Expansion and Functional Divergence of Jumonji C-Containing Histone Demethylases: Significance of Duplications in Ancestral Angiosperms and Vertebrates

Shengzhan Qian, Yingxiang Wang, Hong Ma*, and Liangsheng Zhang*

State Key Laboratory of Genetic Engineering and Collaborative Innovation Center of Genetics and Development, Ministry of Education Key Laboratory of Biodiversity Science and Ecological Engineering and Institute of Biodiversity Sciences, Institute of Plant Biology, Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai 200438, China (S.Q., Y.W., H.M.); Institute of Biomedical Sciences, Fudan University, Shanghai 200032, China (H.M.); and Department of Bioinformatics, School of Life Sciences and Technology, and Advanced Institute of Translational Medicine, Tongji University, Shanghai 200092, China (L.Z.)

ORCID IDs: 0000-0001-8717-4422 (H.M.); 0000-0003-1919-3677 (L.Z.).

Histone modifications, such as methylation and demethylation, are crucial mechanisms altering chromatin structure and gene expression. Recent biochemical and molecular studies have uncovered a group of histone demethylases called Jumonji C (JmjC) domain proteins. However, their evolutionary history and patterns have not been examined systematically. Here, we report extensive analyses of eukaryotic JmjC genes and define 14 subfamilies, including the Lysine-Specific Demethylase3 (KDM3), KDM5, JMJ6, Putative-Lysine-Specific Demethylase11 (PKDM11), and PKDM13 subfamilies, shared by plants, animals, and fungi. Other subfamilies are detected in plants and animals but not in fungi (PKDM12) or in animals and fungi but not in plants (KDM2 and KDM4). PKDM7, PKDM8, and PKDM9 are plant-specific groups, whereas Jumonji, AT-Rich Interactive Domain2, KDM6, and PKDM10 are animal specific. In addition to known domains, most subfamilies have characteristic conserved amino acid motifs.

Whole-genome duplication (WGD) was likely an important mechanism for JmjC duplications, with four pairs from an angiosperm-wide WGD and others from subsequent WGDs. Vertebrates also experienced JmjC duplications associated with the vertebrate ancestral WGDs, with additional mammalian paralogs from tandem duplication and possible transposition. The sequences of paralogs have diverged in both known functional domains and other regions, showing evidence of selection pressure. The increases of JmjC copy number and the divergences in sequence and expression might have contributed to the divergent functions of JmjC genes, allowing the angiosperms and vertebrates to adapt to a great number of ecological niches and contributing to their evolutionary successes.

Chromatin-based regulation is an important mechanism of modulating eukaryotic chromatin structure and gene expression by altering DNA and histone modifications rather than changing DNA sequences. The eukaryotic chromatin contains a group of highly conserved proteins called histones, including the core histones H2A, H2B, H3, and H4, as well as the linker histone H1. The core histones, with two copies each, form an eight-subunit complex, which is wrapped by 146 bp of DNA to form the nucleosome, the basic unit of chromatin. In each nucleosome, the hydrophobic C-terminal regions of the eight subunits occupy the interior, whereas the hydrophilic N-terminal regions (the tails) extend outward (Luger et al., 1997). The covalent histone modifications occur on the tails of the core histone proteins and encode epigenetic information that can be passed through mitosis, sometimes even meiosis, and alter chromatin structure and modulate genomic functions. Specifically, histone modifications include methylation, acetylation, phosphorylation, ubiquitination, and sumoylation. Chromatin structure can be regulated by three classes of proteins: DNA methylase and demethylases, chromatin remodelers that regulate nucleosome positioning, and enzymes for histone modifications.

Among various histone modifications, the role of methylation varies in different species (Feng et al., 2010; Liu et al., 2010) but is relatively stable and suited for the transmission of epigenetic information (Strahl and Allis, 2000), sometimes even transgenerationally.
as supported by a recent study in Arabidopsis (*Arabidopsis thaliana*) showing that the vernalized state can be partially transgenerationally inherited due to a defect in the reduction of histone H3 lysine-27 trimethylation (H3K27me3) levels at FLOWERING LOCUS C (FLC; Crevillén et al., 2014). Histone methylation occurs on Arg and Lys residues and is involved in a wide range of biological processes, including gene expression, chromatin structure, dosage compensation, and epigenetic memory (Martin and Zhang, 2005). A Lys residue can be monomethylated, dimethylated, or trimethylated; however, an Arg residue can only be monomethylated or dimethylated. Different histone modifications regulate distinct functional outcomes within an epigenetic marking system: methylation at histone H3 Lys-4 (H3K4), H3K36, and H3K79 is correlated with higher gene expression, whereas methylation at H3K9, H3K27, and H4K20 is associated with lower gene expression, hence often termed activating or repressing marks, respectively, although the causal relationship between histone modification and transcriptional activity is still elusive (Martin and Zhang, 2005; Henikoff and Shilatifard, 2011; Dong and Weng, 2013).

Histone methylation has been regarded as an irreversible modification for a long time, because of the stable nature of the carbon-nitrogen bond; this idea is also supported by the similar half-lives of histone fractions and methyl Lys and methyl Arg marks (Byvoet et al., 1972). In 2004, the discovery of the first histone demethylase, known as LYSINE-SPECIFIC DEMETHYLASE1 (LSD1), provided experimental evidence for enzymatic demethylation (Shi et al., 2004). The single-copy LSD1 mediates oxidative demethylation on monomethylated or dimethylated H3K4 and/or H3K9 but not trimethylated H3K4 (Shi et al., 2004). Subsequently, a second and larger class of demethylases containing a Jumonji C (JmjC) domain was identified (Tsukada et al., 2006). Unlike LSD1, the JmjC proteins do not require a protonated nitrogen and can also reverse the trimethylated histone Lys state (Shi and Whetstine, 2007). Five amino acid residues within predicted cofactor-binding sites in the JmjC domain are conserved and important for enzymatic activity (Klose et al., 2006). Among these, three residues bind to the Fe (II) cofactor and two other residues are required for α-ketoglutarate (α-KG) binding.

Several plant *JmjC* genes are known to have important functions in regulating development and environmental responses (Liu et al., 2010; Chen et al., 2011; Kooistra and Helin, 2012). For example, POLYCOMB REPRESSIVE COMPLEX2, a conserved and key transcriptional regulator in animals and plants, has been demonstrated to have H3K27me3 affinity (Cao et al., 2002). In Arabidopsis, Relative of Early Flowering6 (REF6)/AtPKDM9A and Early Flowering6 (ELF6)/AtPKDM9B are closely related *JmjC* paralogs and have H3K27me3 demethylase activities (Lu et al., 2011a; Crevillén et al., 2014) but have different roles in the regulation of flowering time (Noh et al., 2004). The elf6 mutant flowers early and exhibits reduced expression of the flowering repressor FLC (Noh et al., 2004), due to an elevated level of H3K27me3 at the FLC locus (Crevillén et al., 2014), but the ref6 mutant shows an FLC-dependent late-flowering phenotype (Noh et al., 2004). In addition, a recent study showed that loss of H3K27me3 was observed after salt priming in Arabidopsis seedlings (Sani et al., 2013), suggesting that some *JmjC* genes may be induced upon salt stress. To date, many members of the *JmjC* gene family have not been characterized genetically or biochemically; nevertheless, sequence and evolutionary analyses can help predict their functions in histone demethylation. However, the expression and functional divergence of *JmjC* duplicates have not been studied extensively, especially in response to abiotic stresses, such as abscissic acid (ABA), salt, and drought. In addition, members of the PKDM7 subfamily exhibited distinct expression patterns correlated with their different enzymatic activities (Lu et al., 2008), but the underlying sequence divergence and dynamic evolution are still unknown.

An early study of the *JmjC* gene family using sequences from Arabidopsis, rice (*Oryza sativa*), and human (*Homo sapiens*) defined 12 subfamilies according to phylogenetic relationships, with support from the presence of other domains (Zhou and Ma, 2008). These domains play potential regulatory roles in the demethylation process, such as the recognition of methylated histone marks (e.g. Plant homeodomain and Tudor domain), protein-protein interaction (e.g. F-box), and DNA binding (e.g. Cys2His2 zinc finger; Zhou and Ma, 2008). However, the previous reports have focused on a single species or a few species (Lu et al., 2008; Zhou and Ma, 2008); thus, there has not been a systematic study of the *JmjC* family in plants or animals. Recent progress in genome sequencing provides an opportunity to gain additional insights into the evolution of the *JmjC* family during the histories of angiosperms and vertebrates.

Histone methylation is mainly mediated by SET (for Suppressor of variegation [Su(var)]3-9), Enhancer of zest [E(z)] and Trithorax [Trx]) domain protein methyltransferase, and genome duplication has resulted in an increase of *SET* copy numbers. In Arabidopsis, five duplicated gene pairs are retained after recent genome duplication events, and 19 pairs are retained in poplar (*Populus trichocarpa*; Lei et al., 2012). It would also be helpful to learn whether the *JmjC* gene family has a similar evolutionary pattern to that of the *SET* genes. The evolutionary patterns of these two families could inform the balanced influence of demethylation and methylation on species evolution and divergence. In this study, we have characterized the evolution of *JmjC* genes in major eukaryotic lineages, including land plants, using phylogenetic and domain analyses, and we define 14 monophyletic subfamilies: *Jumonji*, AT-Rich Interactive Domain2, *JmjD6*, KDM2, KDM3, KDM4, KDM5, KDM6, PKDM7, PKDM8, PKDM9, PKDM10, PKDM11, PKDM12, and PKDM13. We show that some families underwent distinct gene duplication during evolution in angiosperms, especially in the ancestor of extant angiosperms.
RESULTS

Identification of JmjC Genes in Plants, Animals, and Fungi

The complete set of JmjC genes was identified from a comprehensive data set that contains selected plants, animals, and fungi based on HMMER software. In all, 434 sequences, each containing a JmjC domain, were retrieved from 35 different organisms: 11 plants, 16 metazoans, seven fungi, and Monosiga brevicollis (a unicellular choanoflagellate, a protist related to animals; Fig. 1; Supplemental Tables S1 and S2). The identified JmjC proteins range in size from 266 to 2,740 amino acids. Among major lineages of plants, JmjC genes are present in algae, Bryophyta, Pteridophyta, and angiosperms. The copy number of JmjC genes varies considerably among plants, ranging from two in the green algae Chlamydomonas reinhardtii and Volvox carteri to 17 in rice (monocot) and 21 in Arabidopsis (eudicot), with the highest number of 27 in poplar (eudicot). JmjC genes are also widespread in animals, from simple invertebrates, such as the sponge Amphimedon queenslandica, to mammals, such as human, with the gene copy number ranging from six to 28, as well as in the unicellular M. brevicollis. Further investigation reveals that the copy number variation in animals is mainly due to gene duplications in vertebrates; also, the distribution of subgroup PKDM12 in different organisms is complex. In fungi, there are fewer than six genes for each species, such as one in Schizosaccharomyces pombe, three in Saccharomyces cerevisiae, and five in Agaricus bisporus var. bisporus; altogether, 27 JmjC gene sequences were retrieved from seven fungi.

In order to standardize gene names and be consistent with the literature, we adopted a common nomenclature system based on the names of human genes according to the chromatin-modifying enzyme activities of animal members and Arabidopsis genes designated in previous studies as well as our phylogenetic analysis (Allis et al., 2007; Zhou and Ma, 2008). First, for human and Arabidopsis genes with known functions or previous research, the published gene names were retained and used as a reference. Second, the orthologs of these genes from plants, animals, and fungi were named following established references. Ultimately, recent paralogs were distinguished with an uppercase letter after the number, using the same letter for orthologs between organisms whenever possible.

Phylogenetic Classification of JmjC Genes into 14 Subfamilies

To explore the evolutionary relationships of eukaryotic JmjC genes, we conducted phylogenetic analyses
Figure 2. Phylogeny of representative JmjC genes from plants, animals, and fungi. A, Evolutionary relationship between JmjC genes from eight representative species (Chordata: Hs, Homo sapiens; Dr, Danio rerio. Arthropoda: Dm, Drosophila melanogaster. Basidiomycota: Av, A. bisporus var bisporus. Zygomycota: Ro, Rhizopus oryzae. Angiosperms: At, Arabidopsis; Os, rice. Bryophyta: Pp, Physcomitrella patens). Tree topology constructed using MEGA is shown here. For major nodes, NJ and ML (ML bootstrap [BS] values were generated by PhyML (version 3.0) and RAxML (version 7.0), respectively), BS values above 50%, followed by Bayesian posterior probability (PP) values are shown. For minor nodes, empty diamonds represent BS values of 50% to 69% and solid diamonds mark BS values of 70% to 99% from NJ analysis. For major and minor nodes, asterisks indicate BS/PP values of 100%/1. Substrate specificities for residues of histone tails are indicated for all subfamilies with demonstrated
with an alignment of the conserved JmjC domain from representative species using neighbor-joining (NJ), maximum likelihood (ML), and Bayesian methods. ML and Bayes analyses showed that proteins from different species cluster together in clades with high support values, with support from NJ analysis for most results. According to the results from phylogenetic and motif analyses, the eukaryotic JmjC genes can be divided into 14 subfamilies, designated as JARID2, JMD6, KDM2, KDM3, KDM4, KDM5, KDM6, PKDM7, PKDM8, PKDM9, PKDM10, PKDM11, and PKDM12, and PKDM13 (Fig. 2). Among these subfamilies, KDM3, KDM5, JMD6, PKDM11, and PKDM13 each contain members from plants, animals, and fungi. On the other hand, PKDM12 lacks fungal members, and KDM2 and KDM4 lack plant members. In addition, PKDM7, PKDM8, and PKDM9 are plant-specific groups, and JARID2, KDM6, and PKDM10 are animal specific.

The fact that KDM3, KDM5, JMD6, PKDM11, and PKDM13 subfamilies have members from major multicellular groups of eukaryotes, including plants, animals, and fungi, suggests that these clades originated from five respective ancestral genes in the most recent common ancestor (MRCA) of the three kingdoms. According to our phylogenetic analyses, JMD6, PKDM11, and PKDM13 cluster together with strong support, but some of the internal relationships of these clades are not clear. The KDM2 and KDM4 subfamilies both contain animal and fungal JmjC genes, indicating that the two clades are derived from ancestral genes that were present before the separation of animals and fungi. The KDM4 clade is well supported by all phylogenetic methods, but the classification of KDM2 also relied on the protein domain structures. The PKDM7, PKDM8, and PKDM9 subfamilies are plant specific. According to the tree topology, the PKDM8 subfamily forms a sister group to PKDM9, and together they are from an ancestral gene in the MRCA of plants, animals, and fungi, because they are sister to a large clade with genes from all three kingdoms. Both PKDM7 and PKDM9 subfamilies contain genes from major lineages of land plants, including Bryophyta, Pteridophyta, and angiosperms, revealing that they are retained in land plants, whereas PKDM8 genes are conserved from Pteridophyta to angiosperms, suggesting a likely loss of this subfamily in nonvascular plants. The JARID2, KDM6, and PKDM10 subfamilies are composed of only animal genes. However, the sisterhood of JARID2 and KDM6 to other clades containing genes from all three kingdoms suggests the origins of these subfamilies in the MRCA of three kingdoms; similarly, the ancestral gene of PKDM10 was probably already present in the MRCA of animals and fungi. The early origins of JARID2 and PKDM10 were further supported by the identification of their homologs in M. brevicollis.

In summary, eukaryotic JmjC genes form 14 subfamilies. These subfamilies can be grouped into four categories: five shared by all three major eukaryotic kingdoms; one shared by animals and plants; two found in both animals and fungi, with the counterparts lost from plants; and three plant specific and three animal specific. It is likely that there were at least 12 ancestral JmjC genes present in the MRCA of three major eukaryotic kingdoms, with subsequent duplications and losses in specific kingdoms (Fig. 2C). Additionally, within most subfamilies, except for JARID2, PKDM11, and PKDM13, there are two or more gene copies in plants or animals, suggesting further functional divergence of JmjC genes.

Different Domain Architecture and Conserved Non-JmjC Motifs in Subfamilies

According to the phylogenetic tree shown in Figure 2A, 14 monophyletic subfamilies were identified in the JmjC gene family. These 14 subfamilies represent 12 different domain architectures, as three subfamilies (PKDM11, PKDM12, and PKDM13) possess only the JmjC domain. Among these, the plant-specific PKDM7, PKDM8, and PKDM9 subfamilies have similar domain architectures, but PKDM7 proteins possess extra FYRN and FYRC domains. To determine whether motifs outside the JmjC and other known domains were conserved between members of the same subfamily, we searched for motifs in our data set of JmjC proteins. We found 40 motifs with lengths greater than 10 amino acids that are specific within subfamilies (Supplemental Table S3; Supplemental Fig. S1). The subfamilies exhibited various combinations of motifs; in addition, for subfamilies KDM3 and JMD6, different conserved motifs were found in plant and animal proteins (Fig. 2B). None of the 40 conserved motifs corresponded to known domains in the Pfam database. The highly conserved motifs shared by members of subfamilies of the JmjC domain proteins further support the classification presented here. The conservation of these additional domains in the respective subfamilies implies that they play important roles in the functions of these subfamilies.

Multiple Gene Duplication Events Identified in Angiosperms and Retention in Ancestral Angiosperms

To investigate the evolutionary history of the JmjC gene family, we analyzed the phylogeny of each

Figure 2. (Continued.)

demethylase activity. B, Highly conserved non-JmjC amino acid motifs in each JmjC subfamily. An idealized representation of a typical member of each JmjC subfamily is shown, with known domains drawn as shaded rectangles and other conserved motifs drawn as shaded ovals. The diagrams are not drawn to scale. The sequences of each motif in gene family members are given in Supplemental Table S3. C, Model of the evolutionary history of the JmjC gene family. In different kingdoms, the JmjC gene experienced several specific duplication events associated with the species divergence.
subfamily (Supplemental Figs. S2–S12). In the PKDM7 subfamily, AmPKDM7A from *Amborella trichopoda* (sister to other angiosperms) and AmPKDM7B are in two sister groups, each of which contains genes from both monocots and eudicots (Fig. 3). The Bryophyta and Pteridophyta PKDM7 genes constitute a clade separate from sister groups containing the *A. trichopoda* genes. Therefore, it is most likely that AmPKDM7A and AmPKDM7B were derived from gene duplication before the divergence of extant angiosperms but after the separation of angiosperms from other land plants. Additionally, in the PKDM7 subfamily, each of the PKDM7B, PKDM7C, and PKDM7E genes forms a monophyletic group that includes genes from Brassicaceae species, whereas the homologous genes in *P. trichocarpa* and *Vitis vinifera* form a separate basal clade. This topology is supported by multiple analyses and suggests that the duplication events likely occurred in the ancestor of the Brassicaceae. Similarly, in the PKDM8 subfamily, grass PKDM8A and PKDM8B genes are sister groups likely generated by duplication in the ancestor of grasses (Fig. 3B).

Whole-genome duplications (WGDs) are thought to be common in angiosperms and are also associated with the origin of vertebrate animals (Panopoulou et al., 2003; Jiao et al., 2011). To further ascertain whether these JmjC genes are caused by genome duplication events, we examined genomic regions containing duplicated genes in different species. Most members of these duplicated gene pairs are located in

---

**Figure 3.** Phylogenies of the PKDM7, PKDM8, and KDM5 subfamilies. Phylogenetic trees of the PKDM7 (A) and PKDM8 (B) genes from plants and KDM5 from major eukaryotes (C) constructed by ML using the Jones, Taylor, and Thornton model for amino acid substitution and γ-parameters are shown. For each clade, PhyML (version 3.0) and RAxML (version 7.0) percentage bootstrap values no less than 50% are presented. Asterisks indicate both bootstrap values of 100%.

---
syntenic genomic regions, indicating that these \textit{ImjC} gene copies are likely generated by chromosome segmental duplications or WGDs, such as \( \alpha \) and \( \beta \) in Brassicaceae, \( \gamma \) in early core eudicots, \( \rho \) and \( \sigma \) in grasses, and \( \varepsilon \) in early angiosperms (Jaillon et al., 2007; Barker et al., 2009; Tang et al., 2010; Jiao et al., 2011). Specifically, in the \textit{PKDM7} clade, the duplication of \textit{PKDM7C} and \textit{PKDM7E} in Brassicaceae is likely the result of \( \alpha \), and the combined clade of \textit{PKDM7C} and \textit{PKDM7E} form a sister to \textit{PKDM7B}, likely due to \( \beta \) (Fig. 3A; Supplemental Fig. S13). The duplication in core eudicots that generated \textit{PKDM7A} and \textit{PKDM7D} corresponds to \( \gamma \) (Supplemental Fig. S13C). A more ancient event, \( \varepsilon \), is also detected in our results, generating the aforementioned pair of \textit{AmPKDM7A} and \textit{AmPKDM7B} belonging to the two angiosperm sister groups in the \textit{PKDM7} subfamily.
the PKDM8 clade, the occurrence of the PKDM8A and PKDM8B is consistent with the ρ duplication within Poaceae (Fig. 3B; Supplemental Fig. S13D; Paterson et al., 2004). Additionally, we identified four duplicated gene pairs in the ancestral angiosperm (Fig. 4).

Multiple Gene Duplication Events Identified in Vertebrate Evolution and Lineage-Specific Losses

To better understand the evolutionary relationships of lineages within animals, we carried out similar analyses in subfamilies containing animal members. The KDM5 subfamily has three vertebrate groups named KDM5A, KDM5B, and KDM5C, but invertebrates, plants, and fungi have a single copy (Fig. 3C). KDM5A and KDM5B genes formed two well-supported sister groups that coincide with one of two WGDs (1R and 2R) before the divergence of vertebrates (Dehal and Boore, 2005). The combined clade of vertebrate KDM5A and KDM5B is sister to KDM5C, probably due to the earlier of the two WGDs in early vertebrates (Dehal and Boore, 2005). Further support for the origin of these groups from the WGDs is provided by the finding that HsKDM5A, HsKDM5B, and HsKDM5C are located in syntenic regions of three different chromosomes (Supplemental Fig. S14). In the KDM5B clade, a high degree of conservation of syteny was found in the two chromosome segments DrKDM5B and DrKDM5C (Supplemental Fig. S14B). These two genes were members of two fish-specific sister groups in the phylogenetic tree (Fig. 3C).

The duplication generating DrKDM5B and DrKDM5C corresponds to the WGD (3R WGD) that is specific to a fish lineage (Kasahara, 2007). In addition, phylogenetic analyses were performed using full-length sequences of JmjC proteins in other orthologous subfamilies. Most orthologous groups exhibited duplication events that correspond to WGDs, and 15 duplication events were shared by at least two animal species, in the phylogenetic trees for JMJ6, KDM2, KDM3, KDM4, KDM6, PKDM10, and PKDM12 (Supplemental Table S4; Supplemental Figs. S3–S9, S11, and S15). In addition to duplications produced by the 2R and 3R WGDs, a clade including only human and/or mouse genes was discovered in KDM3, KDM4, KDM5, and KDM6 groups, but the mechanisms of such mammal-specific duplication are unclear. Moreover, JARID2 contains animal-specific single-copy genes, and PKDM11 and PKDM13 subfamily members are single-copy orthologs not only in animals but also in plants and fungi (Supplemental Figs. S2, S10, and S12). Additionally, some of the duplicates were lost in some lineages. In KDM3, four human paralogs were located in a syntenic region, namely KDM3A, KDM3B, KDM3C, and KDM3D (Table I), but KDM3D was lost from birds and fishes. A phylogeny tree supports the origin of the four paralogs as results of the vertebrate WGDs 1R and 2R (Supplemental Fig. S5).

The KDM5 and KDM6 subfamilies are unusual in that they possess paralogs located on X and Y chromosomes, respectively. Our phylogenetic analyses indicate that KDM5D and KDM6C each has a sister group in mammals. Synteny analysis shows that these two JmJC genes are located on the Y chromosome and their respective paralogs are located on the X chromosome. A previous study showed that KDM5D arose about 154 million years ago in the common ancestor of marsupial and placental mammals, and KDM6C differentiated before the placental mammal radiation about 116 million years ago (Cortez et al., 2014). The human sex chromosome evolved from autosomes about 240 to 320 million years ago, which is more ancient than the occurrence of KDM5D and KDM6C (Lahn and Page, 1999; Bellott et al., 2010), suggesting that these paralogs could be the result of transposition from one sex chromosome to the other. On the other hand, the origins of KDM4D and KDM4E are not clear. KDM4D was found in mammals but not in other classes of vertebrates, and the latter is only found in primates. HsKDM4D and HsKDM4E are located close together on the same chromosome, indicating that the two genes were likely generated by a tandem duplication (Table II). KDM4D is only found in mammals and is nested in a clade of vertebrate KDM4B genes (Supplemental Fig. S6); the weak support of the clade with the KDM4D and fish KDM4B genes allows the possibility that the KDM4D genes originated in early vertebrates but were lost in birds and fishes. KDM4D and KDM4E both possess an N terminus but lack the C-terminal sequence that is conserved in KDM4B. About

### Table I. Number of human paralogous genes identified from the Synteny Database using KDM3 loci as seed genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>HsKDM3A</th>
<th>HsKDM3B</th>
<th>HsKDM3C</th>
<th>HsKDM3D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hsa2</td>
<td>Hsa3</td>
<td>Hsa10</td>
<td>Hsa8</td>
</tr>
<tr>
<td>HsKDM3A</td>
<td>6/52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HsKDM3B</td>
<td>39/97</td>
<td>14/76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HsKDM3C</td>
<td>13/70</td>
<td>32/32</td>
<td>59/63</td>
<td></td>
</tr>
</tbody>
</table>

### Table II. Number of human paralogous genes identified from the Synteny Database using KDM4 loci as seed genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>HsKDM4A</th>
<th>HsKDM4B</th>
<th>HsKDM4C</th>
<th>HsKDM4D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hsa1</td>
<td>Hsa19</td>
<td>Hsa9</td>
<td>Hsa11</td>
</tr>
<tr>
<td>HsKDM4A</td>
<td>93/203</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HsKDM4B</td>
<td>84/171</td>
<td>45/87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HsKDM4C</td>
<td>2/10</td>
<td>5/6</td>
<td>3/3</td>
<td></td>
</tr>
<tr>
<td>HsKDM4D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 5. Expression of Arabidopsis and A. trichopoda JmjC genes. A, Expression of Arabidopsis JmjC genes in different stages or tissues. The x axis indicates different representative genes, and the y axis represents reads per kilobase of mRNA length per million mapped reads values. Genes resulting from recent duplications are shown in the same color. B, Expression of A. trichopoda JmjC genes in different tissues. C, Expression of Arabidopsis JmjC genes under drought, cold, high-salinity, and ABA treatment. Asterisks indicate differential expression under stress conditions compared with normal conditions.
20 introns were identified in HsKDM4A, HsKDM4B, and HsKDM4C, but only two in HsKDM4D and none in HsKDM4E, suggesting that the HsKDM4B to HsKDM4D genes derived from HsKDM4B by retroposition. It is possible that KDMA4, KDMA4B, KDMA4C, and KDMA4D resulted from the vertebrate WGDs, if weakly supported relationships are not considered.

**Expression and Functional Divergence of JmjC Duplicates**

To obtain clues about possible functional divergence, we examined RNA-sequencing (RNA-seq) data of Arabidopsis and *A. trichopoda* and found that most JmjC genes were expressed in the developing flower and different tissues, except AtPKDM7C and AmPKDM12B (Fig. 5; Supplemental Table S5). More interestingly, AtKDM3D and AmKDM3D are both expressed more highly than other genes in the same subfamily. Transcriptomic comparison of Arabidopsis and *A. trichopoda* JmjC genes indicated that most were expressed at similar levels in different tissues, suggesting functional conservation in angiosperms. In addition, public tiling array data of Arabidopsis under drought, cold, high-salinity, and ABA treatments, but NaCl treatments, indicate that many duplicates showed differential expression upon cold treatment, including AtPKDM7E, AtPKDM3E, AtPKDM7A, and AtPKD8, and many were affected by other treatments. The number of differentially expressed genes in a 10-h treatment was greater than that found after a 2-h treatment. These results suggest that plant JmjC genes play potential roles in stress responses.

Gene duplication increases the quantity of genetic material and improves the chance of functional innovations during evolution. To obtain clues about functional divergence, we examined some of the recent duplicates for expression changes in different tissues and upon abiotic stresses. In Arabidopsis, a trio of paralogs are retained after the Brassicaceae WGDs, AtPKDM7C, AtPKDM7E, and AtPKDM7B. Among them, AtPKDM7E was differentially expressed under drought, but not AtPKDM7B and AtPKDM7C (Fig. 5C; Supplemental Table S6). Some JmjC genes showed differential expression upon cold treatment, including AtPKDM3E, AtPKDM3F, AtPKDM12A, and AtPKDM8, and many were affected by other treatments. The number of differentially expressed genes in a 10-h treatment was greater than that found after a 2-h treatment. These results suggest that plant JmjC genes play potential roles in stress responses.

To further investigate functional divergence among JmjC genes, we examined transcriptomic data sets of the wild type and several anther mutants, including dysfunctional tapetum1 (*dyt1*), basic helix-loop-helix10 (*bhh10*), bhh189, and bhh91 (E. Zhu, C. You, J. Cui, F. Chang, and H. Ma, unpublished data), as a case study of possible regulation of JmjC genes by such regulatory genes. The expression of most JmjC genes was not altered in these mutants, except for AtKDM6B, AtPKDM12A, AtPKDM7E, and AtPKDM8, with 2-fold or greater changes (Fig. 6; Supplemental Table S7). DYT1 encodes a bHLH-type transcription factor and is important for normal tapetum development (Zhang et al., 2006; Feng et al., 2012), as are other bHLH transcription factors; thus, the expression change of several JmjC genes in these mutants suggests a potential role in tapetum development. Additionally, we also compared JmjC gene expression between the wild type and the *pkdm7d/jmj16* mutant (E. Zhu, C. You, J. Cui, F. Chang, and H. Ma, unpublished data) and found some JmjC genes with distinct expression patterns (Fig. 6; Supplemental Table S7), suggesting functional divergence among these genes. Specifically, AtPKDM12A was dramatically down-regulated in *pkdm7d/jmj16*, indicating that its transcription may be regulated by PKDM7D directly or indirectly. Together, JmjC gene expression from public and our own RNA-seq data strongly supports possible functional divergence.

**Sequence Analysis for Functional Diversification of JmjC Duplicates**

In addition to the expression divergence of JmjC genes, functional diversification (FD) of JmjC paralogs could also occur in coding regions. To estimate FD between two paralogs, we used the DIVERGE3.0 program with amino acid sequence data: FD I was based on evolutionary rate (Gu, 1999), and FD II was based on differences in biochemical properties of amino acids (Gu, 2006; Table III). We found that the angiosperm-wide duplicates, PKDM7A and PKDM7B, have FD I with the highest significance (*P* = 1.9E-16). In addition, two other pairs, Brassicaceae PKDM7B versus PKDM7C (*P* = 3.1E-08) and grass PKDM8A versus PKDM8B (*P* = 8.9E-13), also have highly significant FD I.

Using the same program, amino acid sites likely contributing to FD I could be identified. Between the angiosperm PKDM7A and PKDM7B following the ε WGD, 51 putative sites responsible for FD I were identified (Table III; Supplemental Table S8). In addition, 32 sites were found for the angiosperm pair KDM3B and KDM3C, also generated at the time of the ε WGD. The amino acid residues at these sites were highly conserved between plants within one paralogous group but diverse among the other group of paralogs. Some of the differences between paralogs involved functionally distinct amino acids, such as that at position 268, which is occupied by the hydrophobic Phe in PKDM7A proteins but by polar hydrophilic Ser in PKDM7B. Similarly, position 671 has a polar hydrophilic Arg in PKDM7B but hydrophobic amino acids (e.g. Ile, Trp, and Leu) in PKDM7A.
proteins (Supplemental Fig. S16). Next, we examined the amino acid sites identified using FD analysis for known functions and found that many such FD I-associated amino acids are located in the JmjC or zinc finger domain, suggesting that functional specialization could have occurred in enzymatic and/or binding activities (Supplemental Table S8). At the same time, other sites showing FD I have no defined role; further experiments are needed to test the possible functional differences at these sites.

DISCUSSION

WGDs Contributed to JmjC Gene Expansion Especially Near the Origin of Angiosperms and Vertebrates

WGDs are found widely, including in both plants and animals, such as those associated with the origins of angiosperms (ε), core eudicots (γ), and Brassicaceae (α and β), and those in the ancestral vertebrates (2R) and supported by synteny, age estimates of gene duplications, and gene family phylogeny (Jaillon et al., 2007)

Table III. Analysis of FD by DIVERGE

<table>
<thead>
<tr>
<th>FD</th>
<th>WGD</th>
<th>Subfamilies</th>
<th>Coefficient θ ± se (P)</th>
<th>No.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>α/β Brassicaceae PKDM7C versus PKDM7E</td>
<td>0.225383 ± 0.101966 (2.23E-02)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Brassicaceae PKDM7B versus PKDM7C</td>
<td>0.529252 ± 0.101985 (3.09E-08)</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Brassicaceae PKDM7B versus PKDM7E</td>
<td>0.360678 ± 0.117604 (1.45E-03)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>α/β</td>
<td>Brassicaceae KDM3E versus KDM3F</td>
<td>0.101601 ± 0.074635</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>Eudicot PKDM7A versus PKDM7D</td>
<td>0.008360 ± 0.152924</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ρ</td>
<td>Grass PKDM8A versus PKDM8B</td>
<td>0.547235 ± 0.057572 (8.92E-13)</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grass KDM3C versus KDM3D</td>
<td>0.33263 ± 0.083005 (1.22E-05)</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>ε</td>
<td>Angiosperm PKDM7A versus PKDM7B</td>
<td>0.401872 ± 0.059802 (1.86E-16)</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiosperm PKDM9A versus PKDM9B</td>
<td>0.128948 ± 0.038607 (3.21E-04)</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiosperm KDM3A versus KDM3D</td>
<td>0.349834 ± 0.149399 (1.37E-02)</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiosperm KDM3B versus KDM3C</td>
<td>0.343037 ± 0.089236 (1.04E-05)</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiosperm KDM3AD versus KDM3BC</td>
<td>0.115225 ± 0.063442 (4.91E-02)</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Type II</td>
<td>α/β Brassicaceae PKDM7C versus PKDM7E</td>
<td>0.055129 ± 0.045874</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Brassicaceae PKDM7B versus PKDM7C</td>
<td>0.027313 ± 0.049140</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>α/β</td>
<td>Brassicaceae KDM3E versus KDM3F</td>
<td>0.027313 ± 0.049140</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>Eudicot PKDM7A versus PKDM7D</td>
<td>0.167973 ± 0.084460 (4.67E-02)</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>ρ</td>
<td>Grass PKDM8A versus PKDM8B</td>
<td>0.401872 ± 0.059802 (1.86E-16)</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grass KDM3C versus KDM3D</td>
<td>0.167973 ± 0.084460 (4.67E-02)</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>ε</td>
<td>Angiosperm PKDM7A versus PKDM7B</td>
<td>0.008820 ± 0.035777</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiosperm PKDM9A versus PKDM9B</td>
<td>0.094928 ± 0.010074</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiosperm KDM3A versus KDM3D</td>
<td>0.015185 ± 0.040608 (1.51E-04)</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiosperm KDM3B versus KDM3C</td>
<td>0.035761 ± 0.132655</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiosperm KDM3AD versus KDM3BC</td>
<td>0.181663 ± 0.213708</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

*Number of critical amino acid sites detected as related to FD.
relatively high rate of retention of WGD. Globally, most duplicates generated by the 2R population of gene expression played an important role in animals and their major subgroups (core eudicots, et al., 2012). The SET of the mechanism for the expansion of the results suggest that WGDs could be the main mechanism for the expansion of regulatory genes were due to WGD during the last 150 million years (Maere et al., 2005). Consistently, our study indicates that the lost and possible functional change occurred before the divergence of A. trichopoda from other angiosperms. The paralogs KDM3A and KDM3D (Supplemental Fig. S5) were due to the duplication in early angiosperms, indicating that KDM3A had lost the conserved residues after gene duplication. Moreover, these genes are differentially expressed in various tissues and in response to environmental signals, indicating that they might have lost the ancestral functions but gained new functions after gene duplication.

Increase in the Complexity of the JmjC-Mediated Histone Demethylation for Angiosperm and Vertebrate Evolution

The SET genes have experienced many duplication events, particularly in the Sue, Absent, small, or homeotic discs1, Trx, and E(z) subfamilies, which are responsible for catalyzing specific Lys methylation at H3K9, H4K36, H3K4, and H3K27 (Zhang and Ma, 2012). Our results showed that the JmjC genes share a similar evolutionary pattern, with expansion in many well-supported large clades, including lineage-specific subfamilies KDM3 to KDM6 and PKDM7 to PKDM9 (Fig. 2A). The JmjC gene family also has similar copy numbers to the SET family, suggesting that the two gene families might have coevolved, which is an important question for further investigation. In addition, the Arabidopsis JMJ6A and PKDM11 proteins are histone Arg demethylases (Table III), and their genes exist as single-copy genes belonging to highly conserved subfamilies (Supplemental Figs. S3 and S10). Similarly, histone Arg methylation is catalyzed by a small group of protein Arg methyltransferases, including PRMT5, PRMT10, and PRMT4 (Ahmad and Cao, 2012). The phylogenetic tree showed that members of these three subfamilies are also highly conserved and mostly single copy, except for two copies of PRMT4 (PRMT4A and PRMT4B) due to the lineage-specific expansion in Brassicaceae (Supplemental Fig. S18). Together, the highly stable single/low copies of these genes suggest that Arg methylases and demethylases are more conserved than those for Lys modification. In addition to duplication patterns, the JmjC and SET proteins also share some domains, such as Plant homeodomain, Tudor domain, and Really Interesting Gene, suggesting that they might have similar interactive partners.

Sequence Divergence of Specific JmjC Proteins Suggests Changes in Enzyme Activity and Biological Functions

The JmjC domain contains five conserved amino acid residues within cofactor-binding sites that are important for enzymatic activity. Among the five conserved residues, His, Glu, and His bind to the Fe(II) cofactor and Phe and Lys are required for binding to αKG. According to the phylogeny, AtpKDM7C, JMJ14/AtpKDM7B, JMJ16/AtpKDM7D, JMJ18/AtpKDM7E, and JMJ19/AtpKDM7A belong to an H3K4-specific demethylase group (Table IV); however, AtpKDM7A lacks the conserved residues for αKG binding, suggesting that enzymatic activity is lost in this protein (Lu et al., 2008). Similarly, the other proteins encoded by core eudicot members of the PKDM7A clade in the PKDM7 subfamily, CrPKDM7A, VvPKDM7A, and PiPKDM7A, also lack the same conserved residues; in contrast, other angiosperm proteins of the PKDM7 subfamily have all five conserved amino acids within the cofactor-binding sites (Supplemental Fig. S17), suggesting that the ancestral core eudicot PKDM7A gene had lost the conserved residues. Specifically, the two clades (PKDM7A and PKDM7D) are due to the γ WGD in early core eudicots, and PKDM7D proteins still have the conserved residues, suggesting sequence and functional divergence following the γ WGD but before the divergence of extant core eudicots. We also found that members of the KDM3A clade, including AmkD3M3A, lack the conserved residues for Fe(II) and αKG binding; their paralogs and other homologs still retain the conserved amino acids, showing that the lost and possible functional change occurred before the divergence of A. trichopoda from other angiosperms. The paralogs KDM3A and KDM3D (Supplemental Fig. S5) were due to the duplication in early angiosperms, indicating that KDM3A had lost the conserved residues after gene duplication. Moreover, these genes are differentially expressed in various tissues and in response to environmental signals, indicating that they might have lost the ancestral functions but gained new functions after gene duplication.


Qian et al.

Kasahara, 2007; Barker et al., 2009; Jiao et al., 2011; Amborella Genome Project, 2013). We have identified four pairs of JmjC paralogs likely due to the angiosperm ancestral ε WGD (Amborella Genome Project, 2013); these pairs of paralogous groups are also supported by genes from A. trichopoda, the sister of all other extant angiosperms (Fig. 4A). These results suggest that duplicate JmjC genes from ancient duplications were associated with the origin and early evolution of angiosperms. Additionally, synteny analysis indicated that most JmjC duplicates in vertebrates were formed by WGD, with additional fish-specific contribution from the 3R WGD. Globally, most duplicates generated by the 2R WGD have been lost (Dehal andBoore, 2005); thus, the relatively high rate of retention of KDM3 duplicates in vertebrates (Supplemental Fig. S5) suggests that this group might be important for chromatin-based regulation in vertebrates. Previous studies proposed that more than 90% of the increases in Arabidopsis regulatory genes were due to WGD during the last 150 million years (Maere et al., 2005). Consistently, our results suggest that WGDs could be the main mechanism for the expansion of the JmjC gene family in angiosperms and vertebrates.

This pattern of evolution is also very similar to that of the SET domain family of histone methylases (Lei et al., 2012). The JmjC and SET genes both control the methylation status of histones and are important regulators of chromatin structure; our study indicates that these regulators share evolutionary patterns and mechanisms with transcription factor genes and likely contribute to the evolution of regulatory networks. The expansion of genes for histone modification following WGDs near the origins of angiosperms and vertebrate animals and their major subgroups (core eudicots, Brassicaceae, and fish) suggests that epigenetic modulation of gene expression played an important role in the evolutionary successes of the dominating groups in both plants and animals.

Copyright © 2015 American Society of Plant Biologists. All rights reserved.
It is worth noting that there are also some important differences between the SET and JmjC families. Among the Arabidopsis SET genes, the SETD and SMYD genes encode proteins with an insertion of 100 to 300 residues in the middle of the SET domain; however, the JmjC family does not have members with an insertion. Also, the human SET family has only 20 members in four subfamilies, due to fewer retained copies after the 2R WGD, compared with 28 human JmjC genes. As mentioned before, some JmjC duplicates were generated by tandem duplication and possibly transpositions, but such mechanisms have not been detected for SET genes.

## Table IV. Functionally characterized JmjC genes from Arabidopsis

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Locus Tag</th>
<th>Other Names</th>
<th>Substrates</th>
<th>Biological Function</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subfamily JMJD6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtJMJD6A</td>
<td>AT5G06550</td>
<td>JMj22</td>
<td></td>
<td>Seed germination</td>
<td>Cho et al. (2012)</td>
</tr>
<tr>
<td>AtJMJD6B</td>
<td>AT1G78280</td>
<td>JMj21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subfamily KDM3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtKDM3A</td>
<td>AT1G09060</td>
<td>JMj24</td>
<td></td>
<td>No activity</td>
<td></td>
</tr>
<tr>
<td>AtKDM3B</td>
<td>AT4G21430</td>
<td>JMj28</td>
<td></td>
<td>No activity</td>
<td></td>
</tr>
<tr>
<td>AtKDM3C</td>
<td>AT3G07610</td>
<td>JMj25/JIMj3</td>
<td></td>
<td>RNA-directed DNA methylation</td>
<td>Saze et al. (2008); Fan et al. (2012)</td>
</tr>
<tr>
<td>AtKDM3D</td>
<td>AT4G09990</td>
<td>JMj27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtKDM3E</td>
<td>AT1G11950</td>
<td>JMj26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtKDM3F</td>
<td>AT1G62310</td>
<td>JMj29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subfamily KDM5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtKDM5</td>
<td>AT1G63490</td>
<td>JMj17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subfamily PKDM7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtPKDM7A</td>
<td>AT2G38950</td>
<td>JMj19</td>
<td></td>
<td>No activity</td>
<td></td>
</tr>
<tr>
<td>AtPKDM7B</td>
<td>AT4G20400</td>
<td>JMj14</td>
<td></td>
<td>H3K4me1/2/3</td>
<td></td>
</tr>
<tr>
<td>AtPKDM7C</td>
<td>AT2G34880</td>
<td>JMj15/MEE27</td>
<td></td>
<td>Salt tolerance, flowering time, female gametophyte development, early embryo and endosperm formation</td>
<td>Pagnussat et al. (2005); Day et al. (2008); Yang et al. (2012b); Shen et al. (2014)</td>
</tr>
<tr>
<td>AtPKDM7D</td>
<td>AT1G08620</td>
<td>JMj16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtPKDM7E</td>
<td>AT1G30810</td>
<td>JMj18</td>
<td></td>
<td>H3K4me2/3</td>
<td></td>
</tr>
<tr>
<td>Subfamily PKDM8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtPKDM8</td>
<td>AT5G49610</td>
<td>JMj13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subfamily PKDM9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtPKDM9A</td>
<td>AT3G48430</td>
<td>JMj12/REF6</td>
<td></td>
<td>H3K9me3, H3K27me2/3, H3K36me2/3</td>
<td>Noh et al. (2004); Yu et al. (2008); Ko et al. (2010); Lu et al. (2011a)</td>
</tr>
<tr>
<td>AtPKDM9B</td>
<td>AT5G04240</td>
<td>JMj11/ELF6</td>
<td></td>
<td>Flowering time, photoperiod pathway, transgenerational inheritance of vernalized state, repression of flowering time</td>
<td>Noh et al. (2004); Yu et al. (2008); Jeong et al. (2009); Crevillen et al. (2014)</td>
</tr>
<tr>
<td>Subfamily PKDM11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtPKDM11</td>
<td>AT5G63080</td>
<td>JMj20</td>
<td></td>
<td>H3R2me3, H4R3me2</td>
<td></td>
</tr>
<tr>
<td>AtPKDM12A</td>
<td>AT3G20810</td>
<td>JMj30/JIMjD5</td>
<td></td>
<td>Circadian system, flowering time</td>
<td>Jones et al. (2010); Jones and Harmer (2011); Lu et al. (2011b); Gan et al. (2014); Yan et al. (2014)</td>
</tr>
<tr>
<td>AtPKDM12B</td>
<td>AT5G19840</td>
<td>JMj31</td>
<td></td>
<td>Flowering time</td>
<td></td>
</tr>
<tr>
<td>Subfamily PKDM13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtPKDM13</td>
<td>AT3G45880</td>
<td>JMj32</td>
<td></td>
<td>H3K27me3</td>
<td></td>
</tr>
</tbody>
</table>

*These terms are defined as follows: H, histone; R, Arg; K, Lys; and me1/2/3, monomethylation, dimethylation, or trimethylation.*
dimethylation (H3K9me2) is highly concentrated at the centromeric heterochromatin, and this modification is a prerequisite for DNA methylation and gene silencing in Arabidopsis (Jackson et al., 2004; Kavi and Birchler, 2009). In Arabidopsis, 15 SET domain-containing proteins are thought to mediate H3K9me2 deposition, but only six JmjC domain-containing proteins are related to H3K9 demethylation. During angiosperm evolution with additional WGDs, several duplicates in the SET family have been retained, but most of the recent JmjC duplicates have been lost. Our results also suggest that H3K9me2 modification and genome duplication are likely correlated. For instance, an elevated H3K9me2 level repressed redundant gene expression after WGD.

During the evolution of angiosperms with increasing species diversity, it is likely that greater complexity of the histone modification system, due to the expansion of both JmjC and SET genes, provided crucial functions for chromatin-based regulation. The rapidly advancing reverse genetic technologies, including the clustered regularly interspaced palindromic repeats (CRISPR)/CRISPR-associated protein-9 nuclease system, will allow functional tests of the hypotheses raised here on the divergence between paralogs. In particular, the biochemical similarity of closely related JmjC family members can be examined experimentally, whereas functional differences of more distantly related members can be similarly tested. These studies in the near future will likely further improve the understanding of the function and evolution of histone modification enzymes and their impact on development and physiology.

MATERIALS AND METHODS

Data Retrieval

To establish an initial data set with as many JmjC genes as possible, we used several resources. The sequenced genomes and predicted proteomes of plants and animals were downloaded from Phytozome (version 9.0; http://www.phytozome.net/) and Ensembl (release 66; http://www.ensembl.org), respectively. The Joint Genome Institute (http://genome.jgi.doe.gov/) and the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) databases were additional sources for fungal and partial animal sequences with genome annotation. The sequences for Arabidopsis thaliana and Monosiga brevicollis were retrieved from the Amborella Genome Database (http://www.amborella.org/) and the Joint Genome Institute, respectively. In proteome data sets, if two or more proteins are annotated for the same gene from alternative splicing, we selected the longest form. The numbers of genes in each JmjC subfamily in different organisms can be found in Supplemental Table S1.

Homolog Searches

Three steps were carried out to identify all recognizable JmjC family members. First, regardless of the origin of proteome data, the HMMER program (version 3.0; Eddy, 1998) with the hidden Markov model was employed to retrieve all sequences with genome annotation. The sequences for Arabidopsis thaliana and Monosiga brevicollis were retrieved from the Amborella Genome Database (http://www.amborella.org/) and the Joint Genome Institute, respectively. In proteome data sets, if two or more proteins are annotated for the same gene from alternative splicing, we selected the longest form. The numbers of genes in each JmjC subfamily in different organisms can be found in Supplemental Table S1.

from the first and second steps were verified using the Pfam database (http://pfam.xfam.org/search), CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) available from the National Center for Biotechnology Information, and the SMART database (http://smart.embl-heidelberg.de/), with a threshold e-value of less than 1e-10 (Marchler-Bauer et al., 2011; Letunic et al., 2012; Finn et al., 2014). The genes verified by all three steps were then included in our study.

Sequence Alignment

Preliminary multiple sequence alignment was performed using MUSCLE (version 3.8.425) with default parameters (Edgar, 2004). The alignment was then used to generate a preliminary ML tree using FastTree (version 2.1.7; Price et al., 2009). According to the tree topology, the JmjC family was divided into several subgroups. A second round of multiple sequence alignment was carried out for sequences in each subgroup and then adjusted manually in Jalview (version 2.8; Waterhouse et al., 2009). Subsequently, these alignments were combined together using the profile alignment function of MUSCLE.

Phylogenetic Analysis

Systematic phylogenetic analysis of the JmjC family was performed using NJ, ML, and Bayesian methods. For NJ analysis, we used MEGA (version 5.0) with the pairwise deletion option and Poisson correction model (Tamura et al., 2011). The reliability of internal branches was evaluated with a bootstrap test of 1,000 replicates. In this study, ProtTest (version 2.4) was used for model selection, crucial for ML and Bayesian analysis (Abascal et al., 2005). PhyML (version 3.0) and RAxML (version 7.0) were employed to construct ML trees with the Whelan and Goldman amino acid substitution model, γ-distribution, and 100 nonparametric bootstrap replicates (Guindon and Gascuel, 2003; Stamatakis, 2006). Bayesian trees were constructed using MrBayes (version 3.2.1), with the fixed Whelan and Goldman model, four Markov chains, and an average so of 0.01 (Ronquist and Huelsenbeck, 2003).

Motif and Synteny Analyses

All verified JmjC genes were used to search against the PFAM, CDD, and SMART databases to uncover other known domains or motifs apart from the JmjC domain. Additionally, to discover novel conserved patterns in the amino acid sequences of JmjC proteins, all sequences were analyzed using the software MEME (version 4.9.0; Bailey and Elkan, 1994). The length of a motif was set between 10 and 120 amino acids, and the number of motifs was limited to no more than 10. All sequences within each subfamily were analyzed separately by MEME to identify conserved motifs within the group. Every motif was screened against the PFAM database, and the majority of these motifs are not recorded in the database. In addition, genome synteny between duplicate gene pairs was examined by phylogenetic analysis using the Plant Genome Duplication Database (http://chibba.agtec.uga.edu/duplication/; last accessed in February 2015; Tang et al., 2008) and the Synteny Database (http://syntenydb. uoregon.edu/synteny_db/; last accessed in February 2015; Catchen et al., 2009) for plants and animals, respectively.

Expression Analysis

RNA-seq data were of the same sources and treatment as our previous work, including the following Arabidopsis (Arabidopsis thaliana) tissues: seedling, stage 4 flower, inflorescence meristem, stage 1 to 9 flowers, stage 12 flowers, and meiosis (Zhang and Ma, 2012; Zhang et al., 2015). RNA-seq data sets of the following A. thaliana tissues (ftp://amborella-project.huck.psu. edu/Public/Amborella/ transcriptome/) were mapped to the A. thaliana genome sequences with Periodic Seed Mapping (https://coge.google.com/p/ perm/); apical meristem, female buds, whole-plant normalized-1, and whole-plant normalized-2. Only the uniquely mapped sequence reads were used further. The gene expression levels were measured by reads per kilobase of mRNA length per million mapped reads. Tiling array data of Arabidopsis JmjC homologs under drought, cold, high-salinity, and ABA treatment (Matsumoto et al., 2008) were also included in our analysis.

Functional Divergence Analysis

The analysis of FD between JmjC genes of paralogous clades was performed using DIVERGE version 3.0 software (Gu et al., 2013). DIVERGE used
the model free estimation calculation of the θ and σt type I and II coefficients of FD, based on the occurrence of altered selective constraints or radical shifts of physiochemical properties, respectively (Gu, 1999, 2006). The program also estimates the posterior probabilities of amino acid sites to be responsible for FD. A value of 0.6 was chosen as a cutoff to measure the degree of FD at the amino acid level.

Accession numbers and gene identifiers for sequences used in this study are provided in Supplemental Table S2.

**Supplemental Data**

The following supplemental materials are available.

**Supplemental Figure S1.** Logos representing motifs conserved within each subfamily.

**Supplemental Figure S2.** The ML tree of the JARID2 subfamily generated by RAxML.

**Supplemental Figure S3.** The ML tree of the JMJD6 subfamily generated by RAxML.

**Supplemental Figure S4.** The ML tree of the KDM2 subfamily generated by RAxML.

**Supplemental Figure S5.** The ML tree of the KDM3 subfamily generated by RAxML.

**Supplemental Figure S6.** The ML tree of the KDM4 subfamily generated by RAxML.

**Supplemental Figure S7.** The ML tree of the KDM6 subfamily generated by RAxML.

**Supplemental Figure S8.** The ML tree of the PKDM9 subfamily generated by RAxML.

**Supplemental Figure S9.** The ML tree of the PKDM10 subfamily generated by RAxML.

**Supplemental Figure S10.** The ML tree of the PKDM11 subfamily generated by RAxML.

**Supplemental Figure S11.** The ML tree of the PKDM12 subfamily generated by RAxML.

**Supplemental Figure S12.** The ML tree of the PKDM13 subfamily generated by RAxML.

**Supplemental Figure S13.** An illustration of plants with the syntenic regions containing representative duplicated gene pairs from recent polyploidy events.

**Supplemental Figure S14.** An illustration of animals with duplicated gene pairs generated by recent polyploidy events.

**Supplemental Figure S15.** Duplication events in the JmjC family inferred from amino acid sequence analyses.

**Supplemental Figure S16.** Multiple sequence alignments of angiosperm PKDM7A and PKDM7B.

**Supplemental Figure S17.** An ML tree and multiple sequence alignments of PKDM7A and PKDM7D in core eudicots.

**Supplemental Figure S18.** An ML tree of PRMT5, PRMT10, and PRMT4 in green plants.

**Supplemental Table S1.** The number of different JmjC subfamily members in representative species.

**Supplemental Table S2.** A list of all JmjC genes included in this study.

**Supplemental Table S3.** The sequences of conserved motifs within individual subfamilies.

**Supplemental Table S4.** Number of human paralogous genes identified from Synteny Database at Uoregon site using KDM2 and PKDM10, KDM5, KDM6 loci as seed genes and Ciona intestinalis as an outgroup.

**Supplemental Table S5.** Expression profile of JmjC genes at different stages of flower development in Arabidopsis and Amborella.

Duplications of Histone Demethylase Genes in Angiosperms

**Supplemental Table S6.** Tiling array profile of Arabidopsis JmjC genes under drought, cold, high-salinity, and ABA treatments.

**Supplemental Table S7.** Expression of JmjC genes in the wild type and the dy3 (bhh22, bhh10, bhh09, bhh01, and phn4b) mutant lines by RNA-seq.

**Supplemental Table S8.** A list of sites with Functional Divergence type I in plant JmjC genes.

**ACKNOWLEDGMENTS**

We thank Yaqiong Wang, Liping Zeng, Hai Feng Wang, Qi Li, and Fei Chen for comments on the article and helpful discussion.

Received April 8, 2015; accepted June 4, 2015; published June 9, 2015.

**LITERATURE CITED**


Copyright © 2015 American Society of Plant Biologists. All rights reserved.


