RESEARCH PAPER

Genome-wide mapping of 5-hydroxymethylcytosine in three rice cultivars reveals its preferential localization in transcriptionally silent transposable element genes

Xi-liang Wang1,3,*, Shu-hui Song2,*,†, Yong-Sheng Wu4,*, Yu-Li Li1,2,3, Ting-ting Chen2, Zhi-yuan Huang5, Shuo Liu6, Thomas L. Dunwell7, Gerd P. Pfeifer8, Jim M. Dunwell9, Raheema Wamaedeesa9, Ihsan Ullah9,10, Yinsheng Wang6,11 and Song-nian Hu2,†

1 CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
2 Core Genomic Facility, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
3 Graduate University of Chinese Academy of Sciences, Yuquan Road, Beijing 100039, China
4 Mudanjiang Youbo Pharmaceutical Co., Ltd, Heilongjiang 157011, China
5 State Key Laboratory of Hybrid Rice, Hunan Hybrid Rice Research Center, Changsha 410125, China
6 Environmental Toxicology Graduate Program, University of California, Riverside, CA 92521, USA
7 Department of Zoology, University of Oxford, UK
8 Beckman Research Institute, City of Hope Medical Centre, Duarte, CA 91010, USA
9 School of Agriculture, Policy and Development, University of Reading, Earley Gate, Reading RG6 6AR, UK
10 Agricultural Biotechnology Research Institute, Faisalabad 38000, Pakistan
11 Department of Chemistry, University of California, Riverside, CA 92521, USA

* These authors contributed equally to this work.
† To whom correspondence should be addressed. E-mail: songshh@big.ac.cn or husn@big.ac.cn

Received 16 March 2015; Revised 28 May 2015; Accepted 30 June 2015

Editor: Christine Raines

Abstract

5-Hydroxymethylcytosine (5hmC), a modified form of cytosine that is considered the sixth nucleobase in DNA, has been detected in mammals and is believed to play an important role in gene regulation. In this study, 5hmC modification was detected in rice by employing a dot-blot assay, and its levels was further quantified in DNA from different rice tissues using liquid chromatography-multistage mass spectrometry (LC-MS/MS/MS). The results showed large intertissue variation in 5hmC levels. The genome-wide profiles of 5hmC modification in three different rice cultivars were also obtained using a sensitive chemical labelling followed by a next-generation sequencing method. Thousands of 5hmC peaks were identified, and a comparison of the distributions of 5hmC among different rice cultivars revealed the specificity and conservation of 5hmC modification. The identified 5hmC peaks were significantly enriched in heterochromatin regions, and mainly located in transposable elements (TEs), especially around retrotransposons. The correlation analysis of 5hmC and gene expression data revealed a close association between 5hmC and silent TEs. These findings provide a resource for plant DNA 5hmC epigenetic studies and expand our knowledge of 5hmC modification.

Key words: 5-hydroxymethylcytosine, 5hmC profiles, heterochromatin, rice, transposable element.

Abbreviations: BWA, Burrows–Wheeler Alignment Tool; 5caC, 5-carboxylcytosine; ChIP-seq, chromatin immunoprecipitation sequencing; 5fC, 5-formylcytosine; β-GT, β-glucosyltransferase; GLIB, glucosylation, periodate oxidation, biotinylation; HCP, high confidence peak; 5hmC, 5-hydroxymethylcytosine; 5hmdC, 5-hydroxymethyl-2′-deoxycytidine; 5hmU, 5-hydroxyluracil; MRP, more reliable peak; NGS, next-generation sequencing; RPKM, reads per kilobase of exon model per million mapped reads; TE, transposable element; TET, ten-eleven translocation; TIGR, The Institute of Genomic Research Database; TTR, transcriptional termination regions.

© The Author 2015. Published by Oxford University Press on behalf of the Society for Experimental Biology. All rights reserved. For permissions, please email: journals.permissions@oup.com
Introduction

Being a conserved epigenetic mark, methylated cytosine (5-methylcytosine, 5mC) is involved in many important biological processes, including heterochromatin formation, defence against transposon proliferation, genomic imprinting, regulation of endogenous gene expression, and silencing of transgenes (Paszkowski and Whitham, 2001; Bender, 2004; Zhang et al., 2006; Zhang et al., 2014). Cytosine methylation occurs primarily at CG dinucleotide sites, although significant levels of methylation have also been found at CNG and CNN sites in plants (Cao and Jacobsen, 2002). DNA methylation in plants has been implicated in development and transcriptional regulation (Kondo et al., 2010; Marfil et al., 2012).

5-Hydroxymethylcytosine (5hmC), a recently identified epigenetic modification in mammals, is a base generated from the oxidation of 5mC by the TET (ten-eleven translocation) family of Fe(II)- and 2-oxoglutarate-dependent enzymes (Münzel et al., 2011; Zhou et al., 2014), and is now widely considered as the sixth base in DNA (Song and He, 2011). 5hmC, as a new epigenetic mark, has received tremendous attention from epigenetic and other related communities. In mammals, 5hmC is a key intermediate in active DNA demethylation and can be enzymatically deaminated to 5-hydroxymethyluracil (5hmU) or further oxidized to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by the TET family enzymes (Bhutani et al., 2010; Cortellino et al., 2011; Gong and Zhu, 2011; Guo et al., 2011; He et al., 2011). The distribution and function of 5hmC in mammals have also been widely studied. 5hmC mainly exists in promoter and intragenic regions, but is depleted from intergenic regions in human somatic neural tissue (Jin et al., 2011b).

Previous studies showed that 5hmC in gene bodies is preferentially localized to exons (Ficz et al., 2011; Pastor et al., 2011; Williams et al., 2011), and its level is positively correlated with the expression level of the corresponding genes (Laird et al., 2013; Pfeifer et al., 2014). Accumulating evidence reveals the importance of 5hmC in embryogenesis (Ito et al., 2010; Koh et al., 2011) and development (Ruzov et al., 2011; Yao and Jin, 2014). Moreover, 5hmC is strongly depleted in many human cancers relative to the corresponding normal tissue (Jin et al., 2011a; Yang et al., 2013). Many studies have been conducted in mammals regarding profiling the genome-wide distribution of TET and/or 5hmC, as well as assessing the roles of 5hmC in the regulation of gene expression (Iqbal et al., 2011; Robertson et al., 2011; Stroud et al., 2011; Taylor et al., 2014) and in the maintenance of pluripotency and self-renewal of embryonic stem cells (Iyver et al., 2009; Ito et al., 2010; Koh et al., 2011). However, very little is known about the genome-wide distribution and the epigenetic roles of 5hmC in plants.

In the past few years, there has been a debate about whether 5hmC exists in the plant genome, as there are considered to be no functional counterparts of TET family enzymes in plants (Jang et al., 2014). Terragni et al. (2012) estimated the global 5hmC content in the genomes of various plants using recombinant β-GT (β-glucosyltransferase), and found very low levels of 5hmC (<0.07% of the total nucleotides) in Arabidopsis, soybean, and rice. This is the first report of the presence of 5hmC in plant genomes. By using two complementary dot-blot assays, Yao et al. (2012) subsequently reported that genomic DNA from Arabidopsis leaves and flowers contained low but measurable levels of 5hmC, representing ~0.068–0.075% of the total cytosine nucleotides. Moricová et al. (2013) detected 5hmC in DNA isolated from protoplasts of Cucumis sativus and Brassica oleracea, and experimentally proved that oxidative DNA damage was not the main source of 5hmC. Moreover, Liu et al. (2013) reported the detection of ~0.79 molecules of 5-hydroxymethyl-2′-deoxycytidine (5hmDC) per million nucleosides in Arabidopsis genomic DNA using a liquid chromatography-multistage mass spectrometry (LC-MS/MS/MS) method, although a recent study concluded that 5hmC is not present in appreciable quantities in Arabidopsis DNA (Erdmann et al., 2015). Therefore, 5hmC modification has been detected at low levels in plant genomes, but there is no detailed information about its distribution and function.

Current bisulphite conversion-based methodology fails to distinguish 5hmC from 5mC because 5hmC, similar to 5mC, does not undergo deamination during bisulphite treatment (Huang et al., 2010; Jin et al., 2010; Yu et al., 2012). With the development of 5hmC detection methods, such as the use of 5hmC-specific antibodies and chemical labelling (Song et al., 2011), it is feasible to profile the genome-wide distribution of 5hmC. Rice (Oryza sativa L.) is one of the most important world crops with a relatively small genome size (430Mb) (Zhao and Zhou, 2012). The availability of the complete genome sequence of japonica (cv Nipponbare) (Goff et al., 2002) and indica (cv 93-11) (Yu et al., 2002) subspecies of rice has enabled the high-resolution mapping of epigenetic modifications in chromosomes and the investigation of their regulatory roles in transcriptional activity. In this study, 5hmC modification in DNA isolated from different tissues and cultivars of rice was detected and quantified using dot-blot and LC-MS/MS assays. The genome-wide distribution of 5hmC was also explored using a sensitive chemical labelling method (GLIB: glucosylation, periodate oxidation, and biotinylation) followed by NGS (next-generation sequencing). Together with transcriptome data obtained from RNA-sequencing (RNA-Seq), the potential epigenetic roles of 5hmC were further explored.

Materials and methods

Material samples

The material used in this study was Chinese superhybrid rice LYP9, an elite hybrid most widely grown in China, and its paternal lines 93-11, an indica rice (Oryza sativa ssp. indica) variety, and maternal lines PA64s. All three cultivars were planted in a test field of the China National Hybrid Rice R&D Center in Hunan Province in 2011. During the period of ear development, whole panicles at the same developmental stage were sampled using a magnifying lens based on their length (3–4mm) after removing leaves. Four types of tissues, namely leaf, root, stem, and panicle, of 93-11 were collected from a test field in Beijing. Materials were immersed in liquid nitrogen rapidly then stored in a –80 °C freezer.

5hmC detection by immuno-dot-blot assay and LC-MS/MS/MS

All genomic DNAs were extracted using a DNasecure Plant Kit (DP320-03) (TIANGEN, China). Genomic DNA from panicle and leaf was denatured at 95 °C for 10 min, then immediately chilled on
ice for 5 min. The dot-blot analysis was performed on a bio-Dot Apparatus (#170–6545, Bio-Rad). The completely dried charged nylon-based membrane was baked at 80 °C for 2 h, cross-linked with 254 nm UV light for 10 min, and blotted briefly with 5% non-fat milk at room temperature for 1.5 h. The primary rabbit anti-5-hydroxymethylcytosine antibody (1:10k, #39769, Amtive Motif) was applied to the membrane and incubated at room temperature for 1 h or at 4 °C overnight, followed by peroxidase-conjugated anti-rabbit IgG secondary antibody. The signal was visualized by using ECL-Plus (Amersham Pharmacia Biotech).

Genomic DNA (6 μg) isolated from rice tissues was digested using a cocktail of four enzymes, and the 5hmC target nucleoside was enriched on a Beckman HPLC system and subjected to LC-MS/MS/MS analysis, following previously described procedures (Liu et al., 2013).

5hmC-specific chemical labelling and affinity purification

Purified genomic DNA was sonicated into short fragments (200–300 bp) by Covaris DNA shearing with microTUBEs according to the manufacturer’s instructions. The 5hmC labelling reaction was performed in a 75 μl solution containing 50 mM HEPES (pH 7.9), 250 mM MgCl₂, 100 μM UDP-6-N₂-Glu, and 80 U of β-GT, and then incubated at 37 °C for 2 h. After the reaction was complete, the DNA substrates were purified and the buffer exchanged in H₂O via Bio-Rad Micro Bio-Spin according to the manufacturer’s instructions. The click chemistry was performed with the addition of 150 μM biotin into the DNA solution and incubation at 37 °C for 2 h. The DNA samples were then purified by Invitrogen DynabeadsMyOne™ Streptavidin C1.

Sequencing of 5hmC-enriched genomic DNA

5hmC-enriched genomic DNA libraries were generated following the Illumina protocol for ‘Preparing Samples for ChIP sequencing of DNA’. The above-mentioned 5hmC-enriched DNA fragments (50 ng) were ligated with Illumina library adaptors following by a 16 cycle PCR amplification with multiplex PCR primers. The PCR products of 200–300 bp were gel purified and then quantified on an Agilent 2100 BioAnalyzer and by quantitative PCR. The 100 bp single-end sequencing was performed on an Illumina HiSeq 2000 system to obtain 5hmC-enriched DNA fragment sequences.

Read mapping and 5hmC modification peak calling

The adaptor sequences were removed with the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Reads of <25 nucleotides in length or containing ambiguous nucleotides were discarded. The remaining clean reads were aligned to the rice genome sequences (O. sativa subsp. japonica version 6.1) with up to two mismatches allowed by BWA (Burrows-Wheeler Alignment Tool) (Li and Durbin, 2009). All non-redundant uniquely mapped reads were used for peak calling using MACS (P-value < 1 × 10⁻⁷) (Zhang et al., 2008). Annotation of 5hmC peaks was performed based on gene annotation in the TIGR 6.1 database. Genomic distribution of 5hmC peaks was plotted using the SVG program.

RNA-Seq library construction and sequencing

Total RNAs from panicles 3–4 mm in length for LYP9 and its parent lines were isolated using Trizol reagent (Invitrogen). Poly(A)⁺ mRNAs were enriched, and strand-specific RNA-Seq libraries were constructed and sequenced using the Applied Biosystems SOLiD sequencing. The sequenced reads of 38 bp length were aligned to reference sequences including the genome (O. sativa subsp. japonica, TIGR version 6.1) and the exon–exon junction sequences by SOLiD Corona Lite software, where the junction sequences were constructed by joining all possible pairs of non-redundant exons that belong to the same gene. A recursive mapping strategy was adopted to improve the usable sequenced reads information. The sequenced raw reads were mapped simultaneously to reference sequences with at most five, four, three, three, and two colour-space mismatches for 50, 45, 40, 35, 30, and 25 mers, respectively. The 50-mer reads were first mapped to the genome and junction sequences with five mismatches. Reads that failed to be mapped were progressively trimmed off, five bases at a time, from the 3’ end, and mapped to the genome again until a match was found (unless the read had been trimmed by <25 bases). All mapping was carried out in colour-space; reads mapping to multiple genomic locations were discarded. All uniquely mapped reads were used for gene expression quantification with wapRNA (Zhao et al., 2011). RPMK values (measured in reads per kb of exon per million mapped sequence reads, which is a normalized measure of exonic read density) were defined by calculating the read coverage of annotated introns and exons in the rice genome for each sample, an empirical cut-off value was arbitrarily determined and used as a cut-off for detection of gene expression in each mRNA-Seq sample (He et al., 2010).

Validation of 5hmC peaks by real-time PCR

Genomic DNA was extracted using the DNA Secure Plant Kit (Tiangen, China) according to the manufacturer’s instructions. DNA was randomly sheared to an average length of 200–500 bp by sonication. A hMeDIP (hydroxymethylcytosine DNA immunoprecipitation) assay was then performed using the hMeDIP Kit (Active Motif, USA) according to the manufacturer’s instructions. The immunoprecipitated products and 10% of the original input DNA were purified with the Chromatin IP DNA Purification Kit (Active Motif) in parallel. The purified DNA was analysed by real-time PCR on the CFX96™ Real-time system (Bio-Rad, USA) using SYBR® Select Master Mix (Life Technologies, USA). The enrichment of 5hmC in selected peaks was calculated by normalizing the amount of immunoprecipitated DNA to input DNA. Each sample was run in triplicate and the average was plotted. A list of the primers used is provided in Supplementary Table S1 available at JXB online; the online tool Primer3 (http://www.simgene.com/Primer3) was used to design real-time PCR primers.

Accession numbers

The high-throughput sequence data reported herein have been deposited in the Gene Expression Omnibus (GEO) with accession number GSE37242.

Results

A low level of 5hmC modification was detected and quantified in rice DNA, and its level exhibited large variation in different tissues

It has been reported that there is a limited amount of 5hmC in the Arabidopsis genome (Terragni et al., 2012; Yao et al., 2012). In this study, the existence of 5hmC in DNA of rice was detected by dot-blot assay using an anti-5hmC antibody and an LC-MS/MS/MS method. The 5hmC was first detected in leaves at the flowering stage and in panicles at the booting stage in the indica rice cultivar 93-11 (O. sativa L. ssp. indica) by a dot-blot assay. The immuno-dot-blot assay results (Fig. 1A) revealed low but measurable levels of 5hmC, and the abundance of 5hmC differed between the panicle and leaf. It was also verified that this antibody specifically immunoprecipitated oligonucleotides containing 5hmC but not the same sequence containing cytosine or 5mC (Fig. 1A). Using a previously published LC-MS/MS/MS method (Liu et al.,
5hmC was successfully detected in genomic DNA from rice leaf at a level of $1.39 \pm 0.16$ modifications per million nucleosides (Fig. 1B), thereby directly verifying the presence of 5hmC in rice DNA.

The contents of 5hmC in different rice tissues were further quantified using the same LC-MS/MS/MS method (Liu et al., 2013). The levels of 5hmC were $1.39 \pm 0.16$ and $2.17 \pm 0.03$ modifications per million nucleosides in leaf and panicle, respectively, suggesting the tissue-specific occurrence of 5hmC in rice (Fig. 1C). This tissue-dependent distribution of 5hmC in rice is reminiscent of the findings in mammals (Kinney et al., 2011), where the levels varied from 0.01% to 0.7% per total nucleosides (Kriaucionis and Heintz, 2009; Globisch et al., 2010; Szwagierczak et al., 2010; Williams et al., 2012). It is obvious that the content of 5hmC in rice was significantly lower than that reported in mammals. In summary, the large intertissue variation in global 5hmC levels indicated that tissue type was a major modifier of 5hmC content, which is consistent with what has been observed in mammals (Li and Liu, 2011; Nestor et al., 2012).

Genome-wide profiling in three different rice cultivars revealed their specificity and conservation of 5hmC modification

To investigate whether 5hmC is a universal epigenetic mark in rice, 5hmC modification was compared among three different rice cultivars: the superhybrid rice LYP9, and its parental lines 93-11 and PA64s. To minimize systematic bias, immature panicles 3–4 mm in length were simultaneously collected from the superhybrid rice group cultivars grown in a test field of the China National Hybrid Rice R&D Center in Hunan. DNA was extracted from each cultivar and an equal amount
of DNA was used for the dot-blot assay. The dot-blot assay results showed large intercultivar variation in global 5hmC levels, which revealed 5hmC as a prevalent epigenetic modification in rice (Fig. 2A). By qualitatively comparing the levels among the three cultivars, it was found that 93-11 had the highest level of 5hmC. In addition, the level of 5hmC in LYP9 was more similar to that in 93-11 than to that in PA64s, and the former two lines exhibited significantly higher levels of 5hmC than did PA64s.

To obtain a global view of 5hmC modification in rice, high-throughput ChIP-seq (chromatin immunoprecipitation sequencing) analysis was performed using the Illumina HiSeq 2000 platform on 5hmC-specific chemically labelled and affinity-enriched DNA from immature panicles 3–4 mm in length from the superhybrid rice LYP9 and its parental lines (93-11 and PA64s). An in-depth single-end ChIP-seq analysis was conducted and >22 million reads of 100 bp in length were acquired for each sample. The total length of the reads was >2.4 Gb, representing ~7-fold the size of the rice genome. After removing adaptors, all clean reads from the three cultivars were aligned onto the TIGR O. sativa subsp. japonica genome reference (version 6.1) (Yuan et al., 2003; Ouyang et al., 2007), and it was found that ~64.6, 68.97, and 58.03% of the reads could be mapped to the reference genome for 93-11, PA64s, and LYP9, respectively, and ~74, 73.27, and 74.48% of the mapped reads could be uniquely aligned (Table 1). Using a model-based analysis of the ChIP-seq (MACS) software (Zhang et al., 2008), 3970, 2396, and 3489 significantly enriched genomic regions (peaks) associated with DNA 5hmC were identified in 93-11, PA64s, and LYP9, respectively (Fig. 2B). The average lengths of the defined peaks were 142, 141, and 139 bp, and the average fold enrichments were ~10-, 9-, and 10-fold for 93-11, PA64s, and LYP9, respectively. There are ~30 distinct reads in each defined peak (Supplementary Fig. S1 at JXB online), and all of the identified 5hmC peaks are listed in Supplementary Tables S2–S4 for 93-11, LYP9, and PA64s, respectively. All peaks in each cultivar were further divided into four classes based on their statistical significance (i.e. P-value); class I by a P-value between 1.0 × 10^{-5} and 1.0 × 10^{-10}, class II by a P-value between 1.0 × 10^{-10} and 1.0 × 10^{-15}, class III by a P-value between 1.0 × 10^{-15} and 1.0 × 10^{-25}, and class IV by a P-value <1.0 × 10^{-25} (Fig. 2B; Table 2). Three 5hmC peak regions were randomly selected and the existence of 5hmC modification was confirmed in the panicle, leaf, sheath, and root of 93-11. The fold 5hmC enrichment results revealed that the hydroxymethylation level was different among various tissues at selected 5hmC peaks (Supplementary Fig. S2).

The cultivar-specific 5hmC profiles were investigated both in the complete genome (Supplementary Fig. S3 at JXB
Table 1. Statistics of sequenced reads and enriched 5hmC region number

<table>
<thead>
<tr>
<th>Class</th>
<th>93-11 Total reads</th>
<th>PA64s Clean reads</th>
<th>LYP9 Mapped reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>22,804,853</td>
<td>2,139,202</td>
<td>14,733,064</td>
</tr>
<tr>
<td>II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. 5hmC peaks in each cultivar divided into four classes

<table>
<thead>
<tr>
<th>Class</th>
<th>93-11</th>
<th>LYP9</th>
<th>PA64s</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1071</td>
<td>866</td>
<td>579</td>
</tr>
<tr>
<td>II</td>
<td>924</td>
<td>911</td>
<td>544</td>
</tr>
<tr>
<td>III</td>
<td>991</td>
<td>850</td>
<td>669</td>
</tr>
<tr>
<td>IV</td>
<td>984</td>
<td>862</td>
<td>604</td>
</tr>
</tbody>
</table>

online) and at specific loci (Fig. 2C) from the comparative analysis of the sequencing data. There were 40.88% peaks (1,623 out of 3,970) specific in the paternal line 93-11, 37.90% (908 out of 2,396) in the maternal line PA64s, and 31.96% (1,115 out of 3,489) in the superhybrid rice LYP9. More than half of the 5hmC peaks were shared by two or three cultivars, while 1161 peaks were shared by all three cultivars (Fig. 2D), suggesting the similarity and conservation of cytosine hydroxymethylation in different rice cultivars. Further investigation revealed that 1036 peaks of 5hmC were common to LYP9 and its paternal line 93-11, a value which was much greater than the number shared between LYP9 and its maternal line PA64s (177 peaks of 5hmC). Previous studies showed that in both phenotype and gene expression in the panicle, the superhybrid F1 (LYP9) displays a higher level of similarity to the paternal line (93-11) (Song et al., 2007; Ge et al., 2008). The greater similarities of 5hmC content and gene expression between LYP9 and 93-11 indicated that further studies would be needed to explore whether 5hmC modification could regulate gene expression in rice, similar to its role in animals (Ficz et al., 2011).

5hmC modification tends to be located in transcriptionally silent TE genes

Transposable element (TE) genes are known to be enriched in heterochromatin compared with euchromatin (Copenhaver et al., 1999; Turcotte et al., 2001). For example, there is a ~3-fold enrichment of TE genes in the heterochromatin of chromosome 4 of rice when compared with the euchromatin (Yan and Jiang, 2007). To examine the roles of 5hmC in TE genes, the study focused on those peaks that were targeted to genic regions, and it was found that 79.42, 82.77, and 81.67% of 5hmC peaks were located in TE genes in 93-11, PA64s, and LYP9, respectively (Fig. 5A); these values are significantly different from the random data set (P-value <2.2 x 10^-16; Fisher’s exact test) (Supplementary Fig. S8 at JXB online). By surveying the distributions of targets for 5hmC peaks in each class, it was found that the percentages of TE and non-TE were consistent in classes II, III, and IV, but not in class I (Fig. 5B). To increase the reliability of the following results, the 5hmC peaks of class II, III, and IV were defined as being the more...
reliable peaks (MRPs). Based on the TIGR gene annotation, out of 41,415 (72%) non-TE genes, 187 genes containing 228 peaks of 5hmC (10.57% of genic-related peaks) were detected. In contrast, out of 16,209 (28%) TE genes, 1133 genes containing 1930 peaks of 5hmC (89.43% genic-related peaks) were detected in 93-11. Similar results were obtained for PA64s and LYP9. From the chromosome-level statistics, it was also obvious that 5hmC peaks were mostly located in TE genes (Supplementary Fig. S3), a result which was significantly different from the random data set (P-value < 2.2 × 10⁻¹⁶; Fisher’s exact test) (Supplementary Fig. S9), especially on chromosomes 4 and 10. The close relationship between 5hmC modification and TE genes suggested that 5hmC may have a role in influencing the activity of TE genes.

To clarify the influence of 5hmC in TE genes, gene expression information was obtained from RNA-Seq technology.
and expressed genes were defined by RPKM >1.99, 1.83, and 1.7 for 93-11, PA64s, and LYP9, respectively (Supplementary Fig. S10 at JXB online), which were calculated following the methods reported by He et al. (2010). The correlation analysis between 5hmC and RPKM revealed that only a small portion (e.g. 1.52%, 20 out of 1320 in 93-11) of genes with MRPs were actively expressed, and most genes (98.48%, 1300 out of 1320) with MRPs were transcriptionally silent (Table 3; Fig. 5C; typical examples are displayed in Supplementary Fig. S11). Among the 20 actively expressed MRP target genes, three (15%) were TE genes and 17 (85%) were non-TE genes. On the other hand, among the 1300 inactive MRP target genes, 1130 (86.92%) and 170 (13.08%) were TE genes and non-TE genes, respectively (Table 3; Fig. 5D). The correlation results indicated that 5hmC modification may play a different role in TE and non-TE genes, in concert with its dual roles reported in mammals (Wu et al., 2011). Among the 1133 TE genes with MRPs in 93-11, three (0.26%) genes were actively expressed, whereas 1130 (99.74%) were not expressed, and similar results were also observed in PA64s and LYP9 (Table 3). If 5hmC modification does have a role in TE genes, it might repress their activity.
By further investigating 5hmC in different TE gene families, it was found that 5hmC peaks were significantly ($P=1.03 \times 10^{-7}$; Fisher’s exact test) enriched in retrotransposon genes, especially in the Ty3-gypsy family of 93-11 (Fig. 5E; Tables 4, 5), which was very much in agreement with LYP9 and PA64s (Tables 4, 5; Supplementary Fig. S12 at JXB online). All these results suggested that 5hmC may have roles in its direct target genes, especially in retrotransposon genes. Additionally, the Gene Ontology (GO) functional analysis indicated that MRP target genes were significantly enriched in the functional categories of chromatin assembly and chromatin organization ($P$-value < 0.01, Fisher’s exact test) (Fig. 5F), suggesting that 5hmC may have effect on chromosome structure.

**Discussion**

Recent studies have revealed that 5hmC is a bona fide constituent of mammalian DNA, owing to its presence at especially high levels in embryonic stem cells and Purkinje neurons (Globisch et al., 2010; Szulwach et al., 2011). However, the existence of 5hmC modification in plants is controversial. Although some studies have detected low levels of 5hmC in plants using various methods (such as dot-blot and mass spectrometry) (Terragni et al., 2012; Yao et al., 2012; Liu et al., 2013; Moricová et al., 2013), researchers are still doubtful about its existence and biological function because of its very low abundance and possible formation from endogenous reactive oxygen species (Liu et al., 2013; Jang et al., 2014). In this study, 5hmC modification in rice DNA was detected by dot-blot assay and its levels were quantified in various rice tissues using LC-MS/MS/MS. More importantly, thousands of 5hmC peaks were identified in panicles from three different rice cultivars using a GLIB method combined with NGS technology. The high-resolution maps of 5hmC in DNA of three different rice cultivars revealed a large number of cultivarspecific and conserved 5hmC-enriched regions, supporting the conclusion that 5hmC is an authentic DNA modification in rice. It is known that TETs are the writer and primary eraser of 5hmC in mammals (Tahiliani et al., 2009; Zhu, 2009; Law and Jacobsen, 2010; Song and He, 2011) and UHRF2 is a reader of 5hmC in mammals (Spruijt et al., 2013; Zhou et al., 2014). However, there is no functional counterpart of TETs and UHRF2 in the rice genome based on the present sequence comparative analysis. More studies are needed to explore the writer, eraser, and reader of 5hmC in plants.

The number of 5hmC peaks identified here is slightly higher than the number obtained from LC-MS/MS/MS. This inconsistency may be due to (i) the modification of 5mC to 5hmC may not be 100% at some loci; (ii) false-positive peaks introduced by the lack of specificity of the 5hmC antibody and/or bioinformatics analysis; or (iii) the inherent difference between the two techniques. In this vein, the conversion of 5mC to 5hmC at some loci may not be quantitative. Accordingly, the high confidence peaks (HCPs) were defined by setting the $P$-value to < 10$^{-25}$ which is supported by both LC-MS/MS/MS and sequencing.

The discovery of 5hmC, 5fC, and 5caC in mammals has revealed an active demethylation pathway by a successive oxidation process (Zhu, 2009; Gong and Zhu, 2011; Song et al., 2013). In plants, 5hmC may be involved in passive demethylation during cell division. First, it has been reported that *Arabidopsis thaliana* VIM1 protein could recognize 5hmC *in vitro*, and 5hmC at a CpG site may trigger VIM-mediated passive loss of cytosine methylation *in vivo* during *Arabidopsis* DNA replication (Yao et al., 2012). Second, unlike in mammals, the plant genome does not encode TET homologues that oxidize 5mC to 5hmC (Tahiliani et al., 2009; Ko et al., 2010; Wu and Zhang, 2011). Thirdly, plants possess DNA glycosylases that can cleave

<table>
<thead>
<tr>
<th>Gene number with MRPs</th>
<th>93-11</th>
<th>PA64s</th>
<th>LYP9</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE genes</td>
<td>1320</td>
<td>807</td>
<td>1196</td>
</tr>
<tr>
<td>Non-TE genes</td>
<td>1133</td>
<td>688</td>
<td>1037</td>
</tr>
<tr>
<td>Expressed genes</td>
<td>20</td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td>TE genes</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Non-TE genes</td>
<td>17</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>Repressed gene number</td>
<td>1300</td>
<td>794</td>
<td>1174</td>
</tr>
<tr>
<td>TE genes</td>
<td>1130</td>
<td>688</td>
<td>1034</td>
</tr>
<tr>
<td>Non-TE genes</td>
<td>170</td>
<td>108</td>
<td>140</td>
</tr>
</tbody>
</table>

Table 4. A summary of the numbers of MRPs targets, whole-genome and randomly simulated retrotransposon and transposon genes, and the *P*-values obtained from Fisher’s exact test.

<table>
<thead>
<tr>
<th>No. of TEs across the whole genome</th>
<th>Retrotransposon</th>
<th>Transposon</th>
<th>Fisher’s exact test</th>
</tr>
</thead>
<tbody>
<tr>
<td>93-11</td>
<td>12 295</td>
<td>164</td>
<td>8.63E-15$^a$</td>
</tr>
<tr>
<td>No. of MRPs target TEs</td>
<td>969</td>
<td>161</td>
<td>1.03E-07$^b$</td>
</tr>
<tr>
<td>No. of TEs randomly simulated</td>
<td>487</td>
<td>151</td>
<td>1.85E-13$^a$</td>
</tr>
<tr>
<td>LYP9</td>
<td>886</td>
<td>146</td>
<td>3.42E-07$^a$</td>
</tr>
<tr>
<td>No. of MRPs target TEs</td>
<td>440</td>
<td>105</td>
<td>2.75E-08$^a$</td>
</tr>
<tr>
<td>No. of TEs randomly simulated</td>
<td>583</td>
<td>101</td>
<td>1.61E-04$^a$</td>
</tr>
<tr>
<td>PA64s</td>
<td>307</td>
<td>105</td>
<td>3.42E-07$^a$</td>
</tr>
</tbody>
</table>

$^a$ The *P*-value was calculated using the number of whole-genome TEs and the number of retrotransposon and transposon MRPs target TEs.

$^b$ The *P*-value was calculated using the number of randomly simulated TEs and the number of retrotransposon and transposon MRPs target TEs.
5mC directly from the genome (Gehring et al., 2009; Law and Jacobsen, 2010; Jang et al., 2014). Finally, although both DME and ROS1 have glycosylase activities for 5hmC in vitro, DME exhibits higher activity toward 5mC than 5hmC, and the rate of DME-mediated excision of 5mC is 2.6-fold higher than that of 5hmC (Jang et al., 2014). Moreover, 5hmC is one of the numerous substrates of the DME/ROS1 family, just as most DNA glycosylases have a wide range of substrate specificity (Dizdaroglu et al., 1993; Bulychev et al., 1996; Schärer and Jiricny, 2001). In summary, the low abundance of 5hmC in rice may be an intermediate of passive demethylation in local regions.

5hmC is particularly enriched in gene body regions (2190, 1938, and 1326 peaks in the gene body for 93-11, LYP9, and PA64s, respectively), which is similar to the findings in animals (Song et al., 2011; Stroud et al., 2011; Szulwach et al., 2011; Xu et al., 2011). Moreover, most of those 5hmC peaks (1800, 1666, and 1114 for 93-11, LYP9, and PA64s, respectively) were in TE genes which were mainly located in heterochromatin. To increase the reliability of the results, targets of those MRPs were analysed by correlating with RNA-Seq data. It was found that only ~1.52% (20 out of 1320) of genes with MRPs were actively transcribed, and most of these expressed genes (17 out of 20) were non-TE genes. Previous studies of 5hmC in mouse embryonic stem cells revealed that 5hmC occurs on both transcriptionally active and inactive genes, and has dual functions in transcriptional regulation (Wu et al., 2011). The present results are partially in agreement with this finding at a genomic scale and extend to the plant kingdom. As is known, methylation is significantly enriched in repeated sequences such as TEs to preserve genome stability by suppressing TE activity (Zhu, 2009; Paszkowski and Whitham, 2001; Kakutani et al., 2004; Inagaki and Kakutani, 2010; Zhao and Zhou, 2012). Further work is needed to study whether 5hmC has a role in maintaining genome stability by co-operating with methylation to repress the activity of TEs.

In summary, 5hmC is a DNA modification of very low abundance in the plant genome. The detection of 5hmC modification in rice, as well as its intercultivar variation and enrichment in TE genes, will stimulate future investigations about its roles in development, maintenance of the repressive status of TEs, and cytosine demethylation.

### Supplementary data

Supplementary data are available at *JXB* online.

**Figure S1.** Statistical properties of regions enriched for 5hmC modification.

**Figure S2.** Real-time PCR validations for three 5hmC peaks among four tissues of 93-11.

**Figure S3.** Comparative whole-genome distribution of 5hmC peaks among the three cultivars.

**Figure S4.** Genome-wide distribution of those detected 5hmC peaks compared with a 1000 times randomly simulated data set in 93-11 (A), LYP9 (B), and PA64s (C), respectively.

**Figure S5.** Distributions of 5hmC peaks in each class for LYP9 (A) and PA64s (B) in the whole genome.

**Figure S6.** Average number of 5hmC peaks per million bases in heterochromatin and euchromatin of chromosomes 4 and 10 compared with a random data set.

**Figure S7.** Enrichment of 5hmC in heterochromatin (HC) over euchromatin (EC) of chromosomes 4 and 10 in each class for three cultivars.

<table>
<thead>
<tr>
<th>Retrotransposon</th>
<th>Ty1-copia</th>
<th>Ty3-gypsy</th>
<th>LINE</th>
<th>SINE</th>
<th>Centromere-specific</th>
<th>Unclassified</th>
<th>$\chi^2$ test</th>
</tr>
</thead>
<tbody>
<tr>
<td>93-11 No. of TEs across the whole genome</td>
<td>798</td>
<td>3782</td>
<td>114</td>
<td>12</td>
<td>638</td>
<td>6951</td>
<td></td>
</tr>
<tr>
<td>93-11 No. of MRP target TEs</td>
<td>16</td>
<td>536</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>401</td>
<td>2.20E-16</td>
</tr>
<tr>
<td>93-11 No. of TEs randomly simulated</td>
<td>33</td>
<td>164</td>
<td>4</td>
<td>0</td>
<td>19</td>
<td>266</td>
<td>NA</td>
</tr>
<tr>
<td>93-11 No. of MRP target TEs</td>
<td>18</td>
<td>522</td>
<td>1</td>
<td>0</td>
<td>13</td>
<td>332</td>
<td>2.20E-16</td>
</tr>
<tr>
<td>93-11 No. of TEs randomly simulated</td>
<td>30</td>
<td>148</td>
<td>4</td>
<td>0</td>
<td>17</td>
<td>241</td>
<td>NA</td>
</tr>
<tr>
<td>LYP9 No. of TEs across the whole genome</td>
<td>9</td>
<td>355</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>211</td>
<td>2.20E-16</td>
</tr>
<tr>
<td>LYP9 No. of MRP target TEs</td>
<td>9</td>
<td>355</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>211</td>
<td>2.20E-16</td>
</tr>
<tr>
<td>LYP9 No. of TEs randomly simulated</td>
<td>21</td>
<td>104</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>167</td>
<td>NA</td>
</tr>
</tbody>
</table>

**Transposon**

<table>
<thead>
<tr>
<th>Ac/Ds</th>
<th>CACTA, En/Spm</th>
<th>Mutator</th>
<th>Mariner</th>
<th>ping/pong/SNOOPY</th>
<th>Unclassified</th>
<th>$\chi^2$ test</th>
</tr>
</thead>
<tbody>
<tr>
<td>93-11 No. of TEs across the whole genome</td>
<td>2709</td>
<td>6156</td>
<td>114</td>
<td>12</td>
<td>638</td>
<td>6951</td>
</tr>
<tr>
<td>93-11 No. of MRP target TEs</td>
<td>16</td>
<td>536</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>401</td>
</tr>
<tr>
<td>93-11 No. of TEs randomly simulated</td>
<td>33</td>
<td>164</td>
<td>4</td>
<td>0</td>
<td>19</td>
<td>266</td>
</tr>
<tr>
<td>LYP9 No. of TEs across the whole genome</td>
<td>9</td>
<td>355</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>211</td>
</tr>
<tr>
<td>LYP9 No. of MRP target TEs</td>
<td>9</td>
<td>355</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>211</td>
</tr>
<tr>
<td>LYP9 No. of TEs randomly simulated</td>
<td>21</td>
<td>104</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>167</td>
</tr>
</tbody>
</table>

NA indicates that the R software could not calculate the P-value.

* The P-value was calculated using the number of whole-genome TEs and the number of MRP target TEs.

* The P-value was calculated using the number of randomly simulated TEs and the number of MRP target TEs.

Table 5. A summary of the numbers of MRPs targets, the whole-genome and randomly simulated retrotransposon and transposon genes in different species, and the P-values obtained from the $\chi^2$ test.
**Figure S8.** Distributions of 5hmC peaks in TE genes and non-TE genes compared with a random data set for 93-11 (A), PA64s (B), and LYP9 (C), respectively.

**Figure S9.** Distributions of 5hmC peaks in TE genes and non-TE genes in every chromosome of 93-11 (A), PA64s (B), and LYP9 (C) compared with a random data set, respectively.

**Figure S10.** Detection of gene expression in rice by mRNA-Seq.

**Figure S11.** Typical examples for expressed and silent genes with 5hmC peaks in read-level among three cultivars.

**Figure S12.** The numbers of retrotransposons and transposons with MRPs for LYP9 (A) and PA64s (B), respectively.

**Table S1.** A list of the primers used for real-time PCR.

**Table S2.** Identified 5hmC peaks for 93-1.

**Table S3.** Identified 5hmC peaks for LYP9.

**Table S4.** Identified 5hmC peaks for PA64s.

**Acknowledgments**

We would like to thank Dr. Ming Zhou and Dr. Xiao-feng Cao for assistance with obtaining rice materials, Dr. Li-huang Zhu for providing RNA-Seq data, and Dr. Yun-gui Yang for helpful discussion. This work was supported by the National Natural Science Foundation of China (30900831, 31271372 to SS); Beijing Nova Program (Z12110500251200 to SS); Strategic Priority Research Program of the Chinese Academy of Science (XDA08020102 to SS); and the National Institutes of Health of the US [R01 CA101864 to YW]. No conflict of interest is declared.

**References**


Erdmann RM, Souza AL, Clish CB, Gehring M. 2015. 5-hydroxymethylcytosine is not present in appreciable quantities in Arabidopsis DNA. G3: Genes(Genomes)Genetics 5, 1–8.


