ORIGINAL ARTICLE

Genome-wide identification of chromatin regulators in Sorghum bicolor

Yongfeng Hu[1] · Xiaoliang Chen[1] · Chao Zhou[1] · Zhengquan He[1] · Xiangling Shen[1]

Abstract
Chromatin regulators play important roles in plant development and stress response. In this study, we identified totally 231 chromatin regulators including 63 histones, 29 histone chaperones, 101 histone modification enzymes, and 38 chromatin remodeling factors from *Sorghum bicolor* (L.) Moench. Most of these chromatin regulators are homologous to their counterparts in Arabidopsis or rice. However, sorghum genome evolves a few novel histone variants specific to some grass species and a sorghum-unique chromatin remodeling factor that contain the domains belonging to the elongation factor EF-Tu and the histone chaperone SPT16. Finally, we performed co-expression analysis for the chromatin regulator-encoding genes by clustering the expression patterns of these genes. Our results provide useful information for the future studies on the mechanism of epigenetic regulation in sorghum and its roles in development and stress response.

Keywords Histone variant · Histone chaperone · Histone acetylation · Histone methylation · Chromatin remodeling · *Sorghum bicolor* (L.) Moench

Abbreviations

HATs	Histone acetyltransferases
HDACs	Histone deacetylases
HMTs	Histone methyltransferases
HDMs	Histone demethylases
DMTs	DNA methyltransferases
DDMs	DNA demethylases

Introduction

In eukaryotes, DNA resides in the nucleus by associating with histone octamer to form the nucleosome which is the basic unit of chromatin. Histone octamer is composed of two copies of H3, H4, H2A, and H2B. These histones are designated as the core histones. Two H3–H4 dimers interact to create a tetramer, and then two H2A-H2B dimers bind the tetramer at each side to form the octamer. About 147 bp core DNA wraps around the histone octamer to generate the nucleosome core particle, which constitutes the nucleosome together with the linker DNA. Histone H1, also called the linker histone, binds to the linker DNA at both the entry site and the exit site of the nucleosome (Fyodorov et al. 2018). This structure is referred to as the chromatosome. The arrayed nucleosomes can be further compacted into the chromatin fiber with a diameter of about 30 nm. This process requires the linker histone H1 (Song et al. 2014). Histone proteins are encoded by multiple copies of genes, which are transcribed into mRNAs in the nucleus and then translated into proteins in the cytoplasm as other protein-coding genes. Once the histone proteins are synthesized in the cytoplasm, they are escorted by a series of proteins called histone chaperones. Histone chaperones are crucial for proper folding, transportation into the nucleus, and incorporation into the nucleosome of the histone proteins (Hammond et al. 2017). Canonical histones are generally deposited into the chromatin at the S phase of the cell cycle in a DNA replication-dependent manner. Histone variants, differing from the canonical histones by a few amino acids or a larger polypeptide segment, can be incorporated outside the S phase in any cell type (Martire and Banaszynski 2020). All the

Yongfeng Hu and Xiaoliang Chen contribute equally to this work.

✉ Yongfeng Hu
huyongfeng@ctgu.edu.cn

✉ Xiangling Shen
shenxl1982@hotmail.com

1 Key Laboratory of Three Gorges Regional Plant Genetics and Germplasm Enhancement, Biotechnology Research Center, College of Biological & Pharmaceutical Sciences, China Three Gorges University, Yichang 443002, Hubei, China

core histones contain the conserved histone-fold domain composed of three α-helices connected by two loops. The histone-fold domain is important for histone dimerization (Luger et al. 1997). The N-terminal tails of each core histone extend out of the nucleosome and are decorated with various covalent post-translational modifications (PTMs). Among these histone PTMs, histone methylation and acetylation are studied the most (Audia and Campbell 2016). Histone acetylation, which is balanced by histone acetyltransferases (HATs) and histone deacetylases (HDACs), mainly occurs at lysine residues (Lee and Workman 2007; Seto and Yoshida 2014). Histone can be mono-, di-, and tri-methylated at lysine residues and mono-, symmetrically and asymmetrically di-methylated at arginine residues (Jambhekar et al. 2019). Histone lysine methylation is carried out by SET domain family proteins and histone arginine is methylated by protein arginine methyltransferases (PRMTs). Histone methylation is reversed by Jumonji C (JmjC) domain family proteins and LSD1 proteins (Jambhekar et al. 2019). The chromatin DNA can also be methylated at cytosines in different background such as CG in animals and CG, CHH, and CHG in plants (Law and Jacobsen 2010). DNA methylation is catalyzed by DNA methyltransferases (DMTs) and removal of methyl group from DNA is performed by DNA demethylases (DDMs). Chromatin could suffer noncovalent conformational changes catalyzed by chromatin remodeling factors that consume ATP to provide the energy (Zhou et al. 2016). These factors execute diverse ATP-dependent remodeling of nucleosomes including nucleosome sliding, histone ejection, and histone exchange. All of the above factors are considered as chromatin regulators. Different chromatin regulators cooperate to modulate chromatin structure, playing critical roles in the nuclear processes such as transcription, DNA replication and repair, and recombination. Transcription regulation by chromatin regulators is referred to as epigenetic regulation, which is attracting more attention of the epigeneticists to the mechanism how gene expression is inherited (D'Urso and Brickner 2014).

In plants, most of the chromatin regulators have been identified in the two model plants: Arabidopsis and rice. Those include histone variants, histone chaperones, HATs and HDACs, histone methyltransferases (HMTs) and histone demethylases (HDMs), DMTs and DDMs, and ATP-dependent chromatin remodeling factors (Ahmad et al. 2011; Hu and Lai 2015; Hu et al. 2009, 2013; Liu et al. 2012; Lu et al. 2008; Martignago et al. 2019; Ng et al. 2007; Pandey et al. 2002; Probst et al. 2020; Tripathi et al. 2015; Sharma et al. 2009; Zemach et al. 2010). Genes encoding most of these chromatin regulators are conserved in plant genomes. However, plant-specific factors have been evolved, such as HD2 family of HDACs, CMT family of DMTs and histone variant H2A.W (Bourque et al. 2016; Henikoff and Comai 1998; Yelagandula et al. 2014). In the meantime, several

regulators are lost in plant exemplified by JARID2, KDM2, KDM4, KDM6, and PKDM10 subfamilies of JmjC proteins and some chromatin remodeling factors (Qian et al. 2015; Hu et al. 2013). Functional analyses have revealed that the conserved regulators also exert either conserved or plant-specific epigenetic regulation of gene expression, contributing to diverse plant developmental processes and various stress responses (Wagner 2003; Wang and Qiao 2020).

Sorghum is a cereal crop with characteristic developmental and physiological features as well as high tolerance to a variety of stresses, probably due to its evolution in Africa. We speculate that epigenetic regulation must participate in the shaping of these features. The hypothesis can be validated by clarifying function of the chromatin regulators and their interaction network. DMTs and DDMs have been reported previously in sorghum (Yu et al. 2021). In this study, we identified the other chromatin regulators including 63 histones, 29 histone chaperones, 8 HATs, 19 HDACs, 47 HMTs, 27 HDMs, and 38 chromatin remodeling factors and analyzed their gene expression patterns. The vast majority of these chromatin regulators are common among sorghum, rice and Arabidopsis. However, we found that a few histone proteins such as three H3.3-like histone variants and a H2B variant only existed in several different grass species including sorghum but not rice. In addition, a novel sorghum-specific chromatin remodeling factor that combines the domains of the elongation factor EF-Tu, the histone chaperone SPT16 and the helicase-like region of Snf2 protein was revealed for the first time. We clustered the expression profiles of the chromatin regulators genes into four major groups to establish co-expression network of these genes.

Materials and methods

Identification of the protein sequences of sorghum chromatin regulators

The protein sequences of chromatin regulators from Arabidopsis or rice were used as queries to perform BLASTp in Phytozome v12 database (https://phytozome.jgi.doe.gov/pz/portal.html) and NCBI protein database (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Protein sequences of some genes are predicted differently in the two databases. After comparing the two protein sequences encoded by the same gene and predicted in the two databases with their closest homolog in other species, we chose the more homologous one (higher similarity) of the two for further analysis.

Phylogenetic tree and domain architecture analysis

To confirm the identity of sorghum chromatin regulators, sequence alignment was performed using ClustalX, and

phylogenetic trees were constructed for HDAC, SET, JmjC, MSI, and Snf2 family proteins by MEGA 3.1 using neighborhood-joining method with the following parameters: Poisson correction, complete deletion, and a bootstrap test of 1000 replications. Domain architecture was analyzed in SMART (http://smart.embl.de/) database and visualized by the DOG 2.0 programme.

Collection of expression data

RNA-seq expression data were generated by (Davidson et al. 2012; Makita et al. 2015; Wang et al. 2018). FPKM values of gene expression were downloaded from Plant Expression ATLAS (https://www.ebi.ac.uk/gxa/plant/experiments) database. Log$_2$ FPKM values were calculated and visualized by the Heml programme. Cluster analysis of expression patterns were performed by the Cluster 3.0 program and the results were visualized by the Treeview program.

Plant material, RNA extraction, reverse-transcription and real-time qPCR

BTx623 sorghum variety was used in this experiment to obtain tissue samples at different developmental stages. Total RNAs was extracted using TRIzol reagents (Invitrogen). For RT-PCR analysis, 2 μg total RNAs was treated first with gDNA wiper mix and then reverse transcribed in a total volume of 20 μL with Hiscript II qRT SuperMix II (Vazyme). The resulting products were tested by Real-Time PCR with gene specific primers (Table S1).

Real-time PCR was performed in a total volume of 20 μL with 1.0 μL of the RT, 0.25 μM primers, and 10 μL ChamQ SYBR qPCR Master Mix (Vazyme) on a CFX96 real-time PCR machine (Bio-RAD) according to the manufacturer's instructions. The sorghum *EIF4a* gene was used as the internal control. All primers were annealed at 60℃ and run 40 cycles. The expression level of target genes was normalized with that of *EIF4a*: $2^{(\text{Ct of } EIF4a - \text{Ct of target})}$.

Results and discussion

Histones

H3

In mammals, H3 mainly include H3.1, H3.2, H3.3, and CENH3 (Martire and Banaszynski 2020). H3.1 and H3.2 are canonical histones while H3.3 and CENH3 are histone variants. H3.2 and H3.3 proteins differ from H3.1 by one and five amino acids, respectively, whereas CENH3 shows much more variation. H3.3 is enriched in both euchromatic and heterochromatic regions, suggesting its complicated

role in regulating nucleosome dynamics. CENH3 has been found to localize specifically to centromeres and is crucial for the formation of the kinetochore. In plants, H3 proteins are more variable than animal counterparts. Based on four amino acids that discriminate plant H3.3 from H3.1, H3.1-like, and H3.3-like have also been determined in addition to H3.1, H3.3, and CENH3 (Probst et al. 2020; Hu and Lai 2015). Utilizing the plant H3.1 protein sequence as query we identified 16 H3-encoding genes in sorghum (Table S2). Sequence alignment and phylogenetic analysis indicated that there were eight H3.1 proteins designated as H3.1a–H3.1 h and four H3.3 proteins designated as H3.3a–H3.3d (Fig. 1). We found that a few amino acids were inserted at N-terminal of H3.3d, resulting in lack of Lys4 that is an important modification site in H3 proteins (Fig. S1a). Three proteins carry three of the four amino acids specific to H3.3, thus they are termed H3.3L1–H3.3L3 (Fig. S1a). However, these proteins have many substitutions of amino acids within histone-fold domain compared with the H3.3 proteins, but the substitutions are conserved within the three proteins. As their homologs in rice and Arabidopsis were not found, to clarify if this protein is unique to sorghum BLASTp was performed in plant proteome. Their homologs are present in some of grass species such as *Setaria italica*, *Setaria viridis*, *Digitaria exilis*, *Miscanthus lutarioriparius*, *Eragrostis curvula*, and *Panicum virgatum.* These species are evolutionarily close to each other, so we assume that these genes appear after the origin of the gramineae (data not shown). Finally, sorghum genome encodes one CENfH3 gene which is similar to rice and Arabidopsis (Table S2).

H2A

H2A contains canonical H2A, H2A.Z, H2A.X, MacroH2A, H2A.W, and short H2A variants, making it the most diverse core histone protein (Martire and Banaszynski 2020). H2A.Z is involved in a wide variety of nuclear processes such as transcription regulation, heterochromatin formation, DNA replication and DNA damage repair. In particular, involvement of H2A.Z in transcription regulation seems to be intricate since either active or negative impact of H2A.Z incorporation on transcription has been revealed. Plant lacks MacroH2A and short H2A variants but uniquely possesses H2A.W that is enriched specifically in heterochromatin (Yelagandula et al. 2014). Compared to canonical H2A and animal H2A.Z, plant H2A.Z has a shorter C-terminal tail and a longer N-terminal tail (Hu and Lai 2015). H2A.X and H2A.W have SQEF motif and KSPK motif, respectively, at C-terminal tail (Probst et al. 2020). By alignment of sorghum, rice and Arabidopsis H2A protein sequences and phylogenetic analysis as well as analyzing above specific signatures of each H2A variants, we found that there were two canonical H2As (H2A1 and H2A2), four H2A.Zs
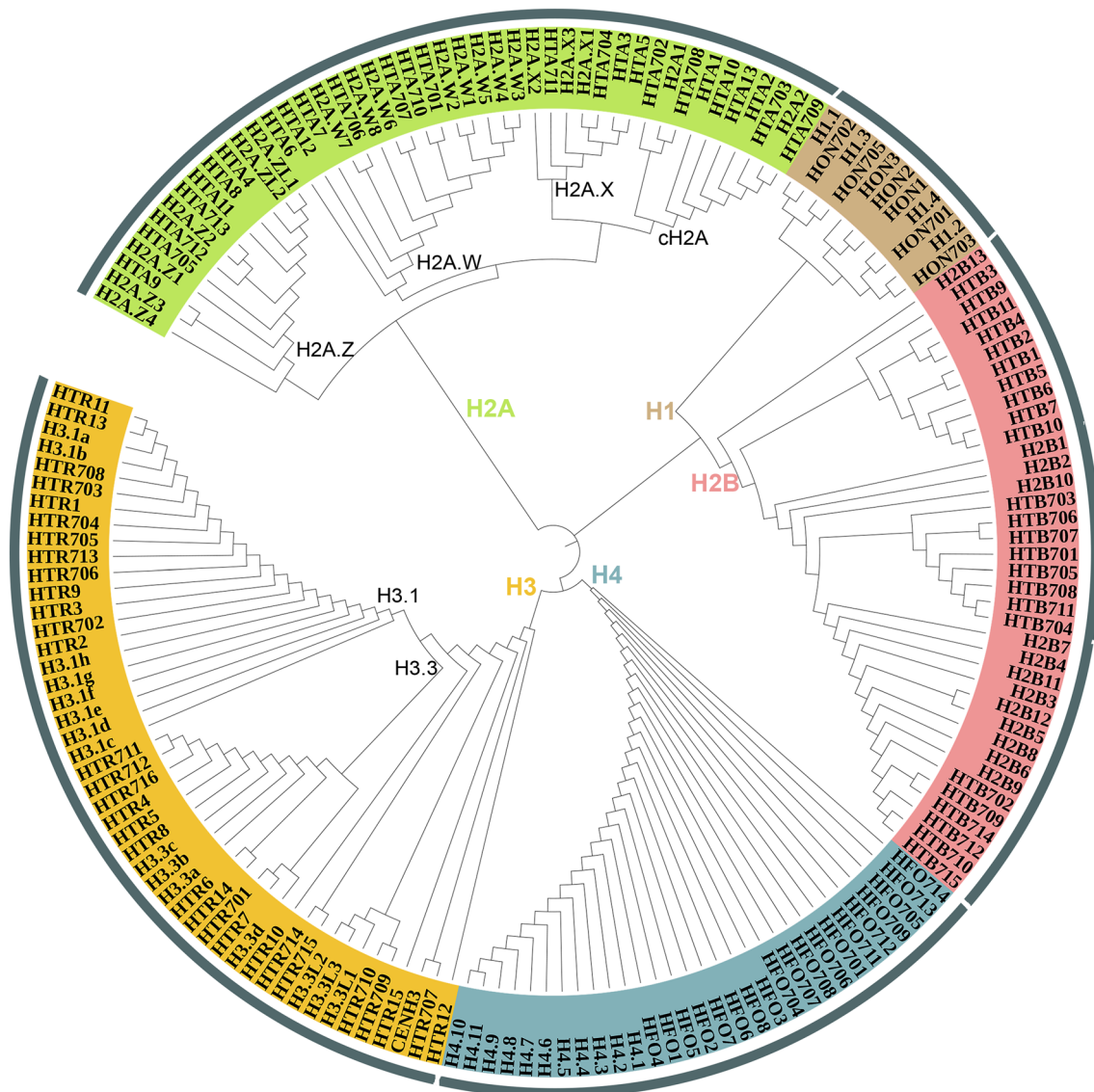
Fig. 1 Phylogenetic analysis of histone proteins. The full sequences of histone proteins from sorghum rice, and Arabidopsis were used for phylogenetic analysis. Five clades of H3, H4, H2A, H2B, and H1 were indicated in the phylogenetic tree. Some histone variants such as H3.1, H3.3, H2A.Z, H2A.X, H2A.W, and cH2A (canonical H2A) were also marked

(H2A.Z1-H2A.Z4), three H2A.Xs (H2A.X1–H2A.X3) and eight H2A.Ws (H2A.W1–H2A.W8) in sorghum (Fig. 1, Fig. S1b, and Table S2). In addition, two truncated H2A proteins with high sequence similarity to H2A.Z at C-terminal were dubbed H2A.ZL1 and H2A.ZL2 (Table S2, Fig. S1b).

H2B

Comparison of plant and mammalian H2B protein sequences has revealed that plant H2Bs present longer N-terminal tail (Hu and Lai 2015). In addition, plant H2B proteins have two variable regions located at N-terminal. Substantial divergence was observed in these regions between rice and

Arabidopsis. Thirteen H2B proteins termed H2B1–H2B13 in sorghum was identified by homologous search and compared with rice and Arabidopsis homologs (Fig. 1, Fig. S2a and Table S2). The result demonstrated that although the primary amino acid sequences in two variable regions of H2B in rice and sorghum showed higher similarity than in Arabidopsis, they were more conserved within each species, suggesting that histone H2B variants appear independently in different lineages (Fig. S2a). Whether divergence of these regions confers functional differentiation of each H2B variants remains to be explored. Furthermore, a H2B protein, H2B13, twice as long as the others was discovered in sorghum (Table S2). It only has conserved histone-fold domain

of H2B at C-terminal of the protein (Fig. S2a). Homologous search indicated that its homologs existed only in some species of the gramineae similar to H3.3L1–H3.3L3. However, H2B13 and H3.3Ls homologs are present in different sets of grass species, demonstrating that two histone variants are not co-evolved.

H4

Histone H4 proteins are the most conserved histones. Functional H4 variants has not been unveiled until recently a nucleoli-localized human H4 variant, H4G, was reported to drive ribosomal RNA transcription by loosening chromatin structure (Martire and Banaszynski 2020). In plants, two H4 variants have been identified in rice and soybean, and yet their functionality has not been confirmed (Hu and Lai 2015). Furthermore, there are only two amino acids different between plant and mammalian canonical H4 proteins. We found that the sorghum genome contains 11 histone H4 genes termed H4.1–H4.11 (Fig. 1 and Table S2), all of which encodes identical proteins with rice and Arabidopsis homologs (data not shown).

H1

Histone H1 proteins are much more diverse than core histone proteins. Eleven H1 variants have been identified in humans (Fyodorov et al. 2018), while Arabidopsis and rice genomes encode only three and four H1 proteins, respectively (Hu and Lai 2015). H1 proteins among different species exhibit low sequence conservation except the global (GH1) domain that is responsible for the binding of H1 to linker DNA. Here, we identified four H1 proteins in sorghum termed H1.1–H1.4, whose sequences outside the global domain were moderately homologous to rice four H1 proteins (Fig. 1, Fig. S2b and Table S2). Plant H1 proteins can be divided into ubiquitously and stably expressed major variants and stress-inducible minor variants (Rutowicz et al. 2015). Structural analysis of H1 proteins in Arabidopsis indicates that the principal difference between the major and the minor variants are the three amino acids in the global domain and S/TPXK motif at N-terminal and CTD of the protein (Rutowicz et al. 2015). The three amino acids in the major variants represented by H1.1/H1.2 (HON1 and HON2) in Arabidopsis are Glu-66, Arg-112, and Ser-116, while those in the minor variants represented by H1.3 (HON3) are Phe-28, Asn-75, and Lys-79. The major variants bear 1–3 S/TPXK motifs at both ends of the protein while the minor variants do not. In addition, the minor variants have the shortened CTD. The structural difference may lead to their functional discrepancy. Indeed, the dynamics of H1.3 nucleosome-binding is considerably higher than that of H1.1/H1.2. Sequence alignment uncovered that the three amino acids in sorghum H1.2 and H1.4

were Ala, Lys and Ala, which is the typical feature of the major variants together with S/TPXK motifs in CTD (Fig. S2b). The corresponding three amino acids in sorghum H1.1 and H1.3 are Phe, Gly, and Lys and they have the shortened CTD, evidencing that they belong to the minor variants despite that H1.3 has S/TPXK motif at N-terminal.

Expression pattern of the histone genes

To investigate expression pattern of the histone genes, we collected RNA-seq data from plant expression ATLAS database. Sixteen organs or tissues were chosen for the expression analysis. However, *H3.1f*, *H3.1 g*, *H3.1 h*, *H3.3b*, *H2A.ZL2* have not been annotated in Phytozome database and they do not have the gene locus number, thus expression data of these five genes could not be obtained from RNA-seq data. We harvested seven samples at different developmental stages and performed RT-qPCR to test the expression of the above five genes. The data from RNA-seq showed that *H3.1a/b/c/d/e* were poorly expressed in leaves, pericarps, endosperm, anther and pollen, while highly expressed in meristems, root, stem, inflorescence, embryo, spikelet, and pistil (Fig. 2), which is consistent with the expression pattern of rice canonical *H3* genes (Hu and Lai 2015). Similarly, RT-qPCR result indicated that *H3.1f/g/h* were expressed at comparatively low levels in young and mature leaves (Fig. S3). In contrast, transcripts of all the *H3.3* genes were accumulated ubiquitously (Fig. 2, Fig. S3). Specifically, the average expression level of *H3.3c* was higher than that of the other *H3.3* genes, indicating that it is the major functional H3.3 in sorghum. Three *H3.3L* genes were not expressed in all the analyzed organs (Fig. 2), implying that they may not be functional genes, which needs to be proved experimentally. Three of 11 H4 genes including *H4.1*, *H4.4*, and *H4.6* were ubiquitously expressed while the expression of *H4.3* showed strong tissue-specificity. The expression pattern of the other *H4* genes was analogous to that of *H3.1* (Fig. 2). For *H2A* genes, *H2A1*, *H2A2*, *H2A.W7*, and *H2A.Z2* were expressed constitutively, while *H2A.X1*, *H2A.X3*, *H2A.W1*, *H2A.W8*, *H2A.Z1*, *H2A.Z4*, and *H2A.ZL2* were more prone to be expressed in the specific tissues (Fig. 2, Fig. S3). We could not detect the expression of *H2A.Z3* and *H2A.ZL1* in any tissues (Fig. 2), suggesting that they might not be functional. Besides, the other H2A genes were expressed in a manner similar to *H3.1* (Fig. 1). We found that *H2B1*, *H2B2*, *H2B4*, *H2B5*, *H2B6*, *H2B7* displayed constitutive expression pattern, while the other *H2B* genes were specifically expressed in some tissues (Fig. 2). Particularly, *H2B8* was only expressed in anther and pollen and *H2B13* was expressed specifically in embryos and meristems, cuing that they may have distinct function from the other *H2B*
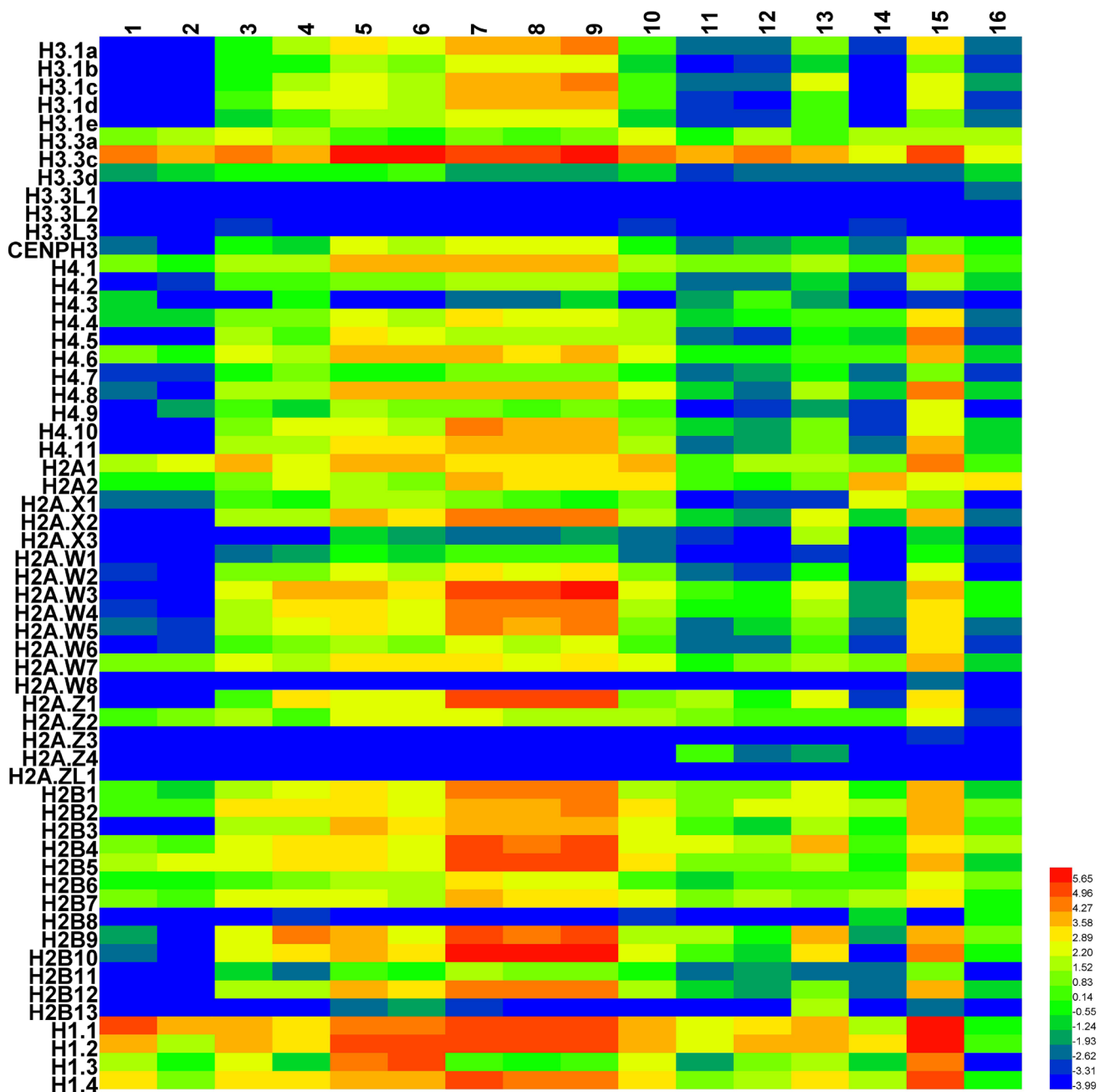
Fig. 2 Expression profiles of histone-encoding genes in sorghum. 1. Young leaves, 2. Mature leaves, 3. Stems, 4. Roots at seedling developmental stage, 5. Vegetative meristem, 6. Floral meristem, 7. Inflorescence (1 to 5 mm), 8. Inflorescence (1 to 10 mm), 9. Inflorescence (1 to 2 cm), 10. Spikelet, 11. Endosperm (20 days after pollination), 12. Pericarp (20 days after pollination), 13. Embryo (20 days after pollination), 14. Anther, 15. Pistil, 16. Pollen (booting stage)

genes. Similar to the expression pattern of the major H1 variants in Arabidopsis (Rutowicz et al. 2015), *H1.2* and *H1.4* in sorghum were expressed in all the chosen tissues (Fig. 2). However, the minor variants were expressed differentially. Sorghum *H1.1* was also ubiquitously expressed like the major variants but *H1.3* was expressed in specific tissues as its counterpart in Arabidopsis (Fig. 2).
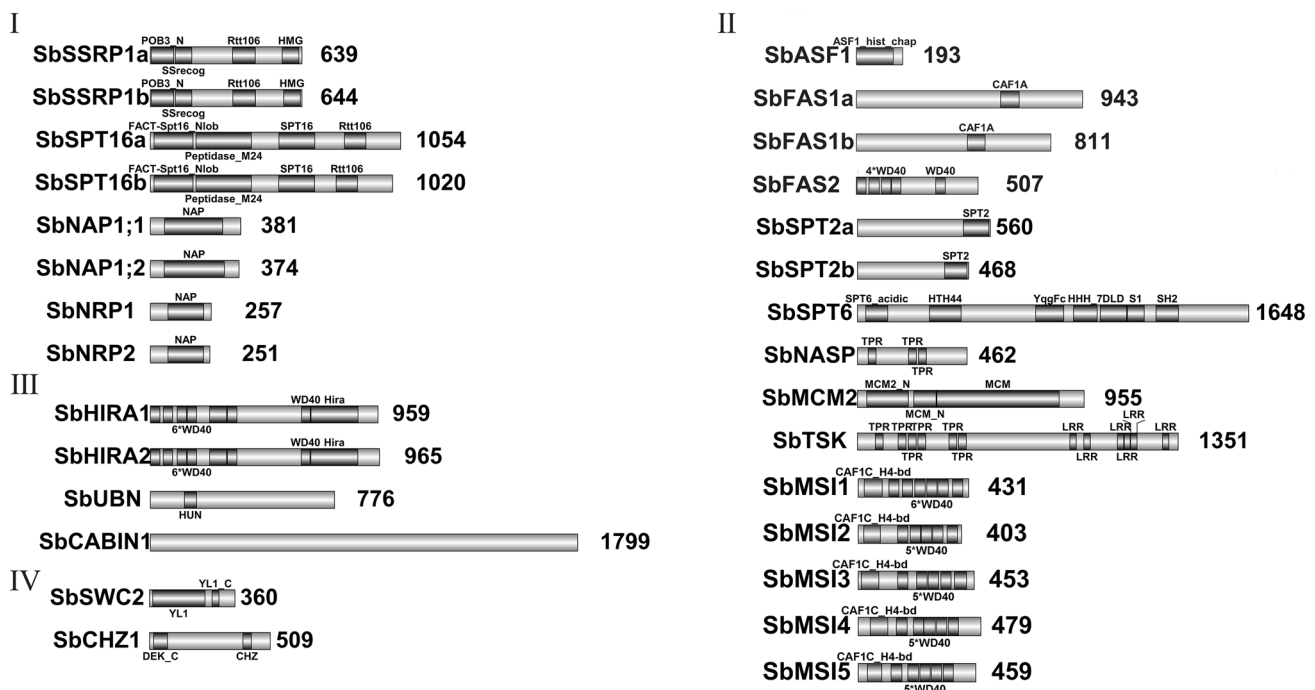
Histone chaperones

Histones are chaperoned by a set of unrelated proteins that use the distinct domains to bind histones (Hammond et al. 2017). Histone chaperones have been systematically identified in Arabidopsis and rice (Tripathi et al. 2015). Here we identified 29 genes encoding histone chaperones in sorghum (Table S3). All the histone chaperones possess conserved

domains as their homologs (Fig. 3). Based on their binding histone substrates which have been reported in other species, we divided histone chaperones into four categories (Fig. 3).

Category I histone chaperones bind both H3–H4 dimer and H2A-H2B dimer, which include FACT complex and NAP family proteins. The FACT complex is composed of SSRP1 and SPT16, each of which is encoded by two genes in sorghum termed *SbSSRP1a/b* and *SbSPT16a/b* (Table S3). Expression analysis indicated that *SbSSRP1a* and *SbSPT16a* were constitutively expressed while *SbSSRP1b* and *SbSPT16b* were specifically expressed in a few tissues (Fig. S4), suggesting that *SbSSRP1a* and *SbSPT16a* constitute the major FACT complex that chaperones histones through sorghum life cycle. NAP1 was initially characterized as a H2A-H2B chaperone, but it has been proved to have H3–H4 binding capacity in Arabidopsis (Gonzalez-Arzola et al. 2017). NAP1 family in sorghum comprises four members named SbNAP1;1, SbNAP1;2, SbNRP1 and SbNRP2 (Table S3). All four *NAP1* genes are highly expressed in all the tissues except pollen, demonstrating their expansive chaperone function in sorghum (Fig. S4).

Category II histone chaperones only bind H3-H4 dimer and include ASF1, CAF1 complex, SPT2, SPT6, NASP, MCM2, and TONSL. ASF1 plays a central role in the delivery of newly synthesized H3-H4 dimer from the cytoplasm to the nucleus and hands over H3.1–H4 dimer and H3.3-H4 dimer to CAF and HIRA complexes, respectively, for

deposition (Hammond et al. 2017). ASF1 are encoded by two genes in Arabidopsis and rice (Tripathi et al. 2015). Two *ASF1* genes *AtASF1a* and *AtASF1b* in Arabidopsis have redundant function in chromatin replication and controlling plant development (Zhu et al. 2011). However, we only discovered one ASF1-encoding gene termed *SbASF1* in sorghum (Table S3), suggesting that sorghum genome has lost the other copy of the ASF1-encoding gene. Additionally, *SbASF1* was poorly expressed in all the analyzed tissues (Fig. S4). We speculate that the low expression level of *SbASF1* reflects that it may be expressed in certain cell types. The CAF1 complex contains FAS1 (CHAF1A), FAS2(CHAF1B), and MSI(RBAP46/48) (Hammond et al. 2017). Sorghum genome contains two *FAS1* genes (*SbFAS1a* and *SbFAS1b*), one *FAS2* gene (*SbFAS2b*), and five *MSI* genes (*SbMSI1-5*) (Table S3). Five MSI proteins belong to four clades when we constructed phylogenetic tree using MSI homologs from four plant species (Fig. S5). MSI is also a subunit of HDAC and Polycomb repressive complex 2 (PRC2) complexes. It is of great interest to reveal whether all of these MSI proteins function redundantly or separately in the three complexes. However, we found that only *SbMSI2* was expressed in a pattern similar to *SbFAS1a*, *SbFAS1b*, and *SbFAS2*, while *SbMSI1*, *SbMSI3*, and *SbMSI4* were highly expressed and *SbMSI5* was poorly expressed in most tissues (Fig. S4). This suggests that they play unequal roles in chromatin regulation. SPT2 is encoded by two genes
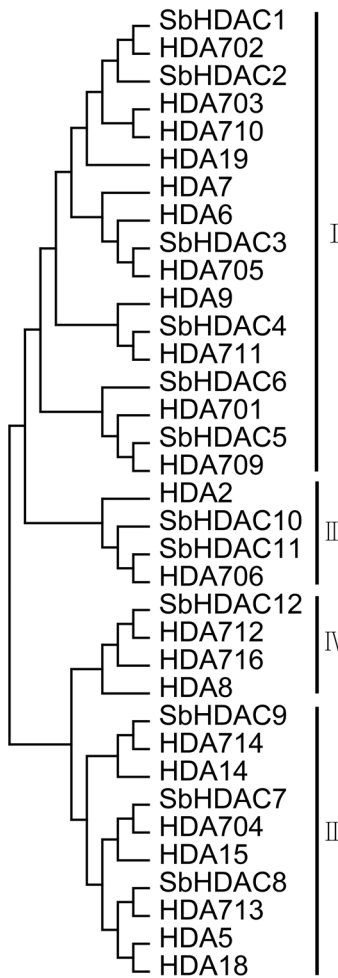

Fig. 3 Domain architecture of histone chaperones. Based on their binding histone substrates, histone chaperones are divided into four categories: I, II, III, and IV. Histone chaperones in Category I bind both H3H4 and H2A–H2B. Histone chaperones in Category II bind only H3–H4. Histone chaperones in Category III and Category IV bind H3.3-H4 and H2A.Z-H2B respectively
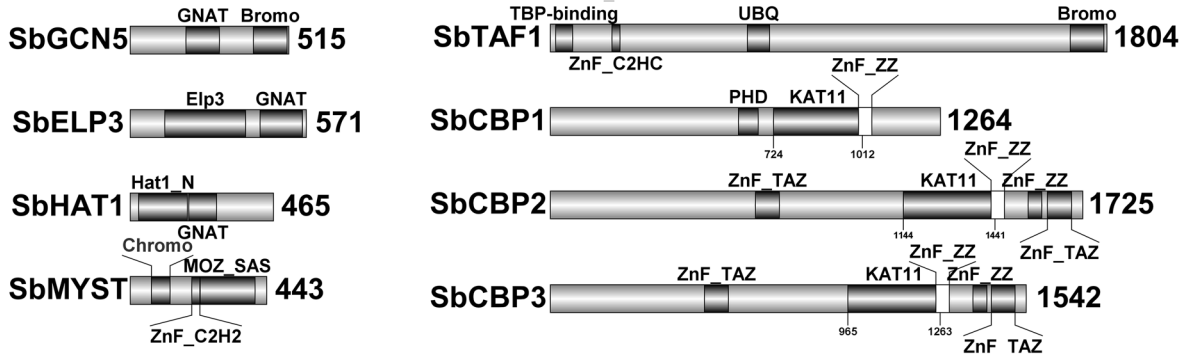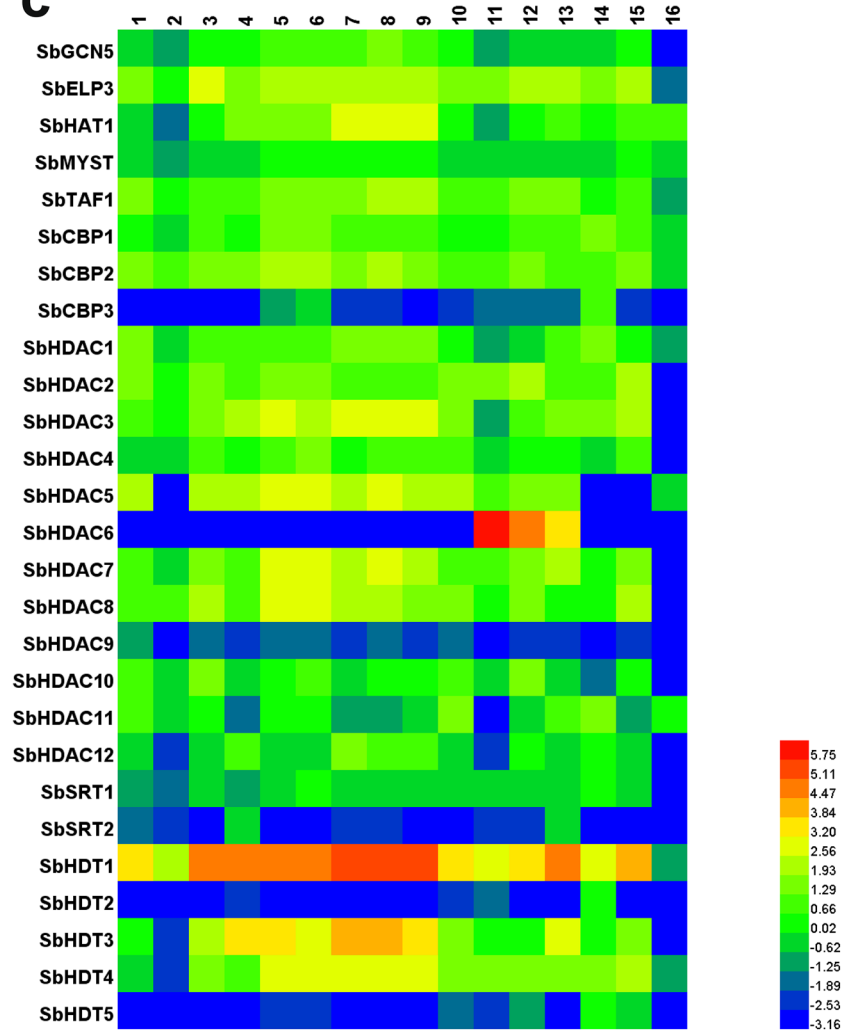
termed *SbSPT2a/b*, and each of SPT6, NASP, MCM2, and TONSL is encoded by a single gene designated as *SbSPT6*, *SbNASP*, *SbMCM2*, and *SbTSK* (Table S3). *SbSPT2b* transcripts accumulate in a few tissues whereas *SbSPT2a* is ubiquitously expressed (Fig. S4), which indicates that the latter is the universal SPT2 chaperone in sorghum. The constitutive expression pattern was also observed for *SbSPT6*

and *SbNASP* (Fig. S4). *SbMCM2* is not expressed in leaves and endosperm and *SbTSK* prefers to be expressed in meristems and inflorescences (Fig. S4).

Category III including HIRA complex binds H3.3-H4 specifically and category IV including SWC2 and CHZ1 bind H2A.Z-H2B. HIRA complex is composed of HIRA, UBN, and CABIN1, which are encoded by *SbHIRA1/2*,

◀**Fig. 4** Identification and expression analysis of histone acetyltransferases (HATs) and histone deacetylases (HDACs) in sorghum. a. Domain architecture of sorghum HATs. b. Phylogenetic tree of HDACs. The sequences of HDAC domains from sorghum, Arabidopsis and rice were used for phylogenetic analysis. Four classes of HDACs were indicated on the right side. Arabidopsis HDAC gene locus numbers: *HDA2*, At5g26040; *HDA5*, At5g61060; *HDA6*, At5g63110; *HDA7*, At5g35600; *HDA8*, At1g08460; *HDA9*, At3g44680; *HDA14*, At4g33470; *HDA15*, At3g18520; *HDA18*, At5g61070; *HDA19*, At4g38130. Rice HDAC gene locus numbers: *HDA701*, LOC_Os01g40400; *HDA702*, LOC_Os06g38470; *HDA703*, LOC_Os02g12350; *HDA704*, LOC_Os07g06980; *HDA705*, LOC_Os08g25570; *HDA706*, LOC_Os06g37420; *HDA707*, LOC_Os01g12310;*HDA709*, LOC_Os11g09370; *HDA710*, LOC_Os02g12380; *HDA711*, LOC_Os04g33480; *HDA712*, LOC_Os05g36920; *HDA713*, LOC_Os07g41090; *HDA714*, LOC_Os12g08220; *HDA716*, LOC_Os05g36930. c. Expression profiles of HATs and HDACs-encoding genes in sorghum. 1. Young leaves, 2. Mature leaves, 3. Stems, 4. Roots at seedling developmental stage, 5. Vegetative meristem, 6. Floral meristem, 7. Inflorescence (1 to 5 mm), 8. Inflorescence (1 to 10 mm), 9. Inflorescence (1 to 2 cm), 10. Spikelet, 11. Endosperm (20 days after pollination), 12. Pericarp (20 days after pollination), 13. Embryo (20 days after pollination), 14. Anther, 15. Pistil, 16. Pollen (booting stage)

SbUBN, and *SbCABIN1* in sorghum (Table S3). SWC2 is a subunit of SWR1 complex that is a chromatin remodeling complex responsible for exerting H2A.Z deposition into nucleosome. Chz1 was identified as a specific chaperone of H2A.Z in yeast but their homolog in mammals does not exist(Hammond et al. 2017). Recently, the function of its rice homolog *Oschz1* has been unveiled (Du et al. 2020). *SbSWC2* and *SbCHZ1* were identified in sorghum to encode the two histone chaperones, respectively (Table S3). All the category III and IV genes show high levels of expression in all the tissues except pollen, reflecting their comprehensive roles in plant development (Fig. S4).

HATs and HDACs

In plants, the HATs can be grouped into four classes: General control nondepressible 5 (GCN5)-related Acetyl Transferase (GNAT), MOZ-YBF2/SAS3-SAS2/TIP60 (MYST), cAMP-responsive element Binding Protein (CBP), and TATA-binding protein Associated Factor 1 (TAF1) (Pandey et al. 2002). We found eight HATs in sorghum, which we named after their Arabidopsis homologs (Table S4). These include three GNATs: SbGSN5, SbELP3, and SbHAT1, one MYST: SbMYST, one TAF1: SbTAF1, three CBPs: SbCBP1-3. These enzymes have specific conserved domains in addition to the catalyzing domain (Fig. 4a). Some of these domains such as bromodomain and chromodomain recognize modified histone tails. Expression analysis indicated that all HATs genes except *SbCBP3* were moderately expressed in nearly all the tissues (Fig. 4c). A lower expression level of *SbCBP3* was only detected in meristems and anther, arguing that it executes HATs function in these tissues.

The HDACs can be classified into three families: Reduced Potassium Dependency 3 (RDP3)/Histone DeAcetylase 1 (HDA1), Silent Information Regulator 2 (SIR2), and the plant-specific Histone Deacetylase 2 (HD2) (Pandey et al. 2002). In addition, based on their homology to yeast HDACs the plant RPD3/HDA1 family is divided into four classes (I, II, III, and IV) (Ueda et al. 2017). In sorghum, RPD3/HDA1 family contains 12 members named SbHDAC1-12, including six class I members, three class II members, two class III members and one class IV member grouped on the basis of phylogenetic analysis (Table S4, Fig. 4b). Two homologs of HDA19 that is involved in multiple developmental processes and stress response in Arabidopsis were identified in sorghum: SbHDAC1 and SbHDAC2. Three homologs of HDA19 in rice have been reported to have redundant function during plant development (Hu et al. 2009). SbHDAC1 and SbHDAC2 showed 89% sequence similarity, which suggests that they may also function overlappingly. Most of class I HDACs were expressed ubiquitously except *SbHDAC6* that was only highly expressed in developing seeds including embryo, endosperm and pericarp (Fig. 4c). Class II members *SbHDAC7* and *SbHDAC8* were also constitutively expressed whereas *SbHDAC9* was poorly expressed in all the tissues. Similarly, *SbHDAC10/11/12* were expressed in most of the tissues (Fig. 4c). Sorghum genome encodes two SIR2 genes named *SbSRT1* and *SbSRT2* and five HD2 genes named *SbHDT1-5* (Table S4). We found that *SbSRT1* was expansively expressed whereas *SbSRT2* was only expressed in root and embryo (Fig. 4c). All sorghum HD2 proteins possess conserved domains including HDAC domain, acidic region (acidic R) and C2H2 type Zinc finger (Fig. S6). However, SbHDT1 and SbHDT2 protein sequences were highly similar to each other although their expression patterns were completely different. *SbHDT1* expression was much higher and more comprehensive, and *SbHDT2* was only moderately expressed in anthers (Fig. 4c). SbHDT4 and SbHDT5 were relatively less conserved in comparison with the other sorghum HD2 proteins. They lack MEFW motif that was reported to be a specific feature of HD2 proteins and KKxK monopartite NLS important for nucleus-localization of the protein (Fig. S6) (Bourque et al. 2016). Thus, biochemical and genetic evidences are required to validate their HDAC activity. *SbHDT4* was expressed in most of the tissues except mature leaves and pollen which is similar to *SbHDT3* (Fig. 4c). *SbHDT5* was moderately expressed in anthers and pistils (Fig. 4c).

HMTs and HDMs

In plants, SET domain family proteins could be divided into seven classes based on their sequence homology and phylogenetic relationships (Ng et al. 2007). The classification of the enzymes to some extent reflects the substrate

specificity. We discovered 34 SET proteins belonging to the seven classes in sorghum genome (Table S4, Fig. 5a). Class I SET proteins which catalyze H3K27 tri-methylation carry SANT and CXC domains in addition to the SET domain (Fig. 6a). The members in this class are encoded by three genes in Arabidopsis: *MEDEA(MEA), CURLY LEAF (CLF),* and *SWINGER (SWN)*. In sorghum, two genes termed *SbEZH1* and *SbEZH2* were identified to be the homologs of *CLF* and *SWN*. *SbEZH1* RNAs are accumulated more in meristems and inflorescence, while *SbEZH2* transcripts are evenly distributed in most of the tissues except pollen (Fig. S7), suggesting both their functional redundancy and divergence. Class II SET proteins containing AWS and postSET domains methylate H3K36 (Fig. 6a). Some members have an additional PHD or Zf-CW domain preceding the AWS domain. There are five class II members named SbASHH1-3, SbASHR3 and SbASHR4 in sorghum. SbASHR4 is a truncated protein devoid of PHD domain, which might be the paralog of SbASHR3 according to phylogenetic and domain analyses (Figs. 5a, 6a). Moreover, it is only weakly expressed in anthers and pollens (Fig. S7), suggesting its specific role during plant development. The other four members-encoding genes were expressed more expansively. Class III SET proteins catalyze H3K4 methylation and could further divided into four groups. Members in each group carry different additional domains (Fig. 6a). In sorghum, group 1 contains SbATX1 corresponding to Arabidopsis SDG27 (ATX1) and SDG30 (ATX2). They bear PWWP, FYRN, FYRC, and PHD domains. Group 2 contains SbATX2 and SbATX3 that are in the same group with Arabidopsis SDG14 (ATX3), ADG16 (ATX4), and SDG29 (ATX5). They have PWWP and PHD domains. SbATXR3 in group 3 and SbATXR5 in group 4 are homologous to SDG2 and SDG25, respectively. We found another protein in this class termed SbATXR6 carrying PHD domain, which has no ortholog in Arabidopsis, therefore, was classified into group 5. All the genes in this class were expressed in the majority of the tissues (Fig. S7). Class IV SET proteins contain only one member named SbATXR4 in sorghum, which have two homologs in Arabidopsis and rice. The proteins in this class are phylogenetically distant from the other SET proteins, although they also possess the PHD domain present in class III proteins (Figs. 5a, 6a). And *SbATXR4* expression was lower at many developmental stages (Fig. S7), reflecting that it was expressed in a few cells as its homologs in Arabidopsis: *ATXR5* and *ATXR6*, which have been reported to accumulate in specific cells and be cell-cycle regulated (Raynaud et al. 2006). SUVHs and SUVRs, having specificity for H3K9 methylation, are assigned to class V. However, the major enzymes characterized in Arabidopsis are SDG33 (SUVH4), SDG9 (SUVH5), and SDG23 (SUVH6) as well as their three homologs in rice (Qin et al. 2010). Fifteen SUVHs and five SUVRs were identified here in sorghum.

We found six SUVHs (SbSVUH1-6) that were closely homologous to SDG9 and SDG23 and two homologs of SDG33 (SbSUVH7 and SbSUVH8) (Fig. 5a), suggesting sorghum genome evolves more H3K9 methyltransferases than Arabidopsis and rice. Nevertheless, SbSUVH4 protein is shorter than the others and it was not expressed in all the tissues (Fig. S7), thus its functionality needs to be confirmed by experimental evidences although it contains SRA, pre-SET and postSET domains characteristic to class V (Fig. 6a). Besides, the other SUVHs and SUVRs were expressed in one or more tissues (Fig. S7). Class VI is represented by SbATXR1, SbATXR2, SbASHR1, and SbASHR2 and Class VII is represented by SbSDG40. The SET domain in these two classes is truncated or interrupted. Expression analysis indicated that *SbASHR2* and *SbSDG40* were expressed more ubiquitous than the other three genes (Fig. S7).

PRMTs can be divided into four types based on the formation of different methylarginines. Nine and eight PRMTs have been identified in Arabidopsis and rice, respectively (Ahmad et al. 2011). We found that sorghum genome encoded eight PRMTs containing SbPRMT1, SbPRMT3, SbPRMT4, SbPRMT5, SbPRMT6a, SbPRMT6b, SbPRMT7, and SbPRMT10, corresponding to the eight PRMTs in rice (Table S4). Arabidopsis genome has two *PRMT1s*, two *PRMT4s* and sole *PRMT6*, suggesting that duplication of these genes occurs after the divergence of monocots and eudicots. All of the sorghum PRMT genes are not expressed in pollens but in all the other tissues, and *SbPRM6a*, *SbPRMT6b*, *SbPRMT7,* and *SbPRMT10* are not expressed in another one or two tissue, indicating that sorghum PRMTs appear to function universally (Fig. S7).

On the basis of phylogenetic and motif analyses, the eukaryotic JmjC genes can be divided into 14 subfamilies, while plant genomes bear nine of them including JMJD6, KDM3, KDM5, PKDM7, PKDM8, PKDM9, PKDM11, PKDM12, and PKDM13. Among these subfamilies, PKDM7, PKDM8, and PKDM9 are plant-specific (Qian et al. 2015). Similar to the feature of SET domain proteins described above, the JmjC proteins in different subfamilies showed substrate specificities for residues of histone tails to a less extent and have characteristic domain organizations. We identified 23 JmjC proteins in sorghum and divided them into nine subfamilies by phylogenetic and domain analyses (Figs. 5b, 6b, Table S4). SbPKDM11, SbPKDM12A/B/C, and SbPKDM13 solely have the JmjC domain and SbJMJD6A/B have an additional F-box domain. *SbPKDM12B* is extremely highly expressed in young leaves but not mature leaves, pointing to its potential roles in leaf development. *SbJMJD6B* transcripts were detectable universally while *SbJMJ6A* was detected in none of the tissues (Fig. S7). The expression of *SbPKDM11* is also restricted to several tissues and lower than *SbJMJ6B* (Fig. S7), which render us to speculate that SbJMJ6B is the major JmjC protein

a

b
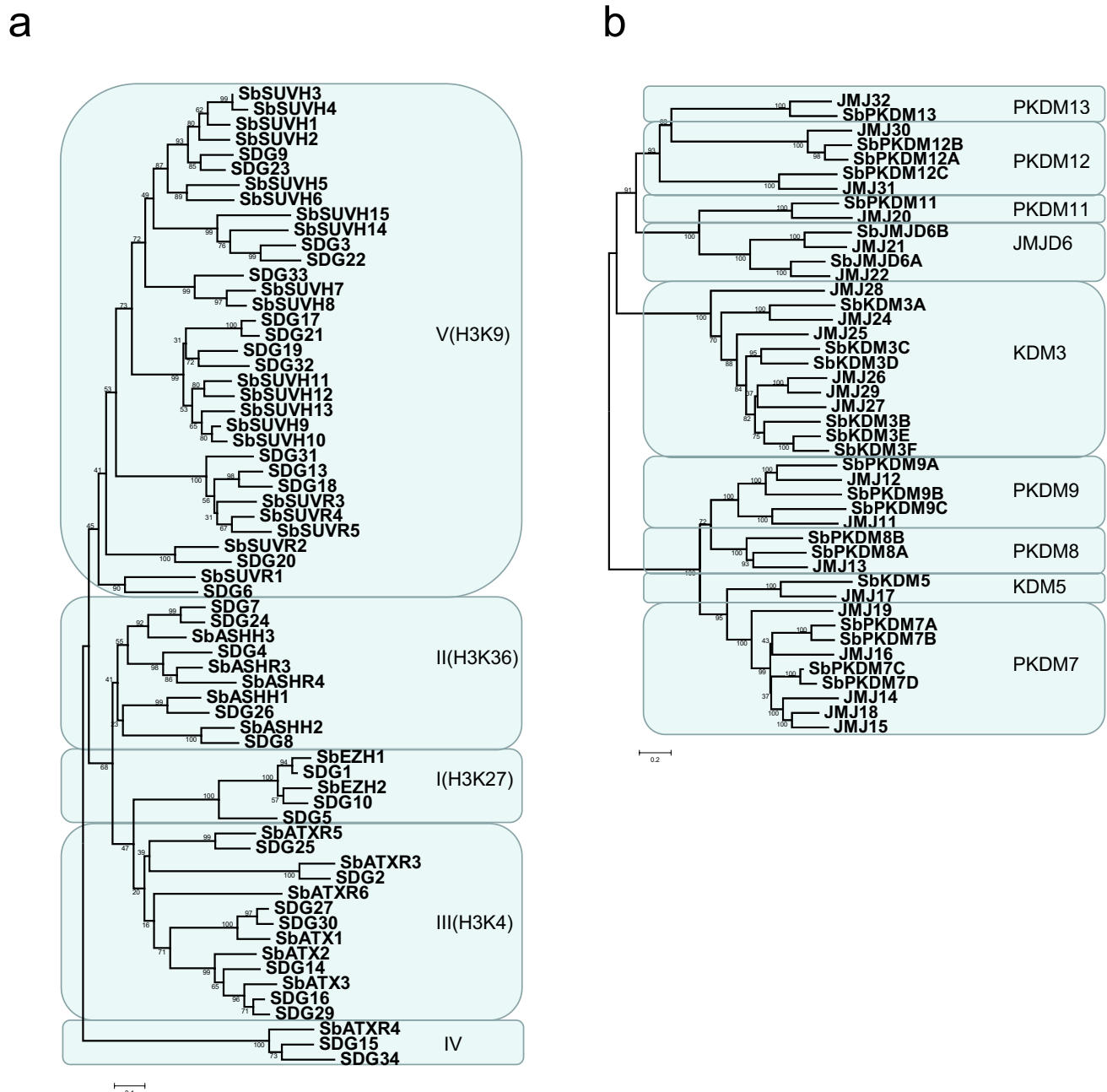
Fig. 5 Phylogenetic analysis of SET domain family proteins and Jumonji C (JmjC) domain family proteins. a. Phylogenetic tree of SET domain proteins. SET domain sequences of Class I-V SET proteins from sorghum and Arabidopsis were used for phylogenetic analysis. Arabidopsis SET gene locus numbers: *SDG5*, At1g02580; *SDG1*, At2g23380; *SDG10*, At4g02020; *SDG7*, At2g44150; *SDG24* At3g59960; *SDG4*, At4g30860; *SDG26*, At1g 76,710; *SDG8*, At1g77300; *SDG27*, At2g31650; *SDG30*, At1g05830; *SDG14*, At3g61740; *SDG16*, At4g27910; *SDG29*, At5g53430; *SDG2*, At4g15180; *SDG25*, At5g42400; *SDG15*, At5g09790; *SDG34*, At5g24330; *SDG32*, At5g04940; *SDG19*, At1g73100; *SDG17*, At1g17770; *SDG21*, At2g24740; *SDG33*, At5g13960; *SDG23*, At2g22740; *SDG3*, At2g33290; *SDG22*, At4g13460; *SDG20*,

At3g03750; *SDG9*, At2g35160; *SDG13*, At1g04050; *SDG18*, At5g43990; *SDG31*, At3g04380; *SDG6*, At2g23740. b. Phylogenetic tree of JmjC domain proteins. The full sequences of nine subfamilies JmjC proteins from sorghum and Arabidopsis were used for phylogenetic analysis. Arabidopsis JmjC gene locus numbers: *JMJ22*, AT5G06550; *JMJ21*, AT1G78280; *JMJ24*, AT1G09060; *JMJ28*, AT4G21430; *JMJ25*, AT3G07610; *JMJ27*, AT4G00990; *JMJ26*, AT1G11950; *JMJ29*, AT1G62310; *JMJ17*, AT1G63490; *JMJ19*, AT2G38950; *JMJ14*, AT4G20400; *JMJ15*, AT2G34880; *JMJ16*, AT1G08620; *JMJ18*, AT1G30810; *JMJ13*, AT5G46910; *JMJ12*, AT3G48430; *JMJ11*, AT5G04240; *JMJ20*, AT5G63080; *JMJ30*, AT3G20810; *JMJ31*, AT5G19840; *JMJ32*, AT3G45880
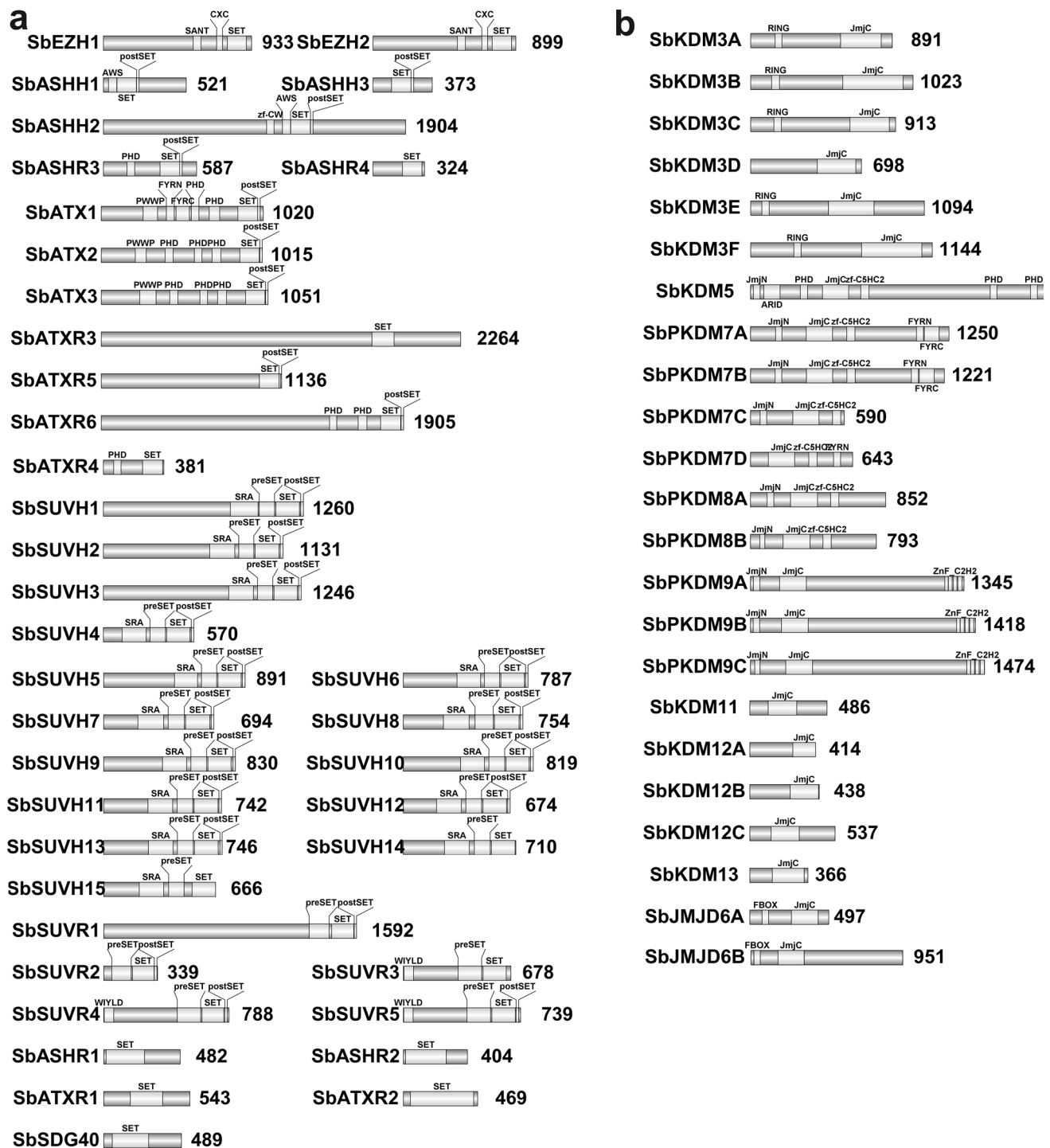
Fig. 6 Domain architecture of SET domain proteins (**a**) and JmjC domain proteins (**b**) in sorghum

exerting histone arginine demethylation, because only the homologs of JMJD6 and PKDM11 in Arabidopsis have been reported to have this catalytic activity. All six SbKDM3s bear an additional Ring domain except SbKDM3D that is shorter than the others (Fig. 6b). We found that *SbKDM3D* was only expressed in endosperm and pericarp, implying its

specific role in these tissues (Fig. S7). All the members of the rest four subfamilies contain JmjN domain as well as the other domains referring to PHD and zf_C5HC2 in KDM5, FYRN, FYRC and zf_C5HC2 in PKDM7, zf_C5HC2 in PKDM8, and ZnF_C2C2 in PKDM9 (Fig. 6b). However, SbPKDM7C and SbPKDM7D are truncated proteins

carrying parts of the conserved domains compared to SbP-KDM7A and SbPKDM7B.

LSD1 is encoded by a single copy gene in human and Drosophila genome, whereas both Arabidopsis and rice have four homologs of LSD1, termed LSD1, LDL1, LDL2, LDL3. LDL3 is less similar to the other three paralogs (Martignago et al. 2019). Similarly, homologous search revealed that there were four LSD1 homologs in sorghum corresponding to the four genes, which we termed *SbLSD1*, *SbLDL1*, *SbLDL2* and *SbLDL3* (Table S4). We found that all these genes were not expressed in pollens (Fig. S7). Low expression level was also detected in mature leaves and endosperm for *SbLSD1*. The other three genes were moderately expressed in all the rest of the tissues (Fig. S7).
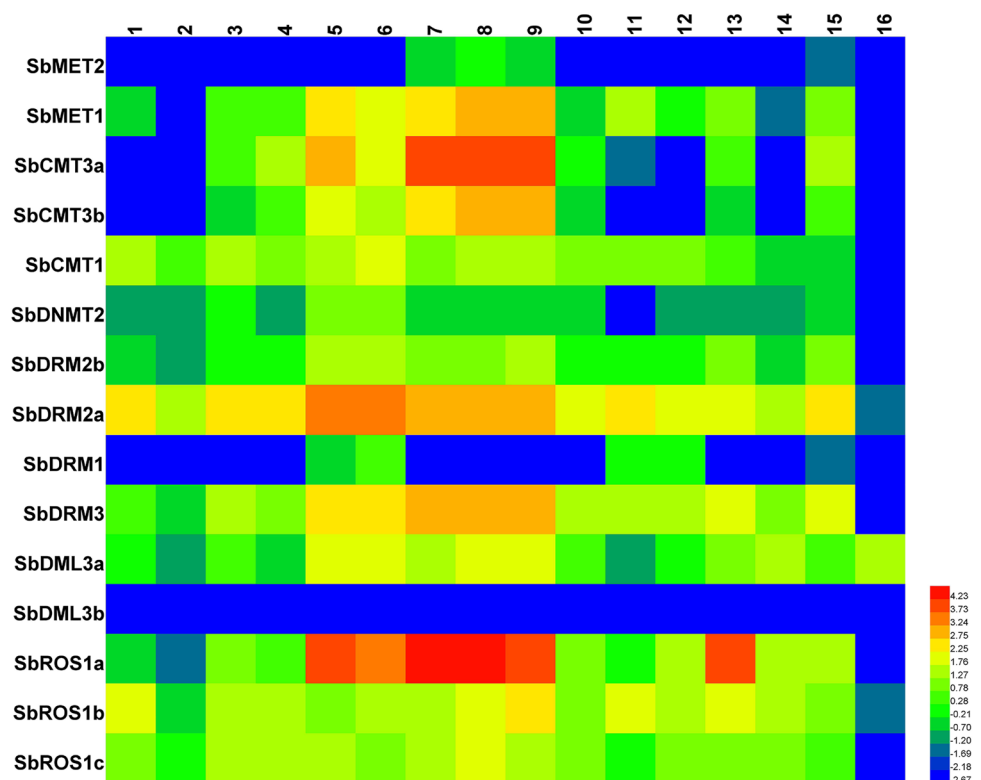
Expression profiles of DMTs and DDMs

DMT and DDM in sorghum have been identified recently (Yu et al. 2021). Ten DMTs were divided into four subfamilies and four DDMs were divided into two subfamilies. However, after BLASTp was performed in sorghum proteome database we discovered another SbDML3, which we named SbDML3b and changed the SbDML3 identified previously to SbDML3a. SbDML3b shows moderate similarity with SbDML3a although it has the conserved DME domains. The expression of *SbDML3b* could not be detected in any of the tissues while *SbDML3a* is expressed constitutively, suggesting that SbDML3a is a universal DNA demethylase (Fig. 7). All DMT genes are not expressed in pollens (Fig. 7). *SbMET1* is expressed specifically in inflorescence and accumulation of *SbMET2* transcripts is more expansive (Fig. 7), which is consistent with the expression pattern of rice homologs *OsMET1-2* and *OsMET1-1*. *SbCMT3a* and *SbCMT3b* are expressed in a similar pattern, revealing the possibility of their functional redundancy (Fig. 7). Specific expression profile was also observed for *SbDRM1*, whose transcripts were more abundant in meristems, endosperm and pericarp (Fig. 7). The other *SbDRM* genes were expressed ubiquitously and at moderate or high levels (Fig. 7), similar to the expression pattern of all the three *SbROS1* genes.

Chromatin remodeling factors

Chromatin remodeling factors are Snf2 family proteins defined by several unique features including a number of conserved motifs and blocks within the helicase-like region. Snf2 family proteins can be divided into six groups, which further subdivided into several subfamilies based on the sequence conservation of the helicase-like region (Hu et al. 2013). The helicase-like region contains two conserved domains: DEADc and HELICc. Members in different subfamilies have additional domains, which may confer diverse function. We identified 38 Snf2 family proteins in sorghum belonging to six groups by phylogenetic analysis and domain organization including: 11 in Snf2-like group, 3 in Swr1-like

Fig. 7 Expression profiles of DNA methyltransferases (DMTs) and DNA demethylases (DDMs)-encoding genes in sorghum. 1. Young leaves, 2. Mature leaves, 3. Stems, 4. Roots at seedling developmental stage, 5. Vegetative meristem, 6. Floral meristem, 7. Inflorescence (1 to 5 mm), 8. Inflorescence (1 to 10 mm), 9. Inflorescence (1 to 2 cm), 10. Spikelet, 11. Endosperm (20 days after pollination), 12. Pericarp (20 days after pollination), 13. Embryo (20 days after pollination), 14. Anther, 15. Pistil, 16. Pollen (booting stage)

group, 7 in Rad54-like group, 10 in Rad5/16-like group, 4 in SSO1653-like, 2 in distant group (Figs. 8, 9, Table S5). Interestingly, we found a novel snf2 protein encoded by *Sobic.004G007133*, which integrates FACT-Spt16_Nlob and Peptidase_M24 domains from SPT16, GTP_EFTU domain from EF-Tu, DEXDc and HELICc from Snf2 protein (Fig. 9). Phylogenetic analysis demonstrates that the helicase-like region of the protein originates from ERCC6 subfamily members (Fig. 8). This protein seems unique to sorghum as its homolog in other species could not be found. It is of great interest to uncover whether combination of the domains from the three proteins could confer the novel protein unique function. Indeed, the gene encoding the novel

protein is expressed in multiple tissues, arguing that it might be functional (Fig. S8).

The members in Snf2-like group and Swr1-like group are the best studied Snf2 proteins. Their function in plants has been unveiled extensively. Many members in these two groups, such as SWI/SNF2, ISWI, SWR1, and INO80, catalyze chromatin remodeling by constituting multi-subunit complex. Plant SWI/SNF2 proteins are encoded by three genes, one of whose products BRM is closely related to the animal ortholog that contain bromodomain at C-terminal. The other two proteins that do not possess bromodomain appear to be plant-specific. Three SNF2 proteins designated as SbSNF2a/b/c in sorghum correspond to their rice and

◄**Fig. 8** Phylogenetic analysis of Snf2 family proteins. The full sequences of Snf2 family proteins from sorghum, Arabidopsis and rice were used for phylogenetic analysis. Arabidopsis Snf2 gene locus numbers: *CHR2*, AT2G46020; *CHR3*, AT2G28290; *CHR12*, AT3G06010; *CHR23*, AT5G19310; *CHR11*, AT3G06400; *CHR17*, AT5G18620; *CHR1*, AT5G66750; *CHR10*, AT2G44980; *CHR5*, AT2G13370; *CHR6*, AT2G25170; *CHR4*, AT5G44800; *CHR7*, AT4G31900; *CHR13*, AT3G12810; *CHR21*, AT3G57300; *CHR19*, AT2G02090; *CHR25*, AT3G19210; *CHR20*, AT1G08600; *CHR35*, AT2G16390; *CHR34*, AT2G21450; *CHR38*, AT3G42670; *CHR42*, AT5G20420; *CHR31*, AT1G05490; *CHR40*, AT3G24340; *CHR22*, At5g05130; *CHR29*, At5g22750; *CHR32*, At5g43530; *CHR37*, At1g05120; *CHR41*, At1g02670; *CHR26*, At3g16600; *CHR27*, At3g20010; *CHR28*, At1g50410; *CHR30*, At1g11100; *CHR33*, At1g61140; *CHR36*, At2g40770; *CHR39*, At3g54460; *CHR16*, At3g54280; *CHR8*, At2g18760; *CHR9*, At1g03750; *CHR24*, At5g63950; *CHR14*, At5g07810; *CHR18*, At1g48310. Rice Snf2 gene locus numbers: *CHR707*, LOC_Os02g02290; *CHR720*, LOC_Os06g14406; *CHR719*, LOC_Os05g05230; *CHR727*, LOC_Os05g05780; *CHR728*, LOC_Os01g27040; *CHR741*, LOC_Os03g51230; *CHR746*, LOC_Os09g27060; *CHR711*, LOC_Os03g01200; *CHR705*, LOC_Os07g46590; *CHR702*, LOC_Os06g08480; *CHR729*, LOC_Os07g31450; *CHR703*, LOC_Os01g65850; *CHR709*, LOC_Os02g46450; *CHR732*, LOC_Os03g22900; *CHR714*, LOC_Os04g47830; *CHR733*, LOC_Os02g52510; *CHR717*, LOC_Os10g31970; *CHR722*, LOC_Os07g49210; *CHR730*, LOC_Os03g06920; *CHR736*, LOC_Os07g25390; *CHR737*, LOC_Os06g14440; *CHR740*, LOC_Os02g43460; *CHR742*, LOC_Os05g32610; *CHR743*, LOC_Os08g14610; *CHR724*, LOC_Os07g44800; *CHR710*, LOC_Os02g32570; *CHR735*, LOC_Os04g09800; *CHR731*, LOC_Os07g32730; *CHR706*, LOC_Os01g57110; *CHR715*, LOC_Os04g53720; *CHR725*, LOC_Os08g08220; *CHR739*, LOC_Os07g48270; *CHR708*, LOC_Os01g72310; *CHR701*, LOC_Os02g06592; *CHR704*, LOC_Os01g01312; *CHR713*, LOC_Os05g15890; *CHR712*, LOC_Os04g59620; *CHR745*, LOC_Os01g44990; *CHR726*, LOC_Os07g40730; *CHR721*, LOC_Os07g44210

Arabidopsis homologous proteins (Fig. 8). ISWI proteins with HAND, SANT, and SLIDE domains are encoded by two genes with highly sequence similarity in Arabidopsis and rice. Two ISWI genes are probably produced by duplication event evolutionarily. It is likely that the duplication of ISWI genes occurs after divergence of monocot and eudicot according to the phylogenetic tree (Fig. 8). The duplication event does not take place in sorghum genome as it has only single copy gene (*SbISWI*), which is homologous to rice *CHR728*. Sorghum genome contain single gene of CHD1, SWR1, and INO80 (*SbCHD1*, *SbSWR1,* and *SbINO80*), two of DDM1 (*SbDDM1a* and *SbDDM1b*) and three of CHD3 (*SbCHD3a/b/c*) resembling their counterparts in rice genome (Fig. 8). The homolog of sorghum SbCHD3c and rice CHR703 in Arabidopsis does not exist, which indicates that this gene is specific to monocots. However, the expression profile of this gene in rice and sorghum is slightly different. High levels of *CHR703* expression were observed in inflorescences and endosperm but *SbCHD3c* is only expressed in inflorescences (Fig. S8). *SbSNF2a/b/c*, *SbISWI*, *SbCHD1*, *SbCHD3a/b*, *SbSWR1,* and *SbINO80* are

expressed at high levels in all the tissues (Fig. S8), echoing that these remodeling factors exert their function through entire life cycle of sorghum. SbDDM1a/b, whose homologs in both rice and Arabidopsis have been revealed to be involved in the promotion of DNA methylation, is expressed in a similar pattern with *SbCMT3s* (Fig. S8), providing the clue for their involvement in DNA methylation in sorghum.

Co-expression analysis of sorghum chromatin regulator genes

Many chromatin regulators function in concert with each other. With the notion that spatial and temporal overlapping of gene expression determines the possibility of functional interaction, we performed co-expression analysis of chromatin regulators by clustering their expression pattern. The result indicates that the expression profiles of sorghum chromatin regulators could be roughly classified into four groups (Fig. S9). Group I contain most of the canonical histone genes that are cell-cycle regulated, suggesting that this group genes tend to be expressed in the tissues capable of cell proliferation. Most of them are poorly expressed in leaves, endosperm, pericarp and anther. Group II genes are constitutively expressed despite that some of them are not expressed in a few tissues especially pollens. Group III contains silent genes and the genes expressed in a few tissues. Group IV genes are actively transcribed mostly in meristems and/or inflorescences as well as a few other tissues.

Conclusions

Chromatin regulators play important roles in various developmental processes and in response to environmental change. Studies from Arabidopsis and rice demonstrate that the majority of chromatin regulators have conserved function in plant. However, specific regulators have been evolved in plant genome. Sorghum is an ideal plant for exploring the mechanism of C4 photosynthesis and plant resistance to stress. In this study, we identified the major chromatin regulators in sorghum. We found that sorghum genome contains most of but not all the plant chromatin regulators-encoding genes that have been identified in Arabidopsis and rice, resulting from differential duplication events of some genes in these species. However, sorghum and a few other grass species evolve some novel histone proteins. It is of great interest to understand whether these proteins could confer specific features to these species. More interestingly, a sorghum-unique protein sharing the domains of Snf2 family proteins, the elongation factor EF-Tu and the histone chaperone SPT16 was discovered. Finally, we categorized expression patterns of all the chromatin regulators into four groups to predict the possibility of their functional
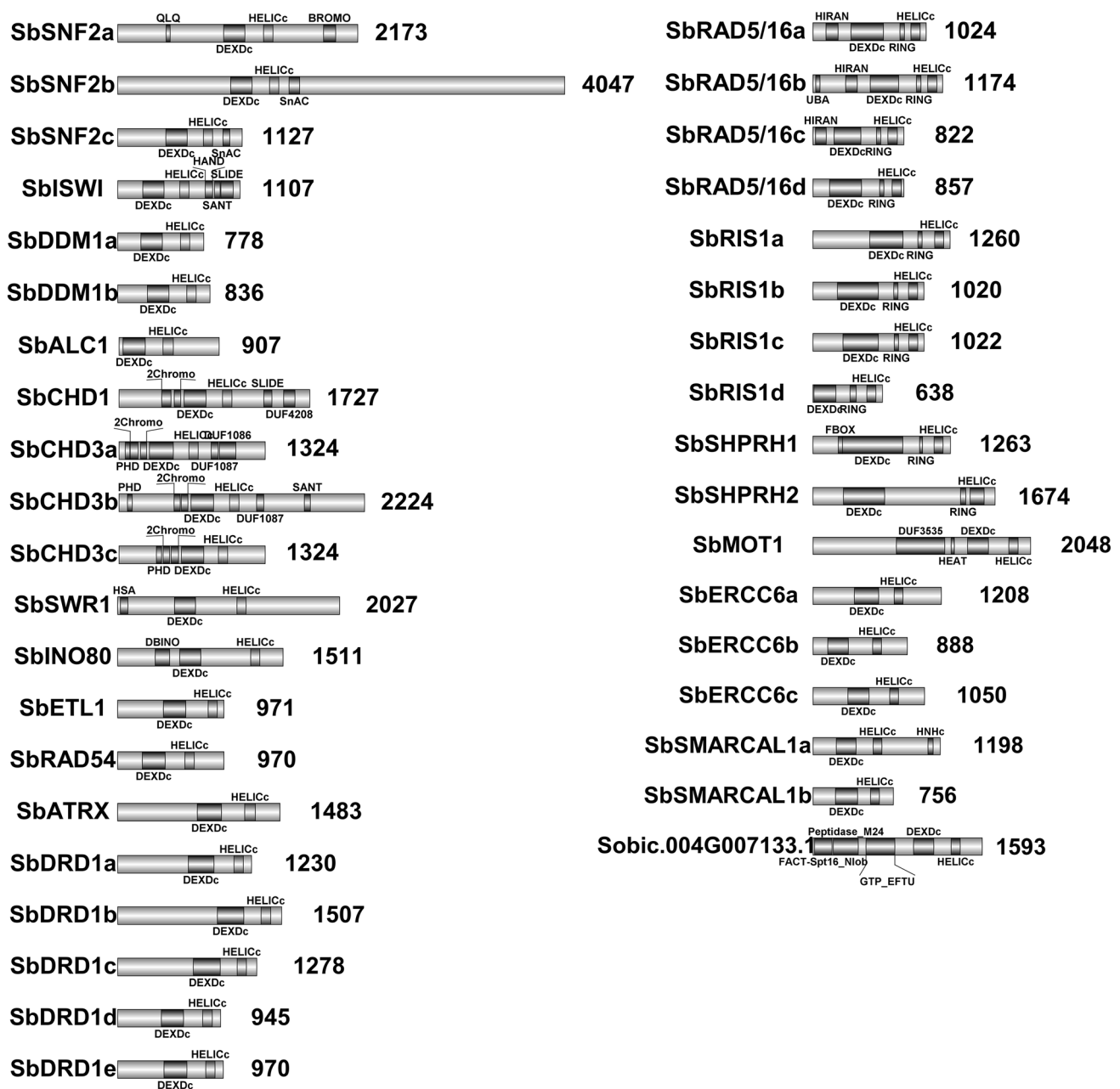
Fig. 9 Domain architecture of Snf2 family proteins in sorghum

interaction. Future works towards investigating how the chromatin regulators cooperatively participate in epigenetic regulation are necessary for fully understanding their roles in development and stress response in sorghum.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

Ahmad A, Dong Y, Cao X (2011) Characterization of the PRMT gene family in rice reveals conservation of arginine methylation. PLoS One 6(8):e22664. https://doi.org/10.1371/journal.pone.0022664

Audia JE, Campbell RM (2016) Histone modifications and cancer. Cold Spring Harb Perspect Biol 8(4):a019521. https://doi.org/10.1101/cshperspect.a019521

Bourque S, Jeandroz S, Grandperret V, Lehotai N, Aime S, Soltis DE, Miles NW, Melkonian M, Deyholos MK, Leebens-Mack JH, Chase MW, Rothfels CJ, Stevenson DW, Graham SW, Wang X, Wu S, Pires JC, Edger PP, Yan Z, Xie Y, Carpenter EJ, Wong GKS, Wendehenne D, Nicolas-Frances V (2016) The evolution of HD2 proteins in green plants. Trends Plant Sci 21(12):1008–1016. https://doi.org/10.1016/j.tplants.2016.10.001

Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu SH, Jiang N, Robin Buell C (2012) Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J 71(3):492–502. https://doi.org/10.1111/j.1365-313X.2012.05005.x

Du K, Luo Q, Yin L, Wu J, Liu Y, Gan J, Dong A, Shen WH (2020) OsChz1 acts as a histone chaperone in modulating chromatin organization and genome function in rice. Nat Commun 11(1):5717. https://doi.org/10.1038/s41467-020-19586-z

D'Urso A, Brickner JH (2014) Mechanisms of epigenetic memory. Trends Genet 30(6):230–236. https://doi.org/10.1016/j.tig.2014.04.004

Fyodorov DV, Zhou BR, Skoultchi AI, Bai Y (2018) Emerging roles of linker histones in regulating chromatin structure and function. Nat Rev Mol Cell Biol 19(3):192–206. https://doi.org/10.1038/nrm.2017.94

Gonzalez-Arzola K, Diaz-Quintana A, Rivero-Rodriguez F, Velazquez-Campoy A, De la Rosa MA, Diaz-Moreno I (2017) Histone chaperone activity of Arabidopsis thaliana NRP1 is blocked by cytochrome c. Nucleic Acids Res 45(4):2150–2165. https://doi.org/10.1093/nar/gkw1215

Hammond CM, Stromme CB, Huang H, Patel DJ, Groth A (2017) Histone chaperone networks shaping chromatin function. Nat Rev Mol Cell Biol 18(3):141–158. https://doi.org/10.1038/nrm.2016.159

Henikoff S, Comai L (1998) A DNA methyltransferase homolog with a chromodomain exists in multiple polymorphic forms in Arabidopsis. Genetics 149(1):307–318

Hu Y, Lai Y (2015) Identification and expression analysis of rice histone genes. Plant Physiol Biochem 86:55–65. https://doi.org/10.1016/j.plaphy.2014.11.012

Hu Y, Qin F, Huang L, Sun Q, Li C, Zhao Y, Zhou DX (2009) Rice histone deacetylase genes display specific expression patterns and developmental functions. Biochem Biophys Res Commun 388(2):266–271. https://doi.org/10.1016/j.bbrc.2009.07.162

Hu Y, Zhu N, Wang X, Yi Q, Zhu D, Lai Y, Zhao Y (2013) Analysis of rice Snf2 family proteins and their potential roles in epigenetic regulation. Plant Physiol Biochem 70:33–42. https://doi.org/10.1016/j.plaphy.2013.05.001

Jambhekar A, Dhall A, Shi Y (2019) Roles and regulation of histone methylation in animal development. Nat Rev Mol Cell Biol 20(10):625–641. https://doi.org/10.1038/s41580-019-0151-1

Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11(3):204–220. https://doi.org/10.1038/nrg2719

Lee KK, Workman JL (2007) Histone acetyltransferase complexes: one size doesn't fit all. Nat Rev Mol Cell Biol 8(4):284–295. https://doi.org/10.1038/nrm2145

Liu X, Luo M, Zhang W, Zhao J, Zhang J, Wu K, Tian L, Duan J (2012) Histone acetyltransferases in rice (*Oryza sativa* L.): phylogenetic analysis, subcellular localization and expression. BMC Plant Biol 12:145. https://doi.org/10.1186/1471-2229-12-145

Lu F, Li G, Cui X, Liu C, Wang XJ, Cao X (2008) Comparative analysis of JmjC domain-containing proteins reveals the potential histone demethylases in Arabidopsis and rice. J Integr Plant Biol 50(7):886–896. https://doi.org/10.1111/j.1744-7909.2008.00692.x

Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389(6648):251–260. https://doi.org/10.1038/38444

Makita Y, Shimada S, Kawashima M, Kondou-Kuriyama T, Toyoda T, Matsui M (2015) MOROKOSHI: transcriptome database in *Sorghum bicolor*. Plant Cell Physiol 56(1):e6. https://doi.org/10.1093/pcp/pcu187

Martignago D, Bernardini B, Policelli F, Salvi D, Cona A, Angelini R, Tavladoraki P (2019) The four FAD-dependent histone demethylases of Arabidopsis are differently involved in the control of flowering time. Front Plant Sci 10:669. https://doi.org/10.3389/fpls.2019.00669

Martire S, Banaszynski LA (2020) The roles of histone variants in fine-tuning chromatin organization and function. Nat Rev Mol Cell Biol 21(9):522–541. https://doi.org/10.1038/s41580-020-0262-8

Ng DW, Wang T, Chandrasekharan MB, Aramayo R, Kertbundit S, Hall TC (2007) Plant SET domain-containing proteins: structure, function and regulation. Biochim Biophys Acta 1769(5–6):316–329. https://doi.org/10.1016/j.bbaexp.2007.04.003

Pandey R, Muller A, Napoli CA, Selinger DA, Pikaard CS, Richards EJ, Bender J, Mount DW, Jorgensen RA (2002) Analysis of histone acetyltransferase and histone deacetylase families of Arabidopsis thaliana suggests functional diversification of chromatin modification among multicellular eukaryotes. Nucleic Acids Res 30(23):5036–5055. https://doi.org/10.1093/nar/gkf660

Probst AV, Desvoyes B, Gutierrez C (2020) Similar yet critically different: the distribution, dynamics and function of histone variants. J Exp Bot 71(17):5191–5204. https://doi.org/10.1093/jxb/eraa230

Qian S, Wang Y, Ma H, Zhang L (2015) Expansion and functional divergence of Jumonji C-containing histone demethylases: significance of duplications in ancestral angiosperms and vertebrates. Plant Physiol 168(4):1321–1337. https://doi.org/10.1104/pp.15.00520

Qin FJ, Sun QW, Huang LM, Chen XS, Zhou DX (2010) Rice SUVH histone methyltransferase genes display specific functions in chromatin modification and retrotransposon repression. Mol Plant 3(4):773–782. https://doi.org/10.1093/mp/ssq030

Raynaud C, Sozzani R, Glab N, Domenichini S, Perennes C, Cella R, Kondorosi E, Bergounioux C (2006) Two cell-cycle regulated SET-domain proteins interact with proliferating cell nuclear antigen (PCNA) in Arabidopsis. Plant J 47(3):395–407. https://doi.org/10.1111/j.1365-313X.2006.02799.x

Rutowicz K, Puzio M, Halibart-Puzio J, Lirski M, Kotlinski M, Kroten MA, Knizewski L, Lange B, Muszewska A, Sniegowska-Swierk K, Koscielniak J, Iwanicka-Nowicka R, Buza K, Janowiak F, Zmuda K, Joesaar I, Laskowska-Kaszub K, Fogtman A, Kollist H, Zielenkiewicz P, Tiuryn J, Siedlecki P, Swiezewski S, Ginalski K,

Koblowska M, Archacki R, Wilczynski B, Rapacz M, Jerzmanowski A (2015) A specialized histone H1 variant is required for adaptive responses to complex abiotic stress and related DNA methylation in Arabidopsis. Plant Physiol 169(3):2080–2101. https://doi.org/10.1104/pp.15.00493

Seto E, Yoshida M (2014) Erasers of histone acetylation: the histone deacetylase enzymes. Cold Spring Harb Perspect Biol 6(4):a018713. https://doi.org/10.1101/cshperspect.a018713

Sharma R, Mohan Singh RK, Malik G, Deveshwar P, Tyagi AK, Kapoor S, Kapoor M (2009) Rice cytosine DNA methyltransferases—gene expression profiling during reproductive development and abiotic stress. FEBS J 276(21):6301–6311. https://doi.org/10.1111/j.1742-4658.2009.07338.x

Song F, Chen P, Sun D, Wang M, Dong L, Liang D, Xu RM, Zhu P, Li G (2014) Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. Science 344(6182):376–380. https://doi.org/10.1126/science.1251413

Tripathi AK, Singh K, Pareek A, Singla-Pareek SL (2015) Histone chaperones in Arabidopsis and rice: genome-wide identification, phylogeny, architecture and transcriptional regulation. BMC Plant Biol 15:42. https://doi.org/10.1186/s12870-015-0414-8

Ueda M, Matsui A, Tanaka M, Nakamura T, Abe T, Sako K, Sasaki T, Kim JM, Ito A, Nishino N, Shimada H, Yoshida M, Seki M (2017) The distinct roles of class I and II RPD3-like histone deacetylases in salinity stress response. Plant Physiol 175(4):1760–1773. https://doi.org/10.1104/pp.17.01332

Wagner D (2003) Chromatin regulation of plant development. Curr Opin Plant Biol 6(1):20–28. https://doi.org/10.1016/s1369526602000079

Wang L, Qiao H (2020) Chromatin regulation in plant hormone and plant stress responses. Curr Opin Plant Biol 57:164–170. https://doi.org/10.1016/j.pbi.2020.08.007

Wang B, Regulski M, Tseng E, Olson A, Goodwin S, McCombie WR, Ware D (2018) A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. Genome Res 28(6):921–932. https://doi.org/10.1101/gr.227462.117

Yelagandula R, Stroud H, Holec S, Zhou K, Feng S, Zhong X, Muthurajan UM, Nie X, Kawashima T, Groth M, Luger K, Jacobsen SE, Berger F (2014) The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in Arabidopsis. Cell 158(1):98–109. https://doi.org/10.1016/j.cell.2014.06.006

Yu Z, Zhang G, Teixeira da Silva JA, Li M, Zhao C, He C, Si C, Zhang M, Duan J (2021) Genome-wide identification and analysis of DNA methyltransferase and demethylase gene families in Dendrobium officinale reveal their potential functions in polysaccharide accumulation. BMC Plant Biol 21(1):21. https://doi.org/10.1186/s12870-020-02811-8

Zemach A, Kim MY, Silva P, Rodrigues JA, Dotson B, Brooks MD, Zilberman D (2010) Local DNA hypomethylation activates genes in rice endosperm. Proc Natl Acad Sci USA 107(43):18729–18734. https://doi.org/10.1073/pnas.1009695107

Zhou CY, Johnson SL, Gamarra NI, Narlikar GJ (2016) Mechanisms of ATP-dependent chromatin remodeling motors. Annu Rev Biophys 45:153–181. https://doi.org/10.1146/annurev-biophys-051013-022819

Zhu Y, Weng M, Yang Y, Zhang C, Li Z, Shen WH, Dong A (2011) Arabidopsis homologues of the histone chaperone ASF1 are crucial for chromatin replication and cell proliferation in plant development. Plant J 66(3):443–455. https://doi.org/10.1111/j.1365-313X.2011.04504.x