoxon rank-sum test. \*\* p-value < 0.01, \*\*\*\* p-value < 0.0001. (D) Venn plot illustrating common TEs (n = 192,678) between Nipponbare (NIP) and the other nine cultivated *Oryza* subpopulations. (E) Mean number per kb of PiMFS-IPs between common and non-common TEs containing PiMFS-IPs<sup>+</sup> (left) or PiMFS-IPs<sup>-</sup> (right). (F) Venn plot illustrating 81,560 conserved TEs between Nipponbare and the 4 wild rice species. (G) Mean number per kb of PiMFS-IPs between conserved and non-conserved TEs containing PiMFS-IPs<sup>+</sup> (left) or PiMFS-IPs<sup>-</sup> (right). (H) Profiling of iM-IP-seq read density across ± 500 bp from the start to the end of common (dotted lines)/non-common (solid lines) TEs containing PiMFS-IPs<sup>+</sup> (blue), PiMFS-IPs<sup>-</sup> (orange) and non-PiMFS (green), respectively. (I) Profiling of iM-IP-seq read density across ± 500 bp from the start to the end of conserved (dotted lines)/non-conserved (solid lines) TEs overlapping PiMFS-IPs<sup>+</sup> (blue), PiMFS-IPs<sup>-</sup> (orange) and non-PiMFS (green), respectively. Significance tests in E, G, H and I were determined by Wilcoxon rank-sum test. \*\*\*\* p-value < 0.0001, \* p-value < 0.05, NS: non significance.





**Figure 6 Differential DNA methylation between TEs overlapping PiMFS-IPs<sup>+/-</sup>.** (A) CG, CHG, and CHH methylation levels around  $\pm 1$  kb of the center of PiMFS-IPs<sup>+</sup> ( $n = 180,608$ ) and PiMFS-IPs<sup>-</sup> ( $n = 180,608$ ) with similar C content. (B) CG, CHG, and CHH methylation levels around  $\pm 1$  kb of the center of TE containing PiMFS-IPs<sup>+</sup> ( $n = 54,716$ ) and PiMFS-IPs<sup>-</sup> ( $n = 54,716$ ) with similar C content (top), and non-TE containing PiMFS-IPs<sup>+</sup> ( $n = 125,892$ ) and PiMFS-IPs<sup>-</sup> ( $n = 125,892$ ) with similar C content (bottom). (C) CG, CHG, and CHH methylation levels between PiMFS-IPs<sup>+</sup> peaks with high (pink) and low (green) peak abundance ( $n = 8,345$ ). (D) Anti-5mC and iMab antibody based dot blotting assays for genomic DNA with or without treatment of M.SssI, a CG specific DNA methyltransferase. The dot signal in each blot was quantified from at least two replicates (right). (E) CD assay of a synthesized oligonucleotide containing 1 C methylation (C1), 3 consecutive C methylations (C3), 5 methylated Cs (C5) or non-C methylation (C0) as indicated. (F) CG, CHG, and CHH methylation levels around  $\pm 1$  kb of the summit of demethylated DNA biased, CK biased and unbiased PiMFS-IPs<sup>+</sup> peaks. (G) CG, CHG, and CHH methylation levels across  $\pm 1$  kb from the start to the end of common (dotted lines) and non-common (solid lines) TEs ( $n = 19,351$ ) containing PiMFS-IPs<sup>+/-</sup> with similar C content. (H) CG, CHG, and CHH methylation levels across  $\pm 1$  kb from the start to the end of conserved (dotted lines) and non-conserved (solid lines) TEs ( $n = 20,491$ ) containing PiMFS-IPs<sup>+/-</sup> with similar C content. Significance tests in A and B were determined by Wilcoxon rank-sum test. \* p-value < 0.05.



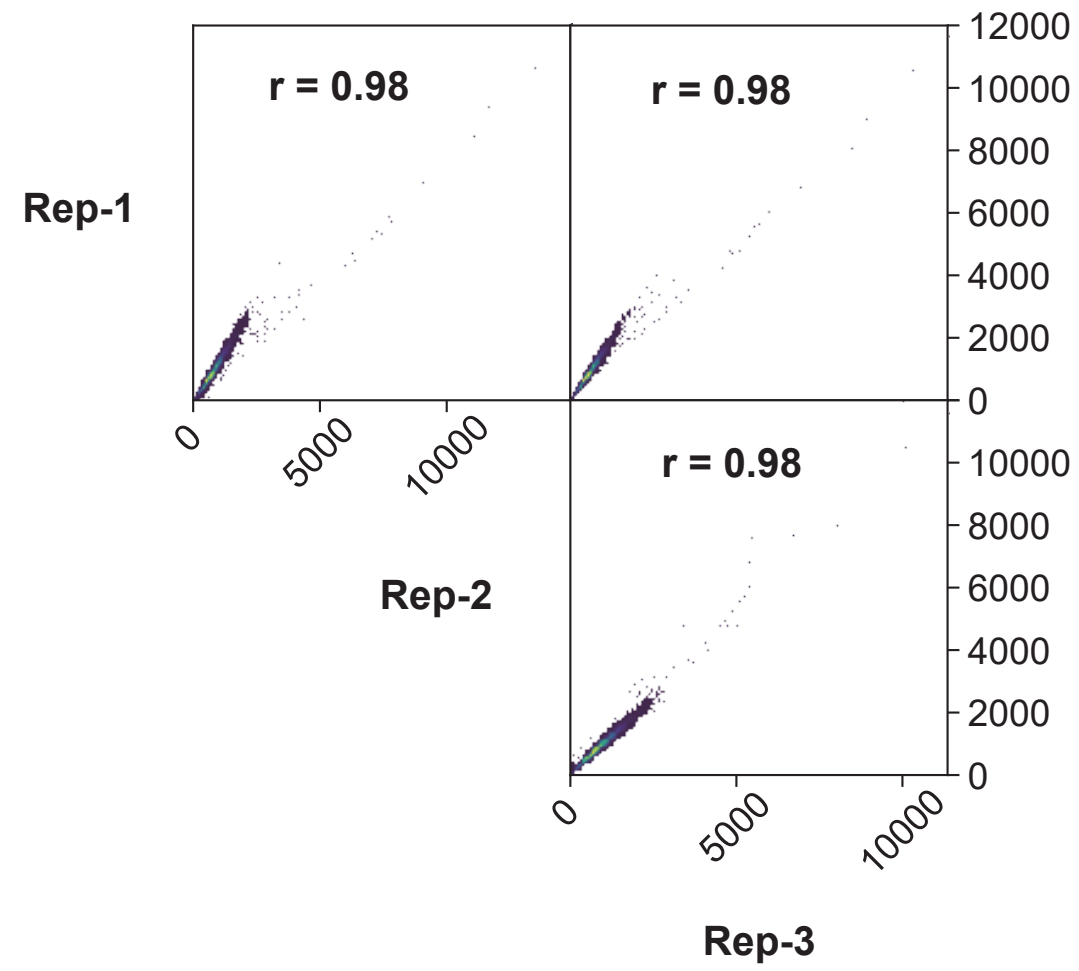
## Supplementary Materials

### **Genome-wide characterization of i-motifs and their potential roles in the stability and evolution of transposable elements in rice**

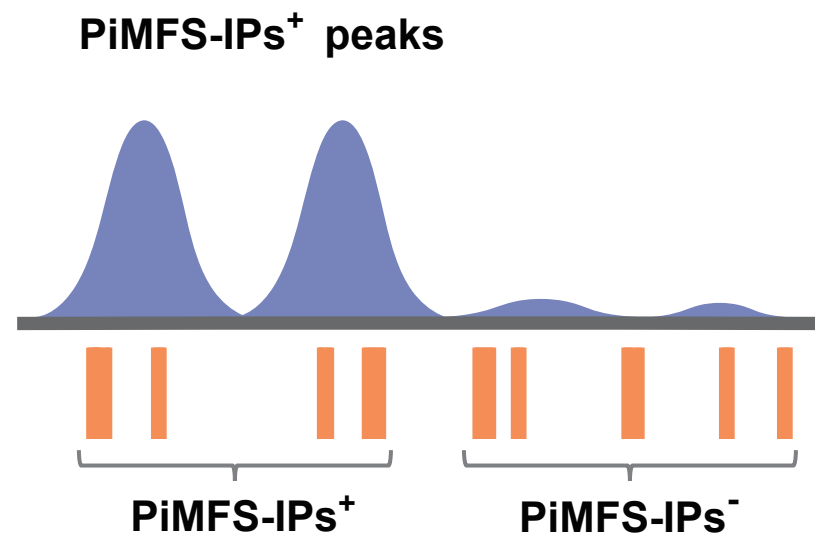
Xing Ma<sup>1,#</sup>, Yilong Feng<sup>1,#</sup>, Ying Yang<sup>1</sup>, Xin Li<sup>2</sup>, Yining Shi<sup>1</sup>, Shentong Tao<sup>1</sup>, Xuejiao Cheng<sup>1</sup>, Jian Huang<sup>3</sup>, Xiu-e Wang<sup>1</sup>, Caiyan Chen<sup>2</sup>, David Monchaud<sup>4</sup>, Wenli Zhang<sup>1\*</sup>

<sup>1</sup>State Key Laboratory for Crop Genetics and Germplasm Enhancement, Collaborative Innovation Center for Modern Crop Production co-sponsored by Province and Ministry (CIC-MCP), Nanjing Agricultural University, No.1 Weigang, Nanjing, Jiangsu 210095, P. R. China. <sup>2</sup>Institute of Subtropical Agriculture, Chinese Academy of Sciences, Changsha, Hunan, 410125, P. R. China. <sup>3</sup>School of Biology & Basic Medical Science, Soochow University, Suzhou, Jiangsu 215123, P. R. China. <sup>4</sup>Institut de Chimie Moleculaire, ICMUB CNRS UMR 6302, UBFC Dijon, France.

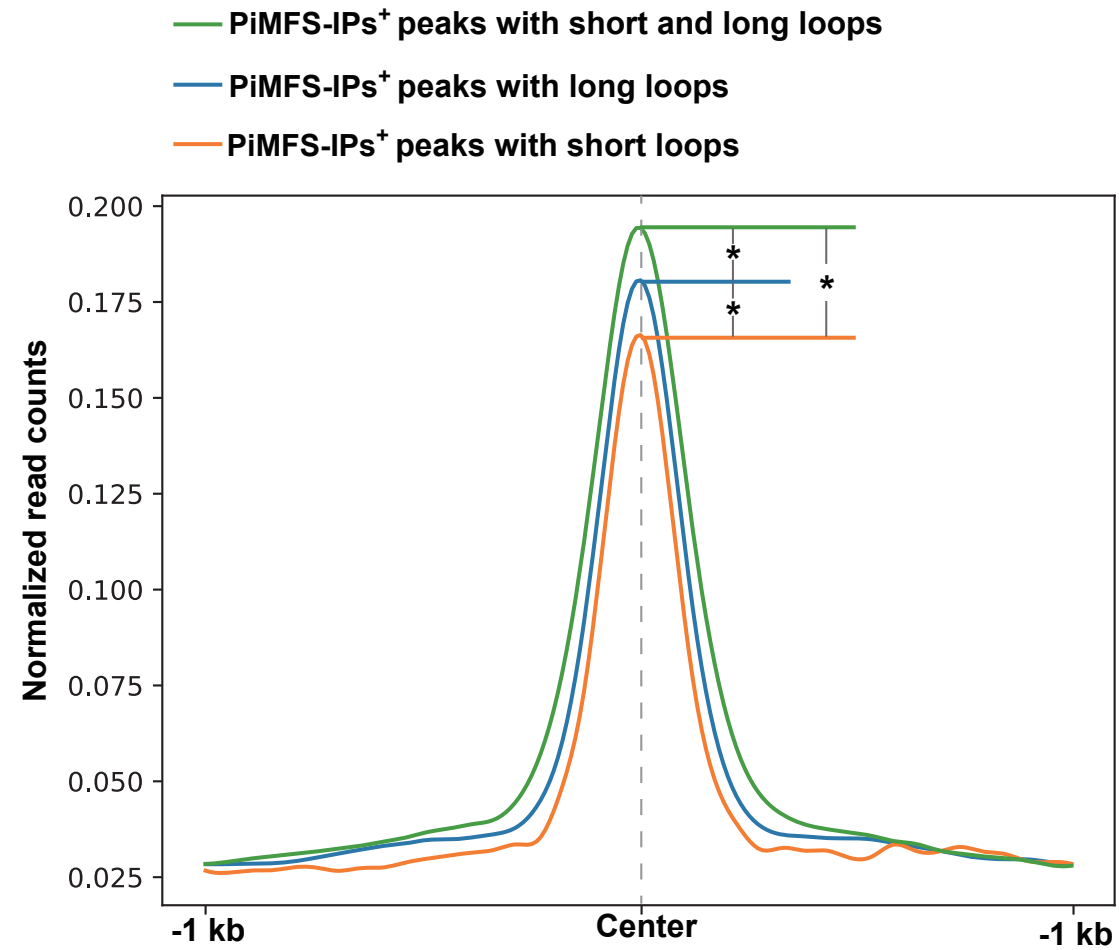
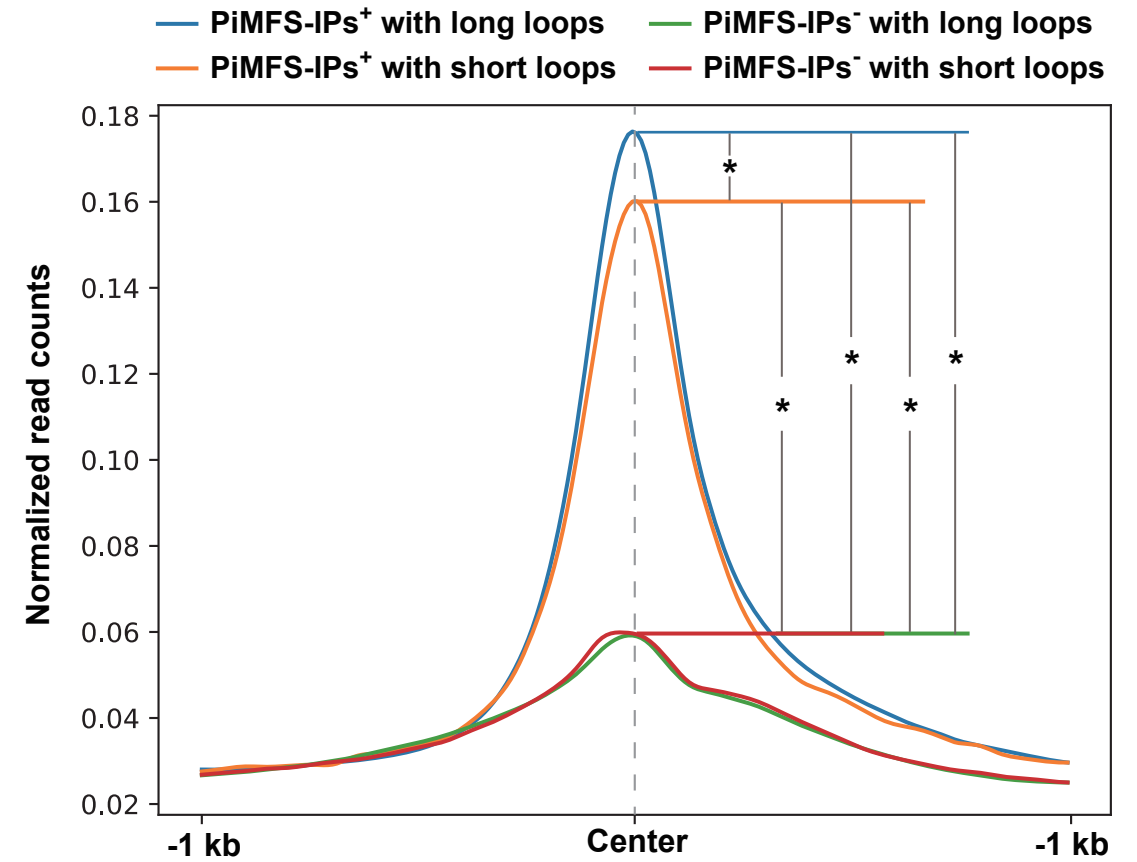




Supplementary Figure S1. Correlation analyses of biologically triplicated iM-IP-seq data sets



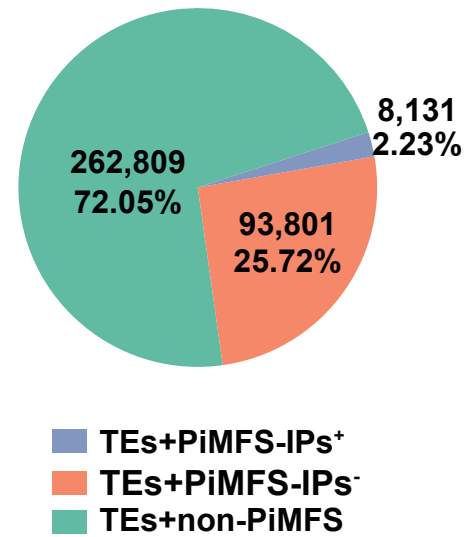
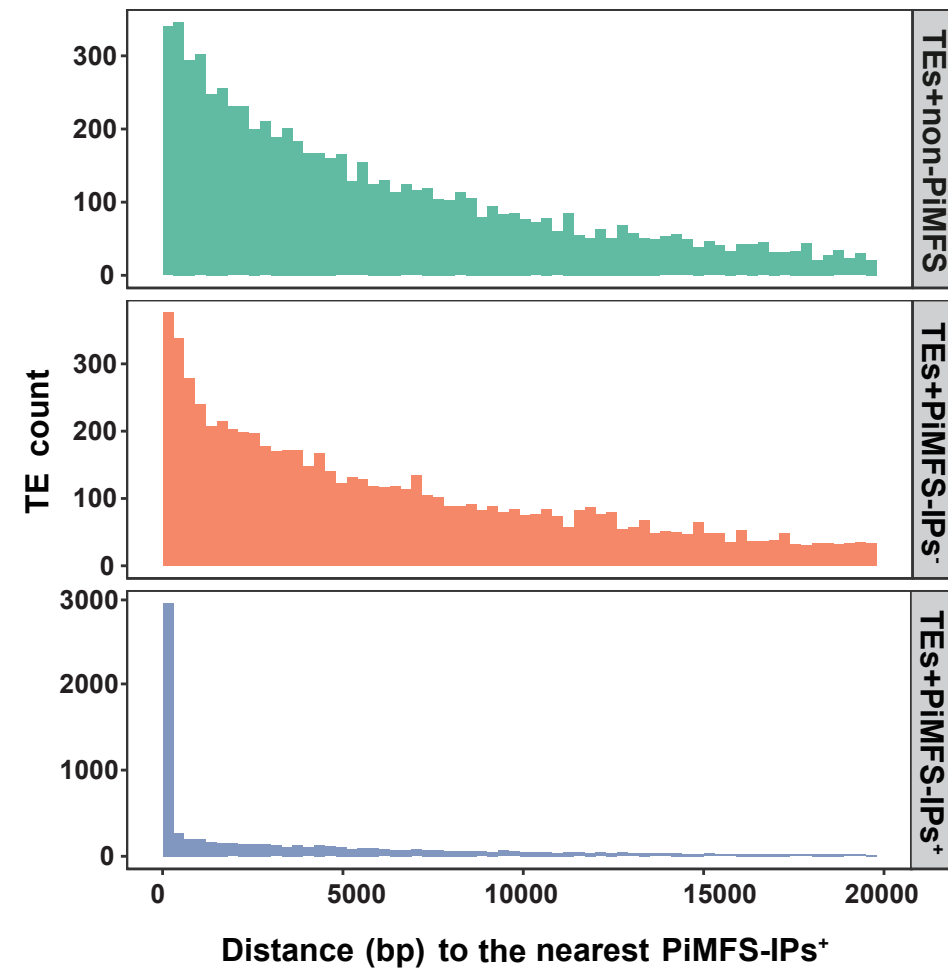
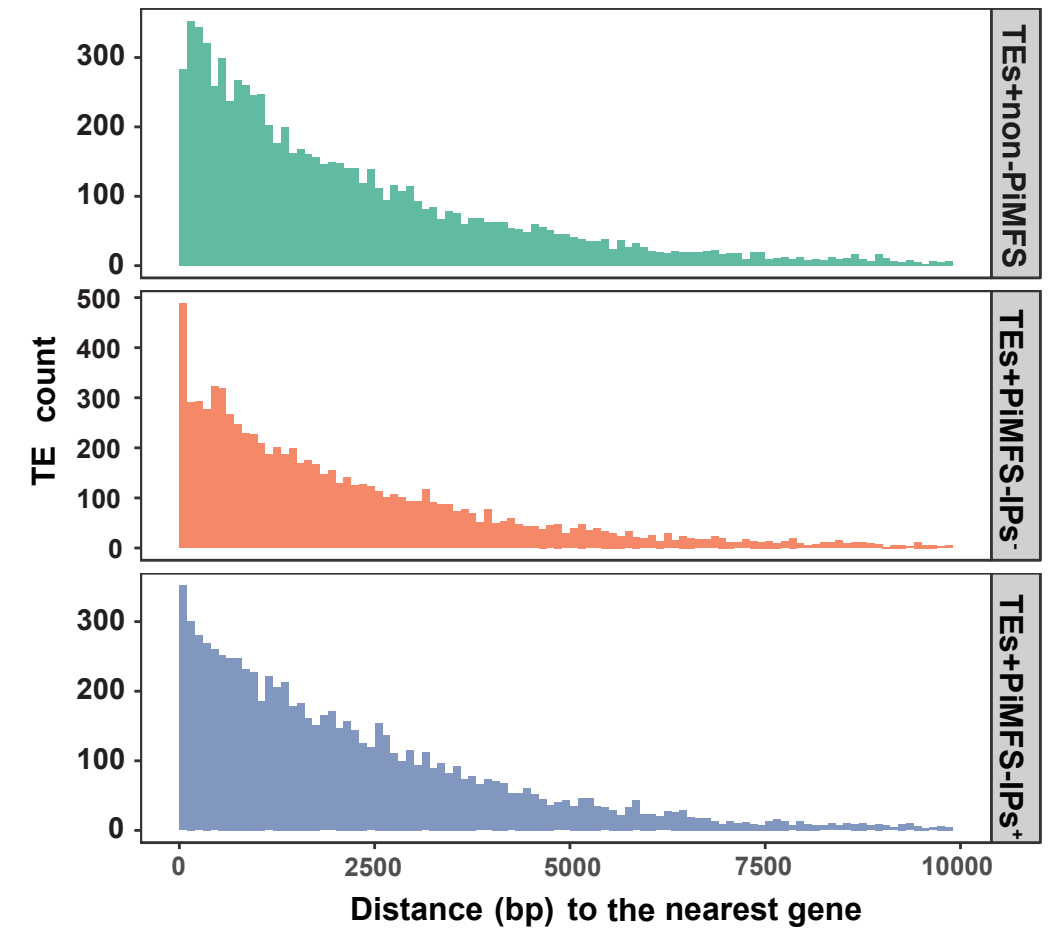
Supplementary Figure S2. Schematic diagram illustrating two subtypes of PiMFSs with or without PiMFS-IPs<sup>+</sup>

**A****B**

**Supplementary Figure S3. Comparison of PiMFS-IPs<sup>+</sup> read intensity among PiMFS-IPs<sup>+</sup> peaks with different loop sizes.**

**(A)** Normalized PiMFS-IPs<sup>+</sup> read counts distributed across  $\pm 1$  kb of the center of PiMFS-IPs<sup>+</sup> peaks with long loops, short loops and both long and short loops.

**(B)** Normalized PiMFS-IPs<sup>+</sup> read counts distributed across  $\pm 1$  kb of the center of PiMFS-IPs<sup>+/-</sup> with long or short loops.

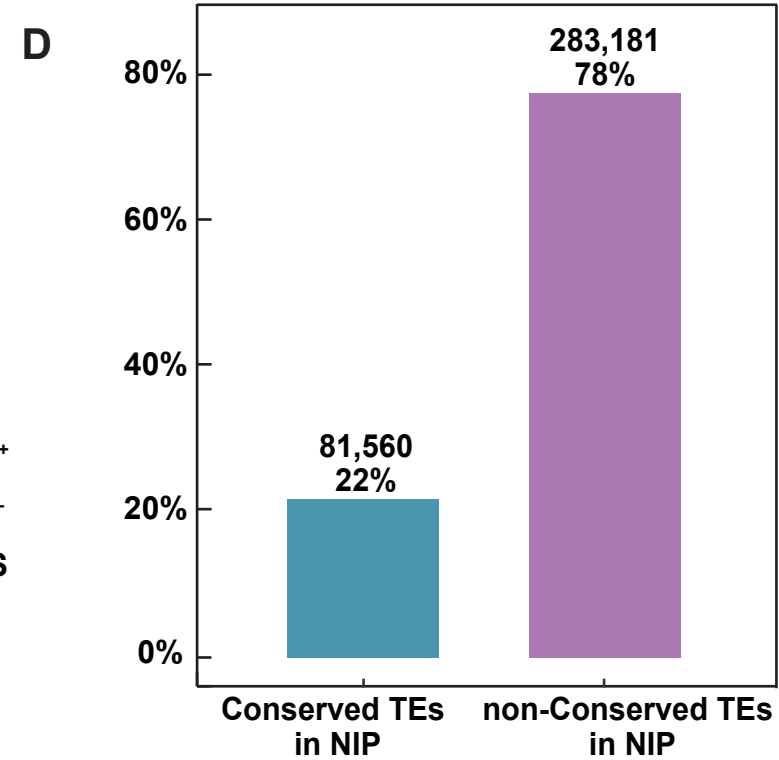
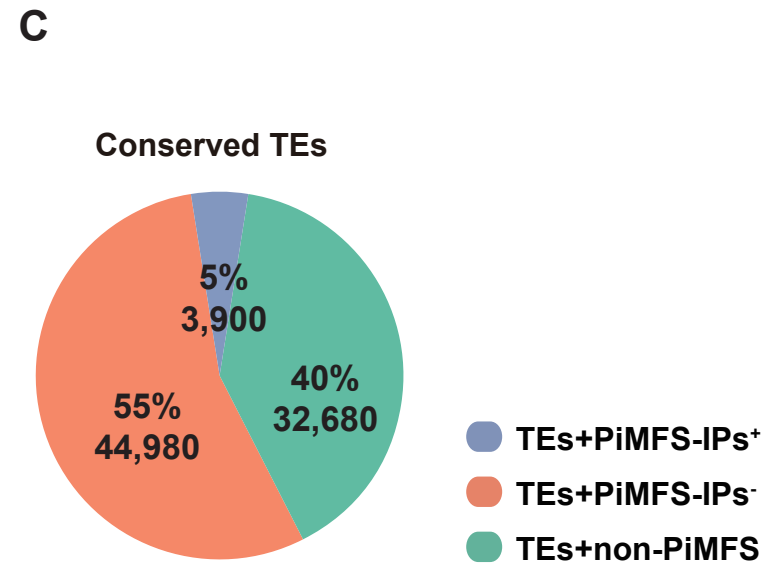
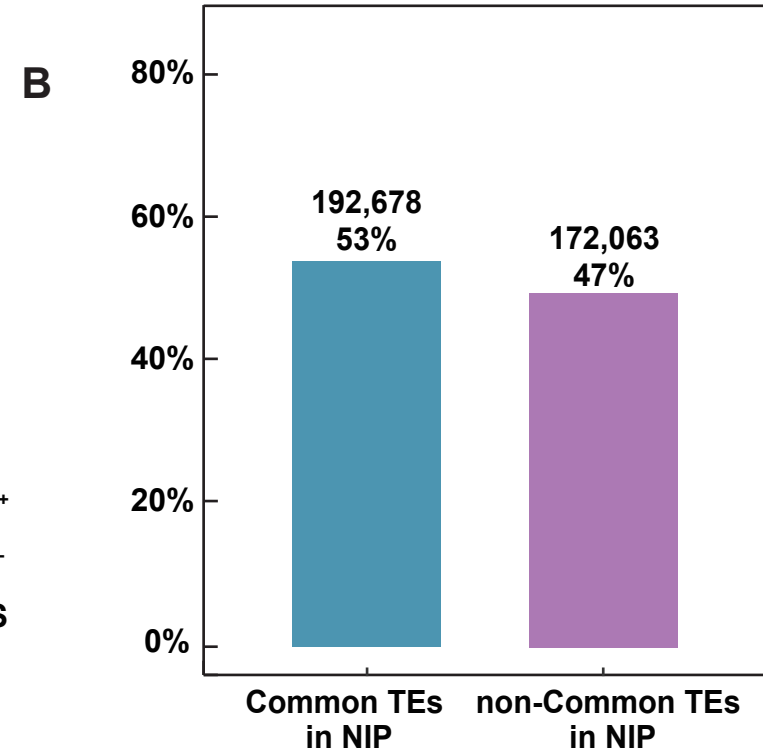
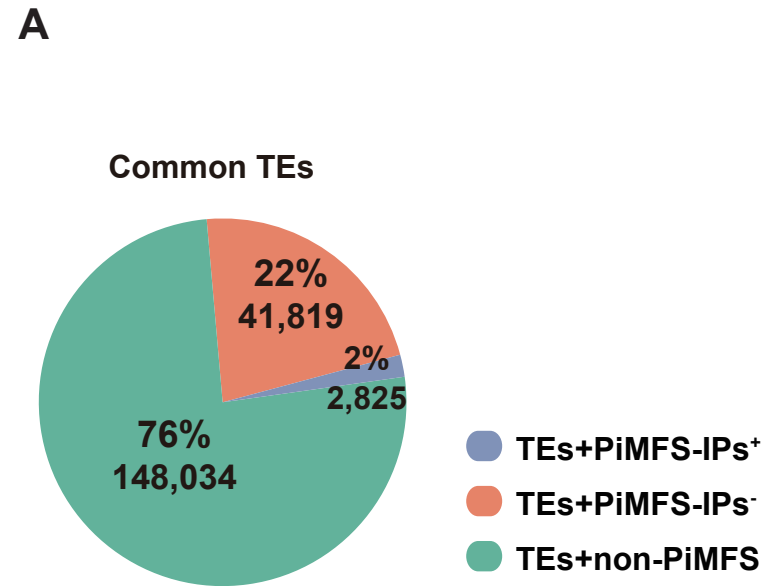
**A****B****C**

**Supplementary Figure S4. Characterization of TEs containing PiMFS-IPs<sup>+/-</sup> or without PiMFSs.**

**(A)** Proportion of TEs containing PiMFS-IPs<sup>+/-</sup> or without PiMFSs.

**(B)** Density map showing the distance from TEs containing PiMFS-IPs<sup>+/-</sup> (bottom/middle) or without PiMFSs (top) to the nearest neighboring PiMFS-IP.

**(C)** Density map showing the distance from TEs containing PiMFS-IPs<sup>+/-</sup> (bottom/middle) or without PiMFSs (top) to the nearest neighboring genes.



**Supplementary Figure S5. Characterization of common/non-common and conserved/non-conserved TEs.**

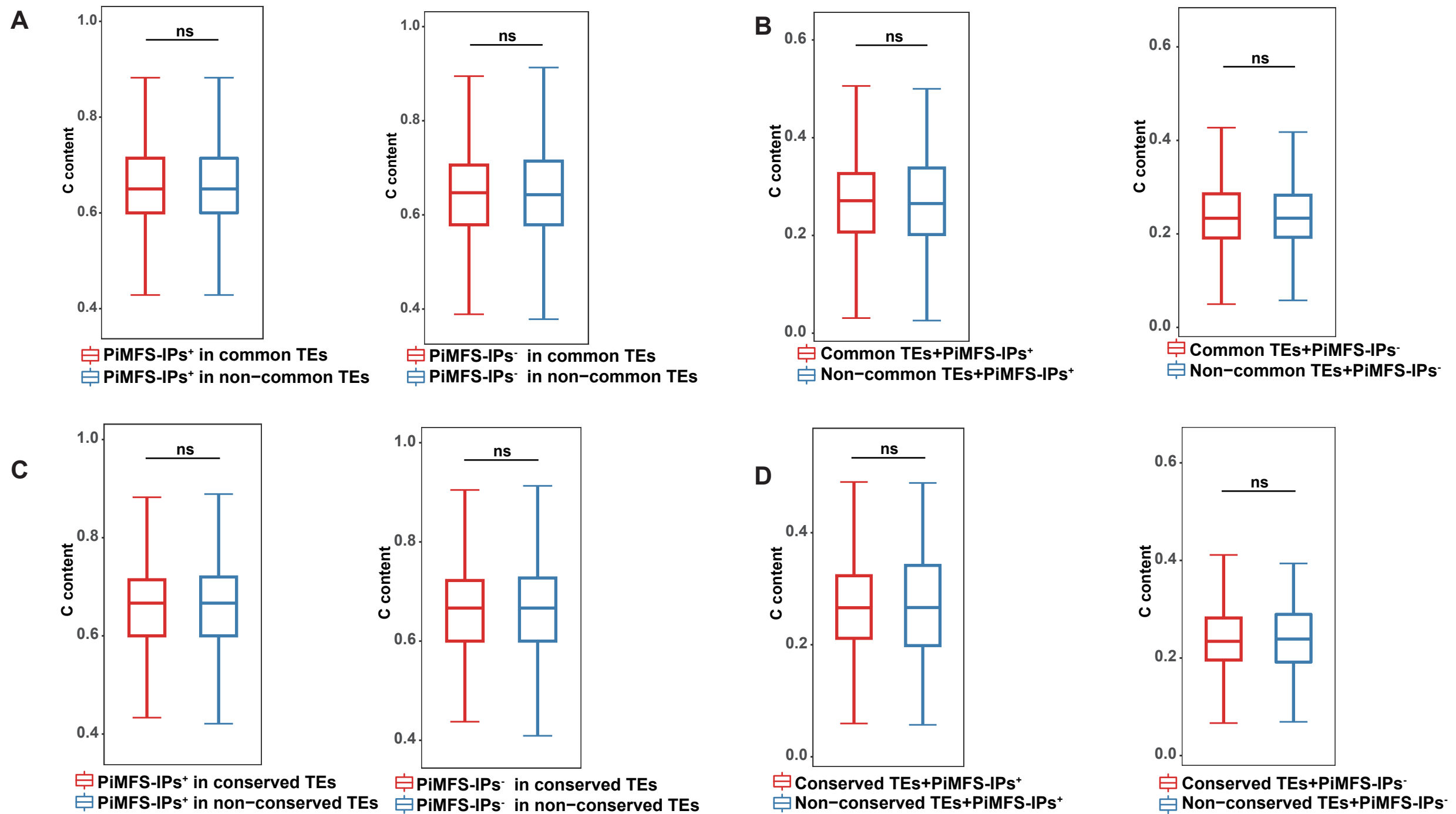
(A) Proportion of common TEs containing PiMFS-IPs<sup>+/-</sup> or without PiMFSs.

(B) Proportion of common and non-common TEs relative to all TEs in NIP (Nipponbare).

(C) Proportion of conserved TEs containing PiMFS-IPs<sup>+/-</sup> or without PiMFSs.

(D) Proportion of conserved and non-conserved TEs relative to all TEs in NIP (Nipponbare).





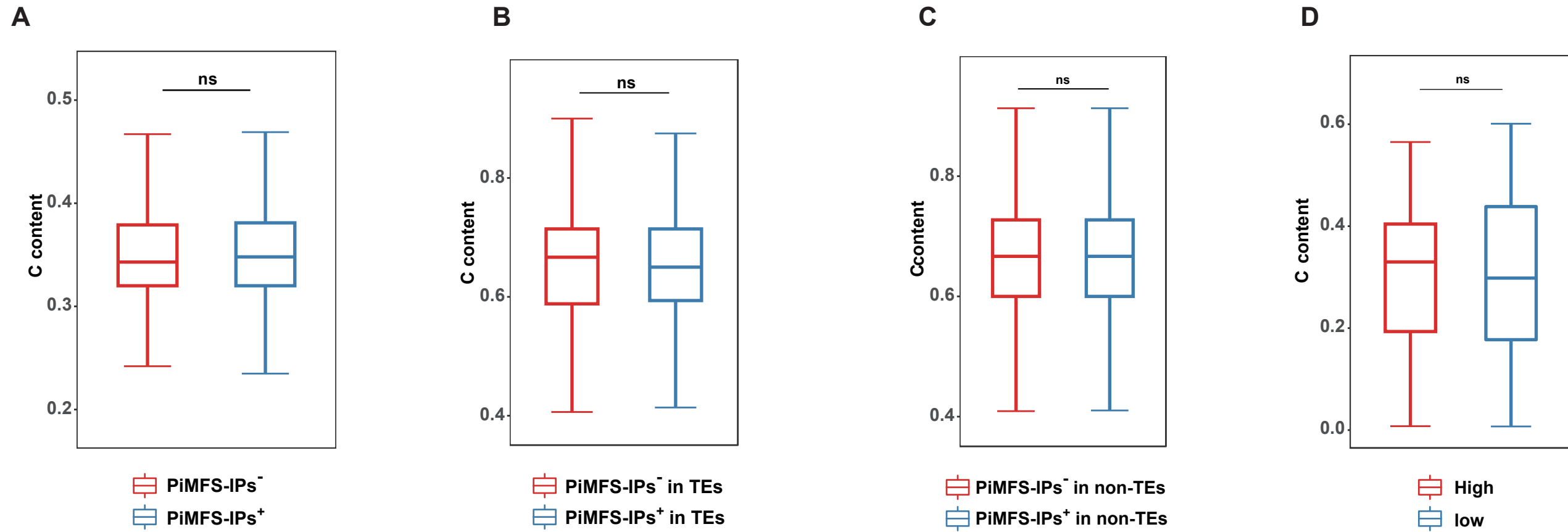
**Supplementary Figure S7. Normalization of C content.**

**(A)** Normalization of C content between PiMFS-IPs<sup>+</sup> (left) and PiMFS-IPs<sup>-</sup> (right) in common and non-common TEs.

**(B)** Normalization of C content between common/non-common TEs containing PiMFS-IPs<sup>+</sup> (left) and PiMFS-IPs<sup>-</sup> (right).

**(C)** Normalization of C content between PiMFS-IPs<sup>+</sup> (left) and PiMFS-IPs<sup>-</sup> (right) in conserved and non-conserved TEs.

**(D)** Normalization of C content between conserved/non-conserved TEs containing PiMFS-IPs<sup>+</sup> (left) and PiMFS-IPs<sup>-</sup> (right).



**Supplementary Figure S8. Normalization of C content.**

(A) Normalization of C content between PiMFS-IPs<sup>+/-</sup>.

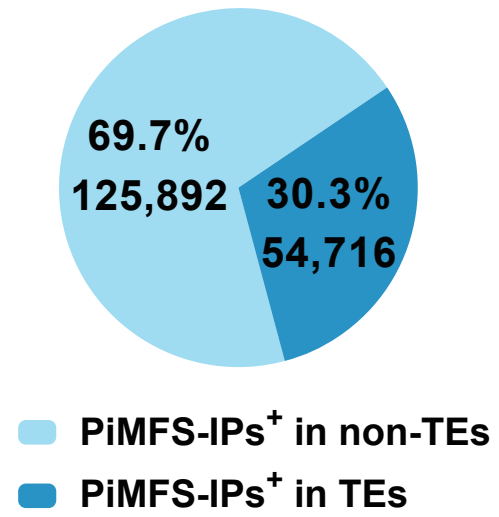
(B) Normalization of C content between PiMFS-IPs<sup>+/-</sup> in TEs.

(C) Normalization of C content between PiMFS-IPs<sup>+/-</sup> in non-TEs.

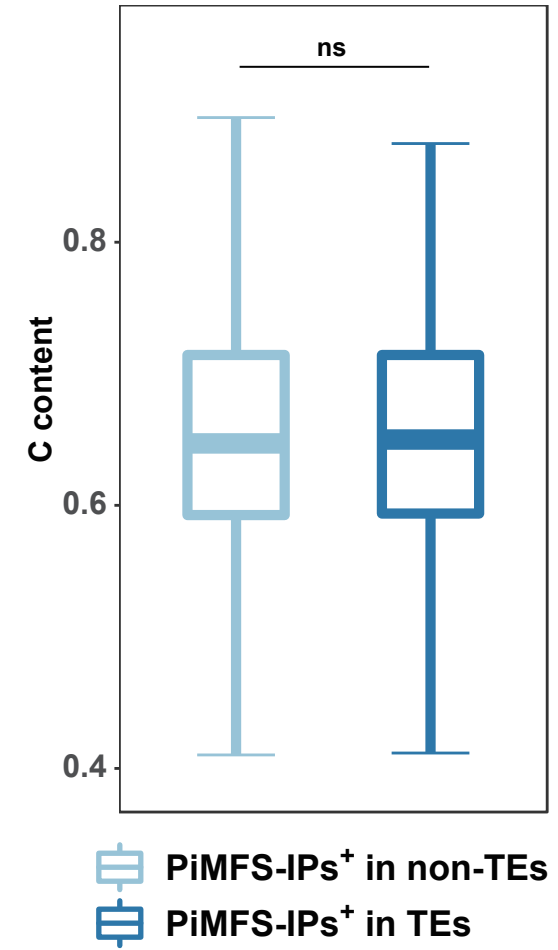
(D) Normalization of C content between PiMFS-IPs<sup>+</sup> peaks with high and low peak abundance.



**A**



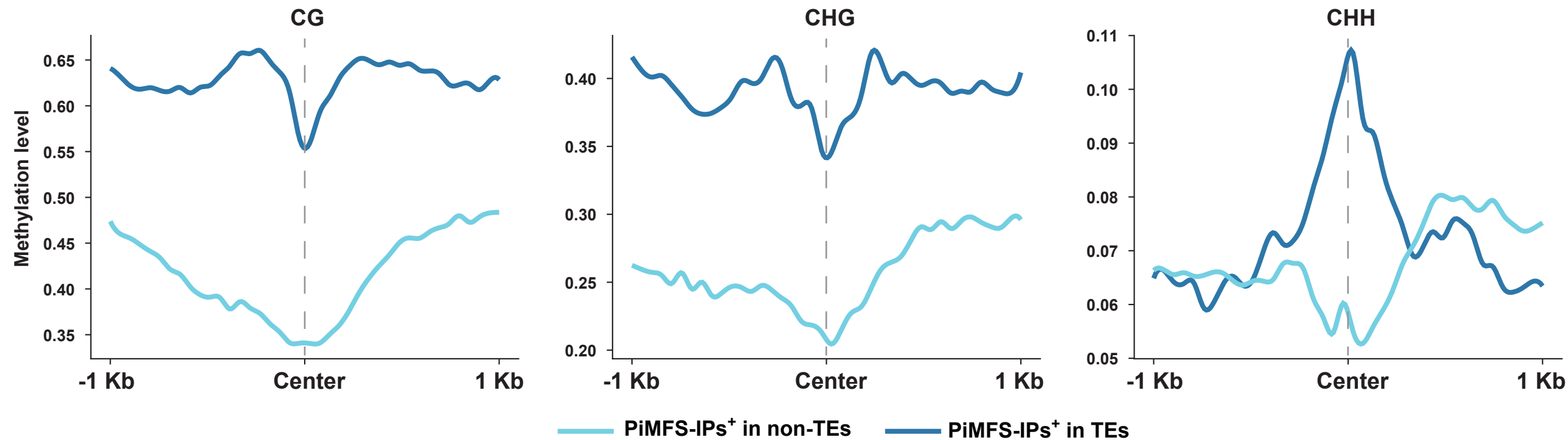
**B**



**Supplementary Figure S9. Characterization of non-TEs and TEs containing PiMFS-IPs<sup>+</sup>.**

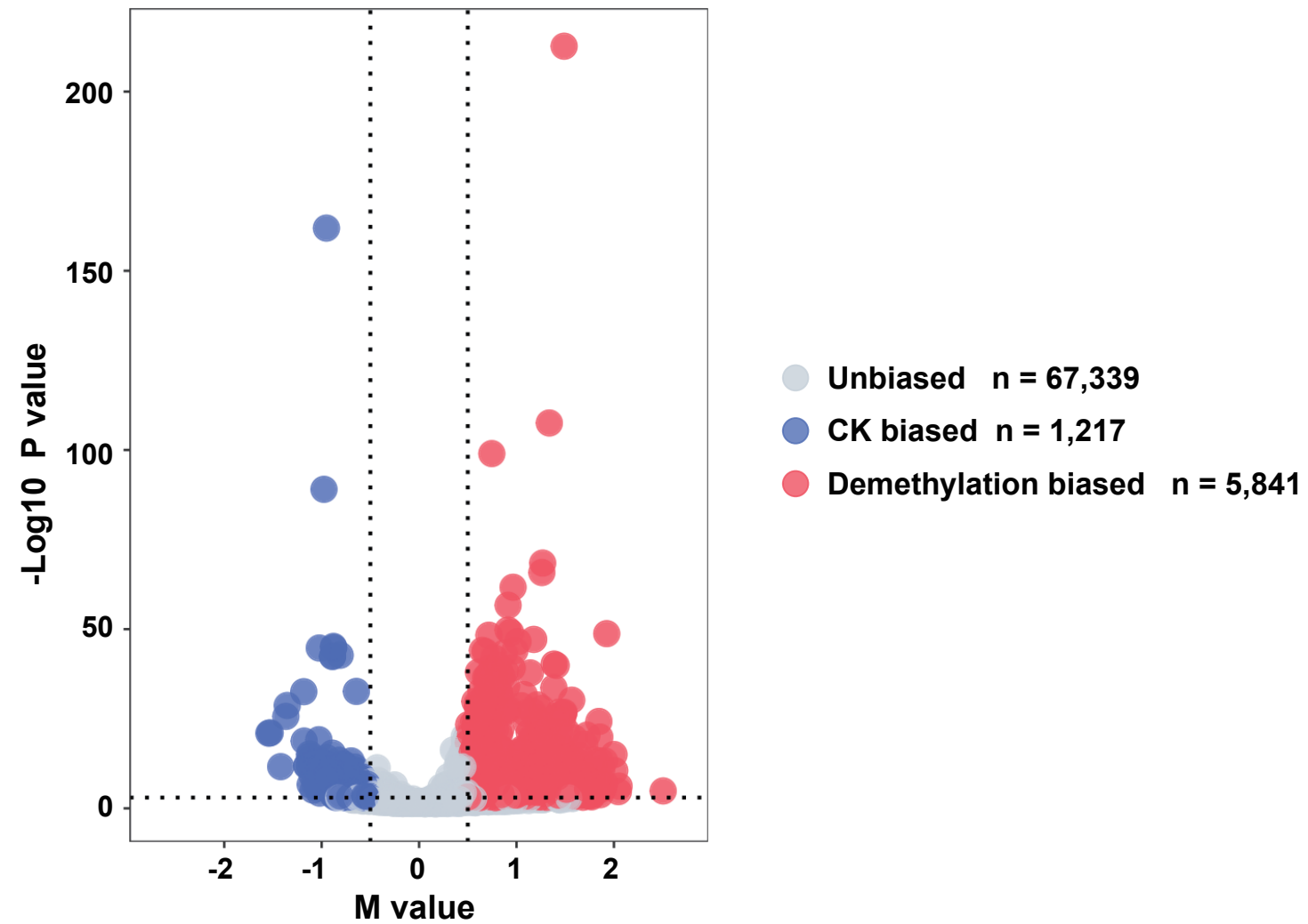
**(A)** Proportion of PiMFS-IPs<sup>+</sup> in non-TEs and TEs.

**(B)** Normalization of C content between PiMFS-IPs<sup>+</sup> in non-TEs and TEs.



**Supplementary Figure S10. Comparisons of CG, CHG and CHH methylation levels between PiMFS-IPs<sup>+</sup> in non-TEs and TEs.**

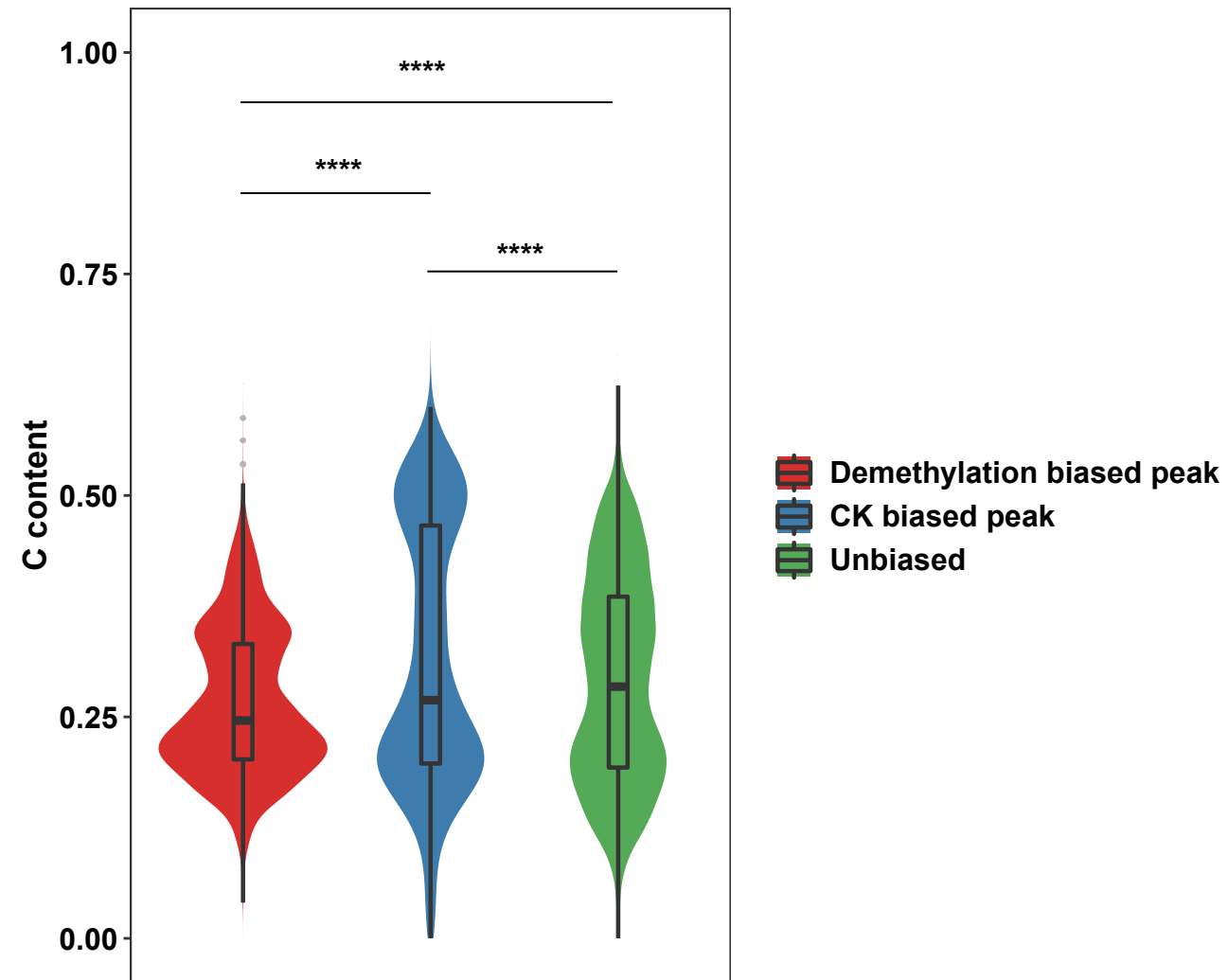
CG (left), CHG (middle) and CHH (right) methylation levels were counted across  $\pm 1$ kb of the center of PiMFS-IP<sup>+</sup>.



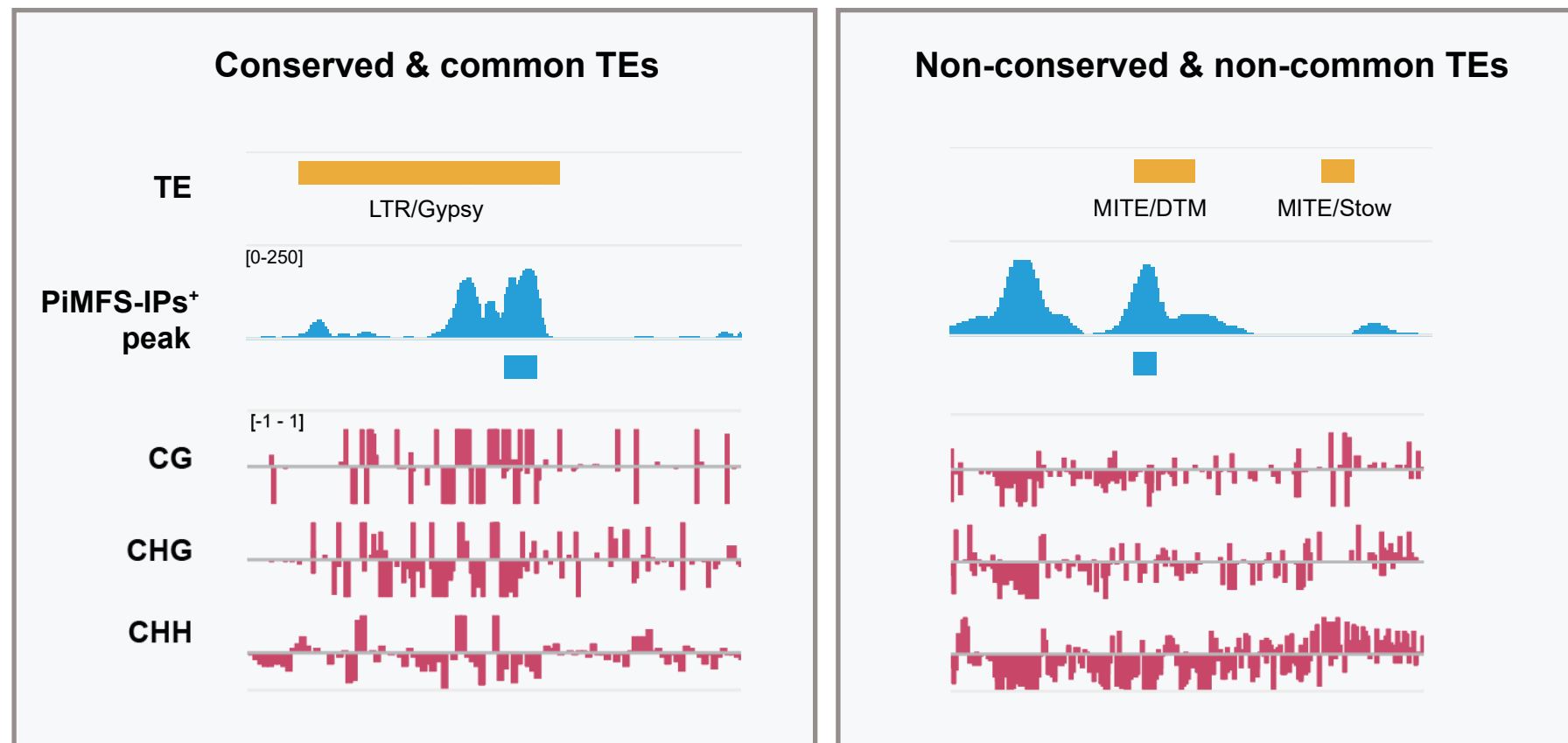
**Supplementary Figure S11. Effects of demethylated DNA induced by Zebularine treatment on PiMFS-IPs<sup>+</sup> formation.**

MA plot illustrating demethylated DNA biased PiMFS-IPs<sup>+</sup> peaks (n = 5,841), CK biased PiMFS-IPs<sup>+</sup> peaks (n = 1,217) and unbiased PiMFS-IPs<sup>+</sup> peaks (67,339). y-axis represents  $-\log_{10}(\text{p-value})$  and x-axis represents M value defined by MAnorm.

Dotted line representing the cutoff with  $y = 3$  and  $x = \pm 0.5$ .



**Supplementary Figure S12. Mean C content of demethylation biased PiMFS-IPs<sup>+</sup> peaks (26.5%), CK biased PiMFS-IPs<sup>+</sup> peaks (31.3%) and unbiased PiMFS-IPs<sup>+</sup> peaks (29.1%).** Significance test was determined by Wilcoxon rank-sum test. \*\*\*\* p-value < 0.0001.



Supplementary Figure S13. Representative Integrative Genomics Viewer (IGV) snapshots for PiMFS-IPs<sup>+</sup> peak and DNA methylation levels corresponding to conserved/common (left) and non-conserved/non-common TEs (right).

**Supplementary Table S1. Summary of sequencing data analyzed in this study**

<b>Samples</b>	<b>No. of reads</b>	<b>Mapped Reads</b>	<b>Mapping Rate</b>	<b>No. of uniquely mapped reads</b>
<b>Input</b>	<b>24,077,398</b>	<b>23,902,105</b>	<b>99.21%</b>	<b>15,319,769</b>
<b>IgG IP</b>	<b>42,995,154</b>	<b>32,188,538</b>	<b>74.84%</b>	<b>14,668,162</b>
<b>iM-IP-Rep1</b>	<b>37,969,720</b>	<b>37,562,907</b>	<b>98.93%</b>	<b>28,839,962</b>
<b>iM-IP-Rep2</b>	<b>42,000,000</b>	<b>41,480,426</b>	<b>98.75%</b>	<b>32,027,103</b>
<b>iM-IP-Rep3</b>	<b>45,318,712</b>	<b>44,983,018</b>	<b>99.32%</b>	<b>33,824,669</b>

**Supplementary Table S3. Summary of regular and non-regular PiMFSs identified in this study**

<b>Subtypes</b>	<b>Pattern</b>	<b>Sequences</b>	<b>PiMFS number</b>	<b>Peak number/ Percentage</b>
<b>Regular</b>	<b>[CC<sub>2-5</sub>L<sub>1-15</sub>]<sub>3</sub> CC<sub>2-5</sub></b>	<b>CCCAGGCCCTCCTCC</b>	<b>1,475,053</b>	<b>22,425 / 88.6%</b>
<b>Regular with Insertion</b>	<b>[CC<sub>2-5</sub>L<sub>1-15</sub>]<sub>3</sub> CC<sub>2-5</sub></b>	<b>CTCTACCCCTCCACC</b>	<b>268,021</b>	<b>521 / 2.1%</b>
<b>3 Runs</b>	<b>[CC<sub>2-5</sub>L<sub>1-15</sub>]<sub>2</sub> CC<sub>2-5</sub></b>	<b>CCCGCCCCAATCC</b>	<b>1,716,374</b>	<b>923 / 3.6%</b>
<b>3 Runs with Insertion</b>	<b>[CC<sub>2-5</sub>L<sub>1-15</sub>]<sub>2</sub> CC<sub>2-5</sub></b>	<b>CGCATAACCCTCCC</b>	<b>262,943</b>	<b>101 / 0.4%</b>
<b>More Runs</b>	<b>[CC<sub>2-5</sub>L<sub>1-15</sub>]<sub>4-7</sub> CC<sub>2-5</sub></b>	<b>CCACCATTCCACCCCT CCATACC</b>	<b>3,352,292</b>	<b>92 / 0.4%</b>
<b>Total</b>			<b>7,074,683</b>	<b>24,062 / 95.1%</b>

**Supplementary Table S5. Summary of published data sets used in this study**

<b>Data</b>	<b>Accession number</b>	<b>References</b>
H3K4me3-ChIP-seq	GSE19602	He et al. 2010
H3K36me3-ChIP-seq	GSE26733	Zhang et al. 2012
H3K4me2-ChIP-seq	GSE26733	Zhang et al. 2012
H4K12ac-ChIP-seq	GSE26733	Zhang et al. 2012
H3K9ac-ChIP-seq	GSE79033	Fang et al. 2016
H3K27ac-ChIP-seq	GSE79033	Fang et al. 2016
H3K4ac-ChIP-seq	GSE79033	Fang et al. 2016
H3K9me1-ChIP-seq	GSE79033	Fang et al. 2016
H3K27me3-ChIP-seq	GSE79033	Fang et al. 2016
H3K9me3-ChIP-seq	GSE79033	Fang et al. 2016
H4K16ac-ChIP-seq	GSE69426	Lu et al. 2015
H3K23ac-ChIP-seq	GSE69426	Lu et al. 2015
H3K9me2-ChIP-seq	GSE81436	Tan et al. 2016