the plant journal

SEB Society for Experimental Biology

The Plant Journal (2016) **87,** 568–582 doi: 10.1111/tpj.13219

The NnCenH3 protein and centromeric DNA sequence profiles of *Nelumbo nucifera* Gaertn. (sacred lotus) reveal the DNA structures and dynamics of centromeres in basal eudicots

Zhixuan Zhu, Songtao Gui, Jing Jin, Rong Yi, Zhihua Wu, Qian Qian and Yi Ding*

Department of Genetics, State Key Laboratory of Hybrid Rice, College of Life Sciences, Wuhan University, Wuhan 430072, China

Received 2 February 2016; revised 15 May 2016; accepted 23 May 2016; published online 2 August 2016.
*For correspondence (e-mail yiding@whu.edu.cn).

SUMMARY

Centromeres on eukaryotic chromosomes consist of large arrays of DNA repeats that undergo very rapid evolution. *Nelumbo nucifera* Gaertn. (sacred lotus) is a phylogenetic relict and an aquatic perennial basal eudicot. Studies concerning the centromeres of this basal eudicot species could provide ancient evolutionary perspectives. In this study, we characterized the centromeric marker protein NnCenH3 (sacred lotus centromere-specific histone H3 variant), and used a chromatin immunoprecipitation (ChIP)-based technique to recover the NnCenH3 nucleosome-associated sequences of sacred lotus. The properties of the centromere-binding protein and DNA sequences revealed notable divergence between sacred lotus and other flowering plants, including the following factors: (i) an NnCenH3 alternative splicing variant comprising only a partial centromere-targeting domain, (ii) active genes with low transcription levels in the NnCenH3 nucleosomal regions, and (iii) the prevalence of the Ty1/*copia* class of long terminal repeat (LTR) retrotransposons in the centromeres of sacred lotus chromosomes. In addition, the dynamic natures of the centromeric region showed that some of the centromeric repeat DNA sequences originated from telomeric repeats, and a pair of centromeres on the dicentric chromosome 1 was inactive in the metaphase cells of sacred lotus. Our characterization of the properties of centromeric DNA structure within the sacred lotus genome describes a centromeric profile in ancient basal eudicots and might provide evidence of the origins and evolution of centromeres. Furthermore, the identification of centromeric DNA sequences is of great significance for the assembly of the sacred lotus genome.

Keywords: sacred lotus (*Nelumbo nucifera*), centromere, NnCenH3, centromere-associated DNA sequences, Ty1/*copia*, basal eudicots.

INTRODUCTION

The centromeres of metaphase chromosomes are responsible for accurate chromosome segregation, and for chromosome stabilization during meiosis and mitosis in eukaryotes. The function and structure of centromeres have been gradually elucidated in recent years. A functional centromere marker, CenH3 (a centromere-specific histone H3 variant, CENP-A in human), consists of a highly variable N-terminal tail domain and a conserved histone fold domain (HFD) containing the CATD (CENP-A targeting domain) region (Black *et al.*, 2004). In epigenetic centromeres, the paradox between the conservation of centromere function and the high divergence of the CenH3 N-terminal tail and centromeric-associated DNA sequences has hindered the elucidation of the mechanisms of centromere formation, maintenance and inheritance (Henikoff *et al.*, 2001).

Recent studies have characterized the centromeric DNA of some species via chromatin immunoprecipitation sequencing (ChIP-seq) analyses using anti-CenH3 antibody (Yan *et al.*, 2008; Sullivan *et al.*, 2011; Gong *et al.*, 2012; Fu *et al.*, 2013; Lefrancois *et al.*, 2013). Two types of centromeric DNA have been proposed: repeat sequences in classical centromeres and unique sequences in neocentromeres. In classical centromeres, rapidly evolving centromeric repeats consist of centromeric satellite DNA and centromere-specific retrotransposons (CRs). The centromeric satellite DNA consists of large arrays of tandem repeats, with monomers ranging in size from one nucleosomal DNA to thousands of base pairs (Gong *et al.*, 2012; Melters *et al.*, 2013). This highly abundant accumulation may form a specific DNA conformation and facilitate

CenH3-nucleosome assembly and stabilization (Verdaasdonk and Bloom, 2011; Garrido-Ramos, 2015). Although the draft genomes of variant species have been completed in recent decades, fine-scale genetic and physical mapping of centromeres remains a challenge because centromeric DNA contains blocks of satellite repeats. CRs in the Ty3/*gypsy* class, which are intermingled with the centromeric satellite repeats, are alternative primary centromeric DNA constituents (Miller *et al.*, 1998; Presting *et al.*, 1998; Nagaki *et al.*, 2005; Liu *et al.*, 2008; Weber and Schmidt, 2009; Li *et al.*, 2012; Luo *et al.*, 2012). Transposons are significant in the stabilization of centromeric structures. For instance, the human centromere protein CENP-B is a homolog of the pogo and Tiggers transposases (Tudor *et al.*, 1992; Smit and Riggs, 1996), and CRM (CRs in maize) elements may be transcribed (Topp *et al.*, 2004). Although there is great variance of the centromeric DNA sequences within and between species, the divergence of centromeric repeats generated during evolution is critical for reproductive isolation and speciation (Malik and Henikoff, 2009). Repeatless-based centromeres are defined as neocentromeres, which may explain the origin of centromeres. Specifically, classical centromeres are thought to have evolved from neocentromeres in a process that involved the emergence and expansion of satellite repeats (Yan *et al.*, 2006; Gong *et al.*, 2012). Moreover, neocentromeres provide evidence that a specific centromeric epigenetic feature, namely centromeric DNA, is unnecessary for kinetochore assembly and centromeric function (Mehta *et al.*, 2010; Westhorpe and Straight, 2015).

In plants, the centromeres of monocots and eudicots have been extensively studied; however, little information regarding the centromeric sequence structure of ancient basal eudicots has been reported, and our knowledge of centromere evolution remains incomplete. Knowledge of the sequence structure of the basal eudicot centromere would provide a critical reference to expand our insight into centromere evolution. Sacred lotus (*Nelumbo nucifera* Gaertn.; $2n = 2x = 16$) is an aquatic perennial basal eudicot that has survived since the Late Cretaceous period, and is important in evolution because of its ancient status in taxonomy. Recent phylogenetic analyses of nuclear and chloroplast genomic sequences and several genes have documented the ancient evolutionary characteristics of *Nelumbo nucifera* (sacred lotus; Ming *et al.*, 2013; Wang *et al.*, 2013; Wu *et al.*, 2014a,b). As a relict species, sacred lotus could be an appropriate model to elucidate the evolutionary similarities and differences of centromeres between basal eudicots and other flowering plants.

Herein, we characterized two isoforms of NnCenH3 that can serve as marks to identify centromeric sequences in sacred lotus. The NnCenH3-associated nucleosomes of sacred lotus were sequenced using ChIP-seq, and the centromere-associated genes, retrotransposons and chromosome distribution of centromere-associated repeats were recovered. These studies describe the particular centromeric profile of this ancient species and provide insights into the composition of centromeric sequences in basal eudicots. The properties of the sacred lotus centromeres display distinct differences from those of other flowering plants and could provide the significant clues to the mechanism of centromere evolution. In addition, with the launch of the *N. nucifera* genome project (Ming *et al.*, 2013; Wang *et al.*, 2013), the systematic study of centromeres will contribute to the integrity of genetic and physical maps of the sacred lotus genome.

RESULTS

Characterization of two NnCenH3 isoforms

We used degenerate primers to amplify homologous sequences of the *CenH3* gene in the sacred lotus genome. Because the N-terminal regions (NRs) of CenH3 are highly variable in eukaryotes, degenerate primers were designed based on the conserved αN-helix and α2-helix domains of the C-terminal region (CR), which was identified based on multiple alignments in *Arabidopsis thaliana* (GenBank accession number Q8RVQ9; Talbert *et al.*, 2002), *Brassica carinata* (GenBank accession number ACZ04983; Wang *et al.*, 2011), *Brassica napus* (GenBank accession number ACZ04984; Wang *et al.*, 2011) and *Brassica rapa* (GenBank accession number ADN92693; Wang *et al.*, 2011) (Figure 1a). We then performed 3′ and 5′ rapid amplification of cDNA ends (RACE) and assembled the amplified overlapping fragments (Figure S1). Subsequently, a 474-bp coding sequence (CDS) of the *CenH3* gene was identified in the sacred lotus genome, and a deduced protein containing 158 amino acids was named NnCenH3-A (*N. nucifera* centromere-specific histone H3 isoform A). Moreover, we obtained a putative histone 3.3 protein, which we designated NnH3.3, and a 132-amino acid protein encoded by a putative splicing variant of NnCenH3 mRNA. According to an analysis of amino acid sequence alignment, NnH3.3 exhibited 48% homology with the predicted protein sequence of NnCenH3-A: it has a shorter N-terminal region and a missing amino acid residue in the loop-1 region (Figure 1a). The splicing variant named NnCenH3-B, which lacked exon 5, and encoded an α1-helix, loop 1 and a small portion of the α2-helix domain, had not been identified in plants until now. Similar deletions were discovered in *Homo sapiens* (GenBank accession number AAH00881.1; Figure 1b).

Because sacred lotus is a basal eudicot, we aligned NnCenH3-A with putative NnH3.3 and two known CenH3 coding sequences obtained from *Oryza sativa* subsp. *japonica* cv. Nipponbare (rice; GenBank accession number AY438639.1) and *A. thaliana* (GenBank accession number Q8RVQ9). The result shows that all CenH3 protein

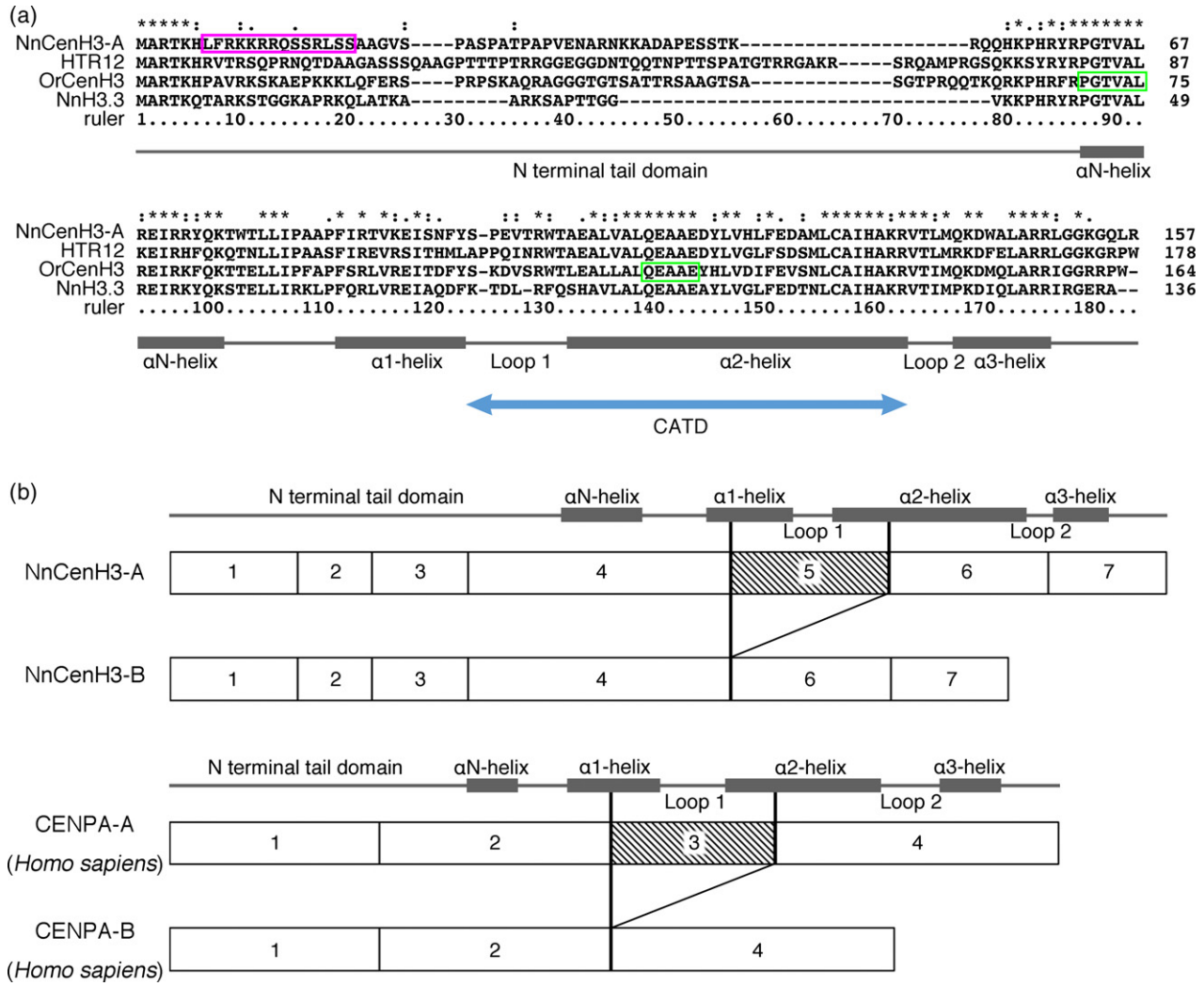

Figure 1. Alignment of CenH3 homologs. (a) Multiple sequence alignment of NnCenH3-A (*Nelumbo nucifera*), HTR12 (*Arabidopsis thaliana*), OrCenH3 (*Oryza sativa*) and NnH3.3 (*Nelumbo nucifera* Histone 3.3). The two green boxes indicate the peptides used to design degenerate primers. Functional domains are shown below the sequences. The blue arrow indicates the CATD-containing loop 1 and α2-helix. The peptide used to produce the NnCenH3 antibody is shown in the magenta box. (b) Diagram of the NnCenH3 alternative splicing variant structure in sacred lotus (*Nelumbo nucifera*) and human (*Homo sapiens*). The numbered boxes represent the exons of *CenH3* homologs. Functional domains are shown above the exons. The diagonal boxes indicate exons that are absent in the *NnCenH3-B* or *CENPA-B* transcript.

sequences contain a conserved C-terminal sequence in the last 96 amino acids, whereas the N-terminal regions of the first 61 amino acids show sequence variations (Figure 1a). Collectively, these results indicate that the NnCenH3 protein shares sequence and structural similarities with other CenH3 proteins in angiosperms.

NnCenH3 expression and specific location on sacred lotus centromeres

To study the correlation between NnCenH3 and sacred lotus centromeres, an anti-NnCenH3 polyclonal antibody was raised against the peptide LFRKKRRQSSRLSS (N-terminal amino acid residues 7–20; Figure 1a). Subsequently, we expressed the 25-kDa NnCenH3 fusion protein in *Escherichia coli* (Figure S2a); the anti-NnCenH3 antibody recognized the 25-kDa major protein band, indicating the high specificity and affinity of this antibody (Figure S2b). In addition, western blot analysis of NnCenH3 from roots, seedlings and leaves did not indicate different levels in these tissues (Figure S2c). Although the anti-NnCenH3 antibody may cross-react with NnCenH-B in the tissues of sacred lotus, no specific cross-reacting bands that migrated more rapidly than the NnCenH3 protein were observed by western blotting, suggesting that the NnCenH-B protein is present at very low levels in these tissues.

To investigate the distribution of the NnCenH3 protein on chromosomes, the root tip of sacred lotus was

immunostained with anti-NnCenH3 antibody. The results revealed strong signals at the centromeric regions of all 16 chromosomes at mitosis (Figure 2a). The NnCenH3 signals were clearly located in the heterochromatin region during the interphase, and showed a non-*Rab*l distribution pattern (Figure 2b) defined by the presence of scattered centromeres throughout the nucleus. The non-*Rab*l pattern of immunostaining of the sacred lotus nucleus is consistent with previous reports in other species (Billia and de Boni, 1991; Zalensky *et al.*, 1995; Dong and Jiang, 1998). Overall, these findings suggest that the deduced NnCenH3 is an authentic centromere-specific protein that exists in a variety of cells and tissues in the sacred lotus.

Distribution of NnCenH3 binding sequences in the sacred lotus genome

To isolate the NnCenH3-binding DNA sequences, ChIP was performed using the anti-NnCenH3 antibody. To test whether the DNA sequences detected by ChIP were located at the centromeres, we performed fluorescence *in situ* hybridization (FISH) using DNA probes obtained from ChIP-ed DNA (Figure 2c). The results indicated clear and specific signals at centromeres on the sacred lotus chromosomes as well as a pair of weak signals on the long arm of chromosome 1 (Chr. 1; Figure 2c); however, anti-NnCenH3 antibody immunofluorescence signals were not detected on the long arm of Chr. 1 (Figure 2a), suggesting

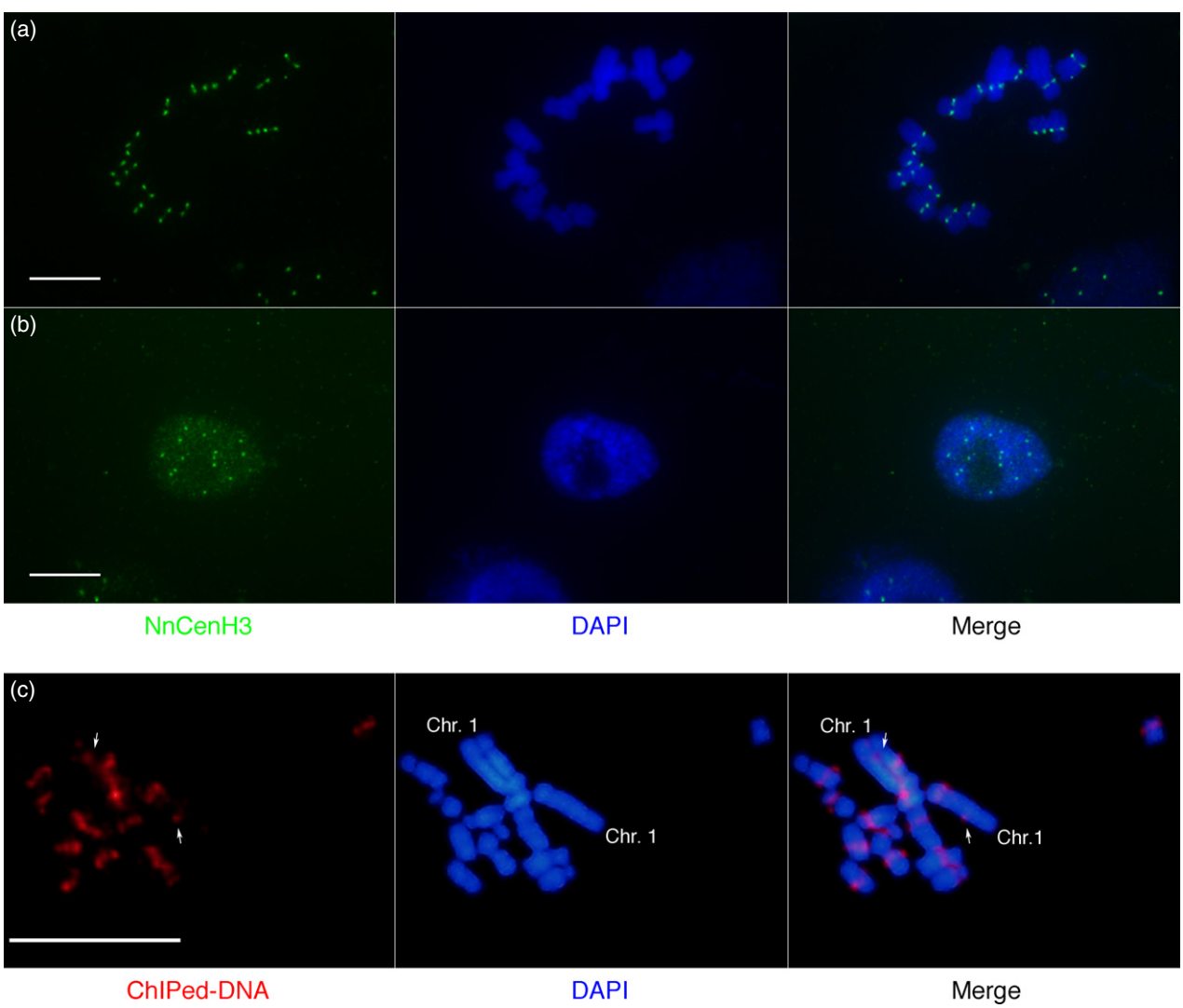

Figure 2. Localization of NnCenH3 protein and ChIP-ed DNA on sacred lotus chromosomes. (a, b) Immunostaining with anti-NnCenH3 antibody (green) on chromosomes (a) and interphase nuclei (b) of sacred lotus. (c) FISH signals (red) for the DNA precipitated by ChIP with the anti-NnCenH3 antibody. The DNA was counterstained with 4′,6-diamidino-2-phenylindole (DAPI; blue). The white arrows in (c) indicate non-centromeric signals on the long arm of chromosome 1 (Chr. 1) of sacred lotus cells. Scale bars: 10 μm.

that some non-centromeric sequences might have similarities to NnCenH3-associated centromeric DNA.

To describe the global NnCenH3-binding profiles, the ChIP-ed DNA was sequenced on the Illumina HiSeq 2500 platform and aligned to the reference sacred lotus genome (Ming *et al.*, 2013) to generate 33.8 million 49-bp reads, including 53.01% (17.9 million) of uniquely mapping reads (Figure S3). We identified 10 967 peaks distributed in 1012 scaffolds. Fifty-nine scaffolds (Figure S4), of which ≥5-kb mapped scaffolds ranging from 6.048 kb (scaffold 1962) to 11.3 Mb (megascaffold 19) were defined using high ChIP-seq read peaks (Figures 3 and S4), showed conspicuous sequence enrichment. These scaffolds were designated 'NnCenH3-binding scaffolds' (NnCBSs) and used for further analysis. The remaining <5-kb mapped scaffolds were arbitrarily considered technical background noise. Generally, CenH3-binding subdomains were separated by histone H3 nucleosomes rather than completely spanning the entire centromeric region (Chueh *et al.*, 2005; Gong *et al.*, 2012; Wang *et al.*, 2014). We found 23 NnCBSs (megascaffolds 19, 46, 47, 48, 84, 88, 89, 129, 133 and 142, and scaffolds 293, 377, 392, 446, 484, 551, 552, 562, 563, 566, 620, 703 and 722) that presented the same interrupted pattern, including at least two peak enrichment blocks (Figure S4). This finding suggested that the distribution of NnCenH3 nucleosomes in the sacred lotus resembles the core centromeres of angiosperms (Yan *et al.*, 2008; Gong *et al.*, 2012) in its interrupted pattern. In megascaffold 19, the discontinuous NnCenH3 subdomains scattered in the 3′ terminal of the scaffold encompassing 0.3 Mb (11.0–11.3 Mb) of DNA, and the remaining peak that lacked the 5′ terminal region comprised the largest NnCBS, being 11.34 Mb in size (Figure 3, megascaffold 19). Although the sizes of centromeres ranged widely from kilobases to megabases (Pluta *et al.*, 1995; Jin *et al.*, 2004; Alkan *et al.*, 2011), core centromeres exceeding 10 Mb have not been found in the monocentric chromosomes of plants or animals.

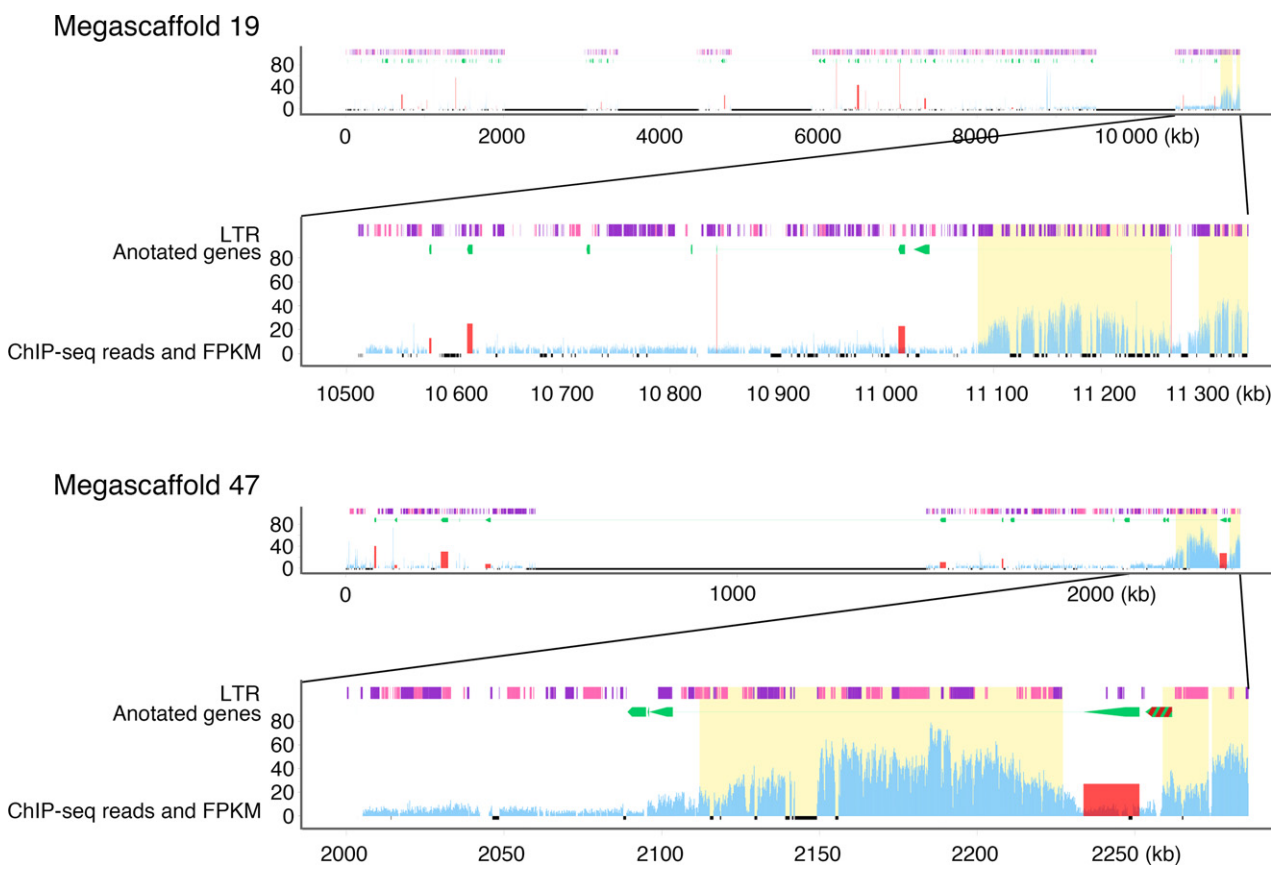

Figure 3. Distribution diagrams of ChIP-ed reads, genes and retrotransposons in megascaffold 19 and megascaffold 47. The blue peaks on the bottom track represent the ChIP-seq reads in 200-bp windows. The red rectangles show the transcription (FPKM) of annotated genes in these scaffolds. The number of ChIP-seq reads and FPKM values are shown on the *y*-axis. The black bars represent gaps. The gaps that are located adjacent to the accumulated peaks are attributed to the NnCenH3 subdomains (yellow boxes). The green arrows in the middle track represent annotated genes, which were less distributed within the NnCenH3 subdomains in megascaffold 19 and megascaffold 47. On the top track, Ty3/*gypsy* (pink rectangles) and Ty1/*copia* (purple rectangles) LTR retrotransposons were concentrated in NnCenH3 subdomains. The enlarged diagrams represent the fine structure of NnCBS. The annotated genes are mapped to the H3 subdomain. One cross-regional gene (gene *NNU_08544* represented by the green and red arrow) in megascaffold 47 is an active gene (Figure 4) with a low FPKM value.

Genes in the NnCenH3 nucleosomes show low density and activity

Although several studies have shown that centromeric heterochromatin is incompatible with gene expression (Schmid *et al.*, 2005; Lomiento *et al.*, 2008), the existence of active genes at the CenH3 subdomain in certain plant species has been proposed (Yan *et al.*, 2008; Gong *et al.*, 2012; Li *et al.*, 2013). Based on the lotus genome annotation database from LOTUS-DB (Wang *et al.*, 2015), 210 annotated non-transposable element (non-TE) genes were discovered in NnCBSs (Figure S4; Table S1) in the present study. Of these genes, 92% (194 of 210) were located in the NnH3 (Histone H3 protein in sacred lotus) subdomains, and 7% (15 of 210) of the NnCBS-associated genes were in the NnCenH3 subdomains; gene *NNU_08544* (indicated by green and red arrows in Figure 3a, megascaffold 47) crossed both subdomain regions in megascaffold 47. We found that the average gene density of the sacred lotus centromere was 64 kb per gene, lower than the density of 30 kb per gene in the sacred lotus genome reported by Ming *et al.* (2013). At the subdomain level, the NnCenH3-binding regions dramatically reduced the density of the sequence (156 kb per gene) compared with NnCenH3-lacking subdomains (56 kb per gene). This gene-sparse condition of the centromere appears to be similar in monocots (Yan *et al.*, 2006) and eudicots (Gong *et al.*, 2012).

We used RNA-seq data (Ming *et al.*, 2013) obtained from the leaf blades, petioles, rhizome internodes and roots to determine whether centromeric genes were transcriptionally competent in sacred lotus (Table S1). The results indicated that a total of 55% (116 of 210) of the putative genes in NnCBSs showed low levels of RNA-seq reads [0 < fragments per kilobase of exon model per million mapped fragments (FPKM) ≤ 15] in four tissues, and that 14% (30 of 210) of the genes (*NNU_01260*, *NNU_10043*, *NNU_15385*, *NNU_03501* and *NNU_15700*) were not expressed (FPKM = 0; Table S1). The remaining NnCBS-associated genes (64 of 210) displayed relatively high expression levels (FPKM > 15) in sacred lotus. Although 33 NnCBS-associated genes in the NnH3 subdomain regions were expressed at relatively high expression levels (FPKM > 15) in leaf tissues, only two NnCBS-associated genes (*NNU_02586* in megascaffold 88 and *NNU_00143* in scaffold 563; Table S1) were associated with NnCenH3 subdomains. The actual expression of all 16 NnCenH3-associated genes, including the subdomain-crossing gene *NNU_08544*, in leaf tissue was validated using RT-PCR (reverse transcription PCR). Four of these genes (*NNU_16348*, *NNU_02585*, *NNU_07763* and *NNU_19813*; Figure S5) were found to share significant sequence similarities with several other genes (*XM_010259541.1*, *XM_010257116.1*, *XM_010249303.1*, *XM_010248825.1* and *XM_010245592.1*) in non-centromeric scaffolds. We designed sequence-specific primers (Table S2)

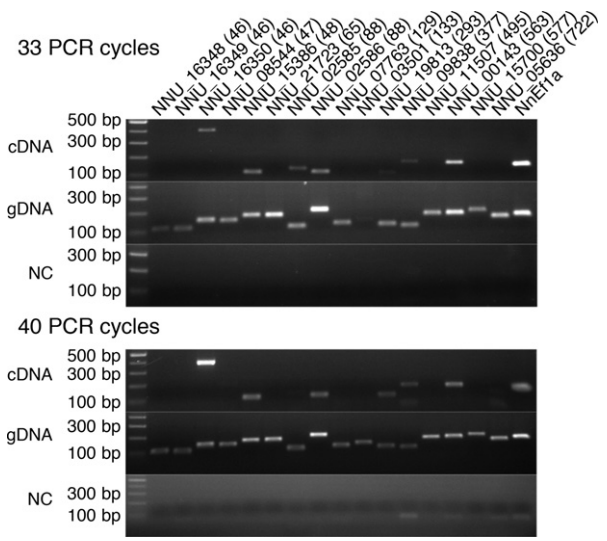

Figure 4. Transcription of NnCenH3 subdomain-associated genes in leaves of *Nelumbo nucifera* (sacred lotus), verified by RT-PCR. The numbers in brackets indicate NnCBS. cDNA was reverse-transcribed from leaf mRNA of sacred lotus. Genomic DNA (gDNA) was extracted from the leaves of sacred lotus. Double-distilled water was used as the negative control (NC), and the *NnEf1a* gene was used as the positive control. No products were obtained using primers *NNU_16348*, *NNU_16349*, *NNU_21723*, *NNU_07763*, *NNU_03501*, *NNU_11507* and *NNU_15700*.

and sequenced the products to differentiate these paralogous genes from the best-matched paralogous gene pairs *NNU_19813* (scaffold 293) and *XM_010245592.1* because only seven separated mispairings remained after alignment (Figure S5). RT-PCR analysis of the NnCenH3-associated genes showed low signals in seven genes (*NNU_16350*, *NNU_15386*, *NNU_02585*, *NNU_02586*, *NNU_19813*, *NNU_09838*, *NNU_00143*) compared with the reference gene *NnEf1a* (GenBank accession number: AB491177) after 33 cycles (Figure 4). We also failed to amplify the other seven genes (*NNU_16348*, *NNU_16349*, *NNU_21723*, *NNU_07763*, *NNU_03501*, *NNU_11507* and *NNU_15700*) after 40 cycles, indicating that these genes are not expressed in sacred lotus leaves (Figure 4). As further confirmed by RT-PCR, the gene *NNU_02585* in megascaffold 88 was expressed despite the fact that its FPKM value in the RNA-seq data from leaf blade tissue was 0. The RT-PCR products were sequenced to verify the genes and assure the specificity of the determination. The sequencing results showed high-purity *NNU_02585* fragments and the complete failure of amplification of the homologous gene *XM_010249303.1,* providing that *NNU_02585* transcripts are actually present in the sacred lotus leaf tissues. Thus the paradox of the detected expression of gene *NNU_02585*, which showed an FPKM value of 0, could result from the presence of a very low concentration of its gene fragments during the RNA library preparation, or from RNA-seq read-mapping errors arising from the high

similarity between *NNU_02585* and the homologous gene *XM_010249303.1*. We used ChIP and quantitative PCR (ChIP-qPCR) to further confirm that these 16 genes were highly enriched in ChIP-ed DNA (Figure S6). Our results confirmed that the gene sequences were present in the NnCenH3 subdomains. In conclusion, gene density and activity decreased in the centromeric regions, and the NnCenH3 subdomains contained fewer active genes than the NnH3 subdomains.

Ty1/*copia* class TEs are prevalent within centromeres

Most eukaryotic centromeres, including core centromeres and pericentromeres, contain blocks of repeated DNA sequences that consist of megabase-scale arrays of satellite DNA and TEs. Both of these elements play important roles in the structure, maintenance, function and evolution of centromeres. To investigate the structure and repetitive DNA distribution of the sacred lotus centromere, we completed a census of repeat sequences in NnCBSs using RepeatMasker. A total of 4 737 179 bp of repetitive DNA sequences was found in NnCBSs. Approximately 79.48% of the 13.3-Mb NnCBS sequence exhibited a higher content of repetitive DNA in centromeres than was found in the sacred lotus genome (38.05%). Retrotransposons, which represented 70.23% of the repetitive sequences, were predominant in the NnCBSs of the sacred lotus genome. The fraction of retrotransposons exceeded that of DNA transposons, which accounted for less than 4.47% of the repetitive sequences (Table 1) in the sacred lotus genome. A similar repeat situation has been found in most plants (Wu *et al.*, 2004; Wolfgruber *et al.*, 2009; Li *et al.*, 2013). Specifically, centromere-specific retrotransposons belonging to the Ty3/*gypsy* class of long terminal repeat (LTR) retrotransposons, which are common major centromeric transposon constituents, have been identified in plants (Neumann *et al.*, 2011). Surprisingly, we observed that the Ty1/*copia* element was the most abundant group of repeats in sacred lotus centromeres: it constituted 29.04% of the centromeric DNA, followed by Ty3/*gypsy* (17%), RC/Helitron (1.65%) and L1 (1.56%) (Table 1).

To compare the centromeric repetitive DNA of sacred lotus, eudicots and monocots, based on uniquely mapping

Table 1 Proportion of repetitive sequences in centromeres of *Nelumbo nucifera* (sacred lotus), *Solanum tuberosum* (potato) and *Oryza sativa* (rice) chromosome 8

	Sacred lotus			Potato			Rice Cen8		
Repetitive DNA	Total (bp)	% of repeats[a]	% on Cens[b]	Total (bp)	% of repeats[a]	% on Cens[b]	Total (bp)	% of repeats[a]	% on Cens[b]
Mobile element	7 448 198	70.23	55.82	4 415 935	96.66	53.81	1 355 279	99.12	68.71
Retroelement	6 851 083	64.60	51.34	4 109 304	89.95	50.08	1 242 941	90.91	63.01
LTR Retroelement	6 635 007	62.56	49.72	3 944 657	86.34	48.07	1 228 544	89.85	62.28
Caulimovirus	7200	0.07	0.05	–	–	–	–	–	–
Ty1/*copia*	3 874 896	36.54	29.04	316 065	6.92	3.85	42 669	3.12	2.16
Ty3/*gypsy*	2 268 103	21.39	17.00	3 626 982	79.39	44.2	973 242	71.18	49.34
Other LTR	484 808	4.57	3.63	1610	0.04	0.02	212 633	15.55	10.78
Other retrotransposon	216 076	2.04	1.62	164 647	3.6	2.01	22 302	1.63	1.13
Non-LTR	–	–	–	33 170	0.73	0.4	3678	0.27	0.19
LINE/L1	208 533	1.97	1.56	105 920	2.32	1.29	9890	0.72	0.5
LINE/RTE	7543	0.07	0.06	20 805	0.46	0.25	–	–	–
SINE	–	–	–	4752	0.1	0.06	829	0.06	0.04
Centromere-specific retrotransposons	–	–	–	–	–	–	7905	0.58	0.4
DNA transposon	597 115	5.63	4.47	306 631	6.71	3.74	104 433	7.64	5.29
En-Spm	20 364	0.19	0.15	125 273	2.74	1.53	31 363	2.29	1.59
hAT	135 795	1.28	1.02	36 303	0.79	0.44	11 611	0.85	0.59
Mariner	–	–	–	6938	0.15	0.08	127	0.01	0.01
MULE	111 355	1.05	0.83	55 201	1.21	0.67	6143	0.45	0.31
Harbinger	15 104	0.14	0.11	18 339	0.4	0.22	12 228	0.89	0.62
Mite	91 387	0.86	0.68	–	–	–	–	–	–
Stowaway	–	–	–	681	0.01	0.01	255	0.02	0.01
RC/Helitron	220 462	2.08	1.65	6053	0.13	0.07	4766	0.35	0.24
Other DNA transposon	2648	0.02	0.02	57 843	1.27	0.7	37 940	2.77	1.92
Simple sequence repeat	92 750	0.87	0.70	68 465	1.5	0.83	9957	0.73	0.5
Other repeats	3 064 322	28.89	22.97	84 124	1.84	1.03	2047	0.15	0.1
Total	10 605 270			4 568 524			1 367 283		

[a]The proportions of various repeat lengths in total masked sequences.
[b]The proportions of various repeats lengths in centromeric sequences.

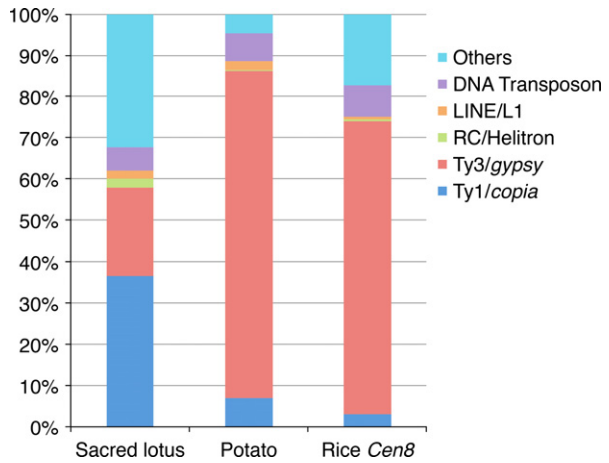

Figure 5. Proportions of repetitive sequences in the centromeres of *Nelumbo nucifera* (sacred lotus), *Solanum tuberosum* (potato) and *Oryza sativa* (rice) chromosome 8.

reads, we selected five neocentromere-associated sequences from *Solanum tuberosum* (potato) centromere DNA (*Cen4*, *6*, *9*, *11* and *12*), a eudicot neocentromere representative that has been ChIP-sequenced using anti-CenH3 antibody, and a rice chromosome 8 centromere (*Cen8*), a monocot neocentromere representative that has been fully sequenced. The ratio of Ty3/*gypsy* to Ty1/*copia* is 0.585:1 in sacred lotus NnCBS, whereas the Ty3/*gypsy* to Ty1/*copia* ratios in potato centromeres and rice *Cen8* are 11.475:1 and 22.809:1, respectively. This difference indicates that the Ty1/*gypsy* class became the predominant TE in the potato and rice genomes (Figure 5). In addition, the ratio of the DNA transposon En-spm to hAT in sacred lotus NnCBS (0.15:1) also differed from those of potato centromeres (3.451:1) and rice *Cen8* (2.701:1) (Table 1). The prevalence of Ty1/*copia* in the sacred lotus centromeres and the ratio of En-spm to hAT DNA transposons indicates the presence of significant discrepancies between the transposon constituents of sacred lotus and the centromere-associated transposons in other species.

Diverse mapping pattern of the centromere-associated repeats on chromosomes

Previous analysis of the sacred lotus genome database (Ming *et al.*, 2013) showed that numerous gaps exist in the sacred lotus draft genome because repeat arrays, especially in the repeat-rich heterocentromeric region, have hampered sequence assembly. It is possible that centromeric repeats were represented by the discarded non-uniquely mapped reads (11.5 million ChIP-seq reads) and the unmapped reads (4.3 million ChIP-seq reads), and that RepeatMasker may have underestimated the repetitive sequence frequency in sacred lotus NnCBSs in this study. To resolve this issue, the centromeric repeats were identified using the method described by Gong *et al.* (2012), which clusters the repeats into putative families by mapping ChIP-seq reads to similarity-based clusters.

We reconstructed the sacred lotus genomic repeats on a randomly sampled proportion (700 000 reads) of Illumina Hiseq2000 re-sequencing reads (GenBank accession number PRJNA269574), corresponding to a genome coverage of approximately 8%. Subsequent hierarchical agglomeration clustering resulted in 31 002 clusters totaling 300 856 reads. The results showed that the repeats occupy approximately 43% (300 856 of 700 000) of the sacred lotus genome. Larger clusters (from CL1 to CL294) with at least 0.001% of the sequence reads were representative of 28.9% of the genome (Table S3). The largest cluster (CL1) in the sacred lotus genome, with approximately 51 000 copies (4091 copies in the set of 700 000 reads data), yielded 74.5% hits in the Ty1/*copia* group (Table S3). The genome showed the greatest prevalence of the Ty1/*copia* LTR element. Similar results were obtained by aligning the sequences with the sacred lotus repeat database using the assembled scaffolds (Figure 5; Table 1).

The NnCenH3-associated ChIP-seq reads were then mapped to the large clusters (genome proportion > 0.01%), and the putative NnCenH3-associated repeats were estimated based on the ratio of mapped ChIP-seq reads to repeats of Illumina Hiseq2000 reads (Figure S7). The results showed that 23 clusters represented a greater than fourfold enrichment. These clusters were used as FISH probes in order to verify that they were NnCenH3-associated centromeric repeats. Although CL249 showed greater enrichment (ratio = 40.6) than any other cluster (Figure S7), no FISH signals were detected in the chromosomes as a result of its dispersed distribution and low copy number (1300 copies) in the sacred lotus genome. The FISH results showed that seven of the 23 ChIP-enriched cluster probes (CL6, CL8, CL17, CL19, CL21, CL24 and CL46; Table 2) generated unambiguous signals located

Table 2 Statistical analysis of centromere-associated repeat clusters

Repeat	RepeatMasker[a]	Re-seq reads (%)[b]	ChIP-seq reads (%) [c]	Ratio (ChIP-seq/ Re-seq)
CL6	LTR/*copia* (94.8%)	0.422	2.152	5.100
CL8	LTR/*copia* (87.2%)	0.379	1.701	4.489
CL17	LTR/*copia* (95.6%)	0.317	1.274	4.018
CL19	LTR/*copia* (91.7%)	0.300	1.901	6.337
CL21	LTR/*copia* (96.7%)	0.292	1.288	4.411
CL24	LTR/*copia* (96%)	0.282	1.284	4.552
CL46	LTR/*copia* (57.4%)	0.197	0.823	4.180

[a]Centromere-associated clusters were aligned to the sacred lotus repeat database by RepeatMasker.
[b]Proportion of re-sequencing genome reads.
[c]Proportion of ChIP-sequencing reads.

at the positions of the centromeres of the majority of sacred lotus chromosomes (Figure 6). Several chromosomes exhibited signals from the CL19 and CL21 probes both at the centromere region and at the distal ends of the chromosomes (Figure 6, CL19 and 21, inset). The CL8 repeats mapped exclusively to the centromere region, but Chr. 1 showed scattered, faint signals that were not specific to centromeres (Figure 6, CL8). This vague configuration was observed universally for other probes (Figure 6, CL17, 19, 21, 24 and 46), indicating that fewer types of centromeric repetitive clusters are presented in the centromere of Chr. 1 of the sacred lotus genome than in the centromeres of the other chromosomes. Interestingly, the CL6 repeats yielded a strong and concentrated signal on the long arm of Chr. 1 (Figure 6, CL6), and these non-centromeric signals co-localized with the NnCenH3-associated ChIP-ed DNA probes (Figure S8), further indicating that homologous centromeric repeats are found at non-centromeric positions. In addition, the CL24 and CL46 probes showed relatively weak signals at the centromeres of all chromosomes except Chr. 1 (Figure 6, CL24 and 46). The similarity-based clustering results indicated that the centromere-associated clusters formed linearly organized parts of the graph layout. But CL46 contributed to the circular layout of the graph, a property of tandem satellite repeats (Figure S9). We also used RepeatMasker to compare the sacred lotus repeat database and the reads of centromere-associated clusters, and this comparison revealed high similarity with the Ty1/*copia* group (Table 2).

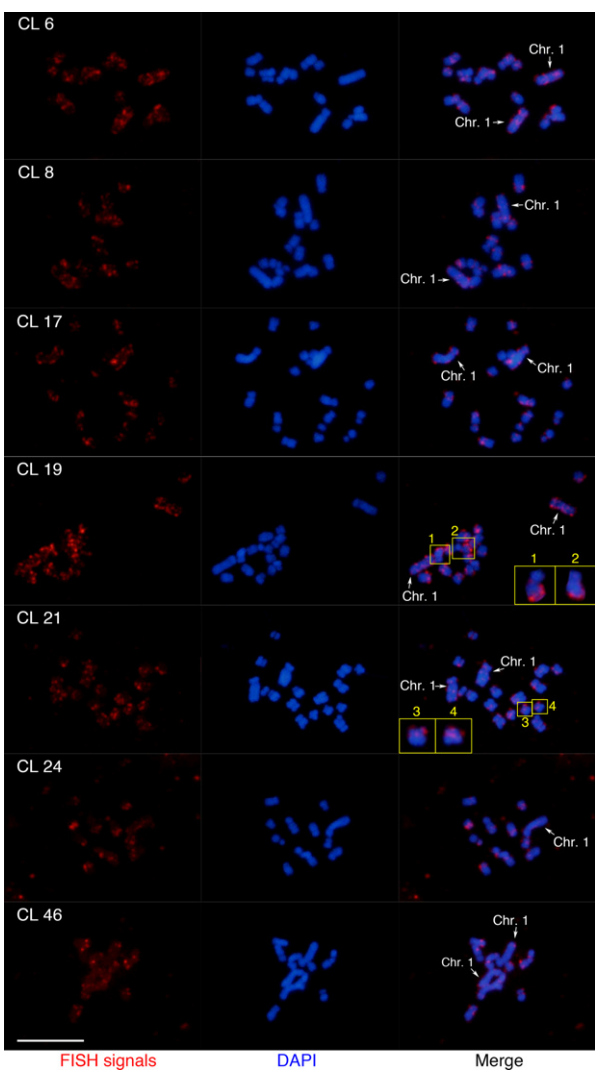

Figure 6. Distribution of centromere-associated clusters on *Nelumbo nucifera* (sacred lotus) chromosomes. The red signals show the (CL6, 8, 17,19, 21, 24 and 46) biotin-labeled centromere-associated repeat cluster probes (red) hybridized to mitotic metaphase chromosomes (blue) of sacred lotus. The DNA was counterstained with 4′,6-diamidino-2-phenylindole (DAPI; blue). The yellow boxes with numbers indicate signals at the terminal portion of sacred lotus chromosomes. The white arrows and the label Chr. 1 indicate chromosome 1 in sacred lotus cells. Scale bar: 10 μm.

DISCUSSION

NnCenH3-B displays a functional defect affecting centromere recognition

Recently, significant progress has been made in identifying the structure of the CenH3 protein. The N- and C-terminal tails and the CATD region comprising the loop-1 linker and the α2-helix of CenH3 have been shown to be critical for the establishment and maintenance of the centromere (Fachinetti *et al.*, 2013; Logsdon *et al.*, 2015; Westhorpe and Straight, 2015). Several groups have demonstrated that CATD, the core domain of CenH3, participates in CenH3 nucleosome assembly mediated by HJURP (Holliday Junction-Recognizing Protein), and is responsible for centromere targeting (Black *et al.*, 2007; Foltz *et al.*, 2009). These significant functions have been proven via lethal mutations in the loop 1 of Cid (a CenH3 homolog in *Drosophila*; Vermaak *et al.*, 2002), a change that is comparable with the loss of CATD in *A. thaliana* (Ravi *et al.*, 2010). More recently, the *A. thaliana* lines with some missense mutations in the CENH3 CATD region were shown to be sterile or to exhibit uniparental inheritance when crossed with wild-type lines, hereditary stability in the CENH3 mutant lines in self-crosses notwithstanding (Kuppu *et al.*, 2015). In the present study, the characterization of NnCenH3 indicated that either its sequential or its functional features are associated with centromeres (Figures 1, 2a and 2b). Two transcripts of NnCenH3, NnCenH3-A and NnCenH3-B were obtained from tender leaves using the RACE method. The sequence alignment showed that the CDS of NnCenH3-A is identical to that of the histone H3-like centromeric protein HTR12 described by Ming *et al.* (2013; GenBank accession numbers *XM_010268168* and *XM_010268169*). Two CenH3 variants were first identified in *Arabidopsis halleri* and *Arabidopsis lyrata* (Kawabe *et al.*, 2006), in allotetraploid rice (Hirsch *et al.*, 2009), and

in *Luzula nivea* (Moraes *et al.*, 2011). In grass, for instance, the two CenH3 isoforms originated from a common ancestor and play essential roles in centromere function, despite discrepancies in their expression in wheat (*Triticum* spp.; Yuan *et al.*, 2015). We failed to detect the NnCenH3-B protein in a western blot assay (Figure S2), however, probably as a result of the decay of its mRNA, or because of low protein expression in the roots, seedlings and leaves. In human cells, the *Arg*80 and *Gly*81 mutations in the CENP-A loop-1 region reduced CENP-A deposition at centromeres (Tachiwana *et al.*, 2011). Conceivably, even if the NnCenH3-B protein is translated in these tissues, the partial loss of CATD in the alternatively spliced isoform suggests that NnCenH3-B nucleosome assembly may be interrupted. This would be expected because of the presence of stable defects in the NnCenH3-B-defined centromeres caused by the loss of loop 1. Although similar truncated transcripts have been found in humans, their function remains unclear (Mammalian Gene Collection (MGC) Program Team. 2002). Thus, our findings for NnCenH3-B suggest that CenH3 alternative transcripts are expressed in plants. Future studies are needed to determine the evolutionary and functional significance of these findings.

Active genes in NnCenH3 nucleosomes

The ChIP analyses performed in earlier studies showed that CenH3 nucleosome-associated genes exist in neocentromeres and exhibit repeatless properties in plants (Yan *et al.*, 2005, 2006); however, the activities of these genes were entirely incompatible with core centromeres. This incompatibility is not observed in the sacred lotus genome. We identified nine NnCenH3 nucleosome-associated genes (*NNU_16350*, *NNU_08544*, *NNU_15386*, *NNU_02585*, *NNU_02586*, *NNU_19813*, *NNU_09838*, *NNU_00143* and *NNU_05636*) with relatively low transcription levels, suggesting the compatibility of these genes with centromeres. This compatibility is supported by the fact that a potato gene (*PGSC0003DMG400012074*) in the repeatless-based centromere of chromosome 11, which shows transcriptional activities and fewer repeats, is also associated with CenH3 nucleosomes (Gong *et al.*, 2012).

Ty1/*copia* and Ty3/*gypsy* retrotransposons in NnCBS

In higher eukaryotes, retrotransposons are conducive to centromere evolution and the stabilization of chromosome structure, a fact that explains the ubiquitous nature of centromeres (Topp *et al.*, 2004; Slotkin and Martienssen, 2007; Neumann *et al.*, 2011). Despite the prevalence of L1 retroelements in human centromeres (Chueh *et al.*, 2005), the chromovirus CRM clade of Ty3/*gypsy* LTR retrotransposons represents the most common class of retroelements in the centromeric regions of plant genomes (Neumann *et al.*, 2011); however, we identified Ty1/*copia*-type centromeric retrotransposons in the sacred lotus genome. Ty1/*copia* retrotransposons are reportedly always dispersed over the chromosomal arms (Brandes *et al.*, 1997). On the contrary, Ty1/*copia* elements are concentrated in the centromere regions of *Cicer arietinum* (chickpea) and *A. thaliana* (Brandes *et al.*, 1997), similar to the situation found in sacred lotus centromeres, in which Ty1/*copia* exhibited a concentrated distribution pattern. Apparently, the proportions of TEs in the centromeres and genome of sacred lotus contributes to the correlation between the statistics of repeats in NnCBS and the genome (Table 1). In fact, although the Ty3/*gypsy* class is the most abundant type of transposon in centromeres, the Ty1/*copia* class is predominant in the *Musa acuminata* (banana) genome (Hřibová *et al.*, 2010). In addition, the Ty3/*gypsy* class has not significantly decreased in abundance in NnCBS, suggesting that the coexistence of Ty3/*gypsy* and Ty1/*copia* in the sacred lotus centromeres confers an adaptive advantage. Ty3/*gypsy*-enriched centromeres have been reported in most plants, suggesting that the Ty3/*gypsy* class might have advantages over the Ty1/*copia* class during centromere evolution.

The centromere dynamics of the sacred lotus genome

The centromere-associated repeats experienced rapid evolution in the absence of selective constraints. Although seven identified repeat clusters in sacred lotus that showed high similarity to the Ty1/*copia* clade of LTR retrotransposons mainly specifically mapped to the primary constriction of chromosomes, these clusters displayed three different distribution patterns on the chromosomes (Figure 6), and these distribution patterns differed from the distribution patterns of the ChIP-ed DNA-based probes (Figure 2c). Owing to the low percentages (ranging from 0.8 to 2.1%) of cluster reads in ChIP-ed DNA (Table 2), the cluster signals were theoretically represented by 0.8–2.1% of the ChIP-ed DNA signals when we used ChIP-ed DNA-based FISH probes. Thus, it was difficult to detect and distinguish the distributions of these clusters from those of FISH signals obtained under the same experimental conditions. When we used highly pure, concentrated DNA fragments of centromere-associated clusters as FISH probes, however, the specific features of the FISH signals of the clusters were amplified, and showed distinctive distribution patterns on the sacred lotus chromosomes that indicated the dynamic nature of the sacred lotus centromeres (Figure 7).

(i) Chromosome 1 contains two domains with CL6 signals (Figure 6, CL6), one of which co-mapped with the signals of ChIP-ed DNA and NnCenH3, the hallmark of functional centromeres (Figures 2 and S8). Additional signals that did not overlap with the NnCenH3 signals were detected by both the CL6 probes and the ChIP-ed DNA probes. This may be because of the presence of a

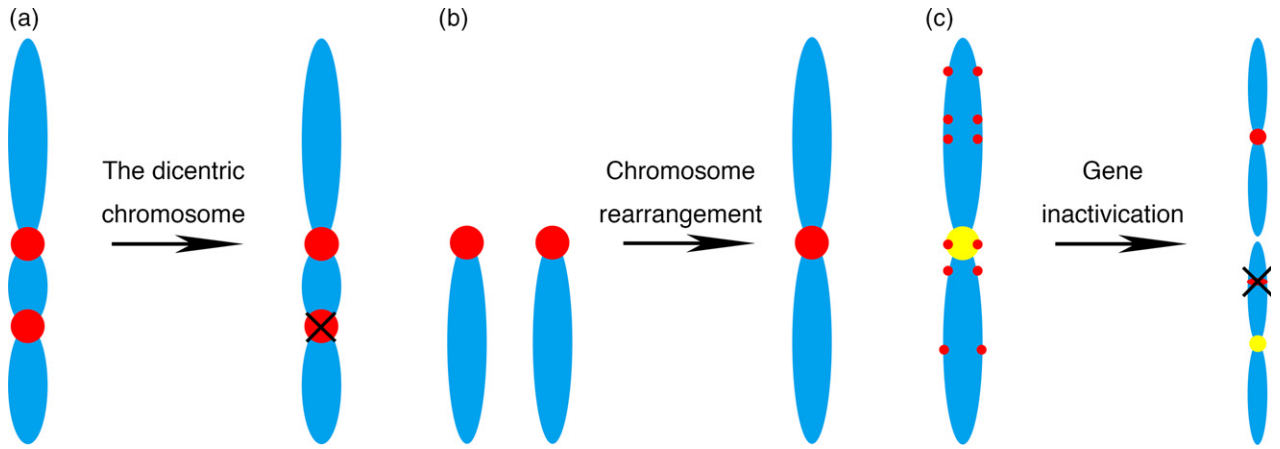

Figure 7. Dynamics of primary centromeric repeats on sacred lotus chromosomes. (a) The centromeres (indicated by the red circle with the black cross) was inactivated on the stable dicentric chromosome (indicated by the long blue ovals with red circles). (b) The telomere-based centromeres (red circles) might be derived from chromosome rearrangements. (c) Gene inactivation might result from large arrays of repeat insertions; the arrangement of these insertions showed that repetitive DNA preferentially resides on centromeric heterochromatin regions. The red circles indicate newly inserted repetitive sequences; the yellow circles indicate the ancient centromeric DNA sequences. The red circle with the black cross indicates inactivated genes containing repetitive sequences.

dicentric chromosome (Figure 7a). In general, dicentric or multicentric chromosomes are unstable as a result of anaphase bridge formation, which leads to faulty chromosome segregation (Gisselsson *et al.*, 2000; Shimizu *et al.*, 2005). Therefore, stable dicentric or multicentric chromosomes containing only one pair of functional centromeres are viable for posterity. In sacred lotus, the non-centromeric signals of CL6 indicated an inactive centromere in Chr. 1.

(ii) The centromere-associated clusters CL19 and CL21 were located at terminal regions of some chromosomes (Figure 6, CL19 and 21, inset). A similar homology between centromeric repeats and telomeric or subtelomeric repeats has been found in other plants (Tek and Jiang, 2004; Torres *et al.*, 2011; He *et al.*, 2013; Emadzade *et al.*, 2014). The 'centromeres from telomeres' hypothesis and chromosome rearrangement (Villasante *et al.*, 2007) may explain the origin of telomere-like centromeric repeats (Figure 7b). Thus, chromosomes containing CL19 or CL21 repeats on centromeres might be subject to chromosome rearrangement.

(iii) Ty1/*copia* retroelements were prone to disperse along the chromosomes (Brandes *et al.*, 1997), and some centromeric LTR retrotransposon families within the Ty3/*gypsy* class were not concentrated in centromeres (Neumann *et al.*, 2011). We observed interspersed CL19 signals on the arms of Chr. 1 (Figure 6, CL19). Genes containing large arrays of repeats in euchromatin regions will be inactive, and this is likely to result in species elimination; however, centromeric heterochromatin regions without selective pressure are conducive to the rapid evolution of repeats (Figure 7c), and this process might have resulted in the centromeric preference of the Ty1/*copia* retrotransposon (CL19) distribution in the sacred lotus genome. Thus, these observed intrachromosomal discrepancies of the centromeric repeat distribution demonstrate the high frequency of dynamic rearrangement, which may have resulted in sequence diversity in sacred lotus centromeres.

Herein, we first describe the centromere profile of an ancient eudicot, i.e. the sacred lotus, and analyzed how it differs from other sequenced angiosperm centromeres. The fully sequenced genome of *Amborella trichopoda*, the most ancient angiosperm, suggests that transposable elements were inserted early in the evolution of the ancient monotypic genus (Amborella Genome, 2013). The sacred lotus centromere may be older than the centromeres of other eudicots as a result of the very slow rate of evolution within its genome (Wang *et al.*, 2013). This study may provide valuable insights into centromere evolution. Because the available information on sacred lotus evolution is limited, additional sequence data on the centromeres in basal eudicots and other ancient species, such as basal angiosperms and gymnosperms, are needed to fully characterize centromere evolution.

EXPERIMENTAL PROCEDURES

Plant materials

The materials used in the experiments were grown in pools at the Department of Genetics, College of Life Sciences, in Wuhan University (Wuhan, China; 30°340″N, longitude 114°17″E).

Cloning of *NnCenH3* and *NnH3.3* cDNA

Total RNA was extracted from the fresh leaves of sacred lotus using pBIOZOL reagent (Bioer Technology, http://www.bioer.

com.cn) according to the manufacturer's instructions. Reverse-transcribed total RNA was generated using the Fermentas RevertAid First Strand cDNA Synthesis Kit (Fermentas, now ThermoFisher Scientific, http://www.thermofisher.com). The degenerate primers PGTVAL-C and QEAAE-C (Table S2) were designed based on multiple alignments of the conserved regions of CenH3 among Brassicaceae species. The partial regions of the CenH3 and H3 genes were obtained by RT-PCR. To obtain the full-length CenH3 and H3 cDNA of sacred lotus, 3′ and 5′ RACE-PCR was then performed using a 3′ and 5′ Full RACE Kit (Takara, now Clontech, http://www.clontech.com). The primers used for RACE-PCR are listed in Table S2. The RACE-PCR products of the expected lengths were cloned into the pMD18-T vector (Takara) and sequenced. The full-length cDNA and predicted protein sequences were analyzed and aligned using CLUSTALW.

Preparation of anti-NnCenH3 antibody and western blot analysis

The full-length NnCenH3 CDS was amplified using the primers pET28a-*Eco*RI and pET28a-*Xho*I (Table S2), inserted into the pET28a vector containing two 6-His tags and transformed into freshly prepared Rosetta (DE3) competent cells. Expression of the fusion protein was induced using 1 M isopropyl-β-D-1-thiogalactopyranoside (IPTG) at 16°C with shaking at 2.05 ***g*** for 24 h; cellular proteins were then analyzed via 15% SDS-PAGE. Anti-NnCenH3 polyclonal antibody was prepared against the synthetic peptide LFRKKRRQSSRLSS, which is located between seven and 20 amino acid residues away from the predicted NnCenH3 protein, and was custom synthesized and purified by CWBio Co., Ltd (http://www.cwbiotech.com). For western blotting analysis, total proteins were separated by SDS-PAGE and transferred to nitrocellulose membranes (Whatman, now GE Healthcare Life Sciences, http://www.gelifesciences.com). The membrane was blocked with 5% non-fat milk. After incubation with the primary (anti-NnCenH3 antibody at a dilution of 1:1000) and secondary (goat anti-rabbit IgG conjugated with alkaline phosphatase at a dilution of 1:2000) antibodies, the blotted proteins were visualized with 5-bromo-4-chloro-3-indolyl phosphate (BCIP)/*p*-nitro-blue tetrazolium chloride (NBT).

FISH and immunostaining assays

FISH was performed as described by Li *et al.* (2001) with minor modifications. The generated random-primed DNA probes for ChIP-ed DNA and centromeric clusters were labeled with either biotin-dUTP (ThermoFisher Scientific) or digoxigenin-dUTP (Roche, http://www.roche.com) by nick translation. The root tips for the chromosome preparation were pre-treated with saturated α-bromonaphthalene at 28°C for 2 h to obtain metaphase cells, which were fixed in Carnoy's solution (ethanol:glacial acetic acid = 3:1, v/v) overnight at 4°C. The root tips were treated with 0.1 M HCl for 1 min at 60°C, followed by 2 h of digestion with a mixture of 2.5% pectolyase and 2.5% cellulose (pH = 4.2) at 37°C. After digestion, the treated root tips were crushed on slides and air-dried. The prepared slides were stored at −20°C. The prepared chromosomes on the slides were treated with 1% formaldehyde for 10 min, denatured in 70% formamide for 5 min at 90°C, and chilled in 70, 95, and 100% ethanol at −20°C for 5 min each. The slides were air-dried at room temperature (20–25°C). Meanwhile, a hybridization mixture containing 50% deionized formamide, 2 × SSC (0.3 M NaCl and 0.03 M sodium citrate), 1 mg ml^{-1} sheared salmon sperm DNA, 1 μg ml^{-1} DNA probe and 10% dextran sulfate was denatured for 5 min at 75°C and immediately cooled on ice. The slides were incubated in the hybridization mixture for 16 h at 37°C. To detect biotin- or digoxigenin-labeled probes, the slides were incubated with 50 μl of streptavidin-Cy3 (Jackson Immunoresearch, https://www.jacksonimmuno.com) or sheep anti-digoxin FITC (Roche) for 1 h at 37°C. After incubation with the primary antibodies, the slides treated with biotin-labeled probes were incubated with biotinylated anti-streptavidin antibody (Vector Laboratories, http://vectorlabs.com), and the slides with digoxigenin-labeled probes were incubated with rabbit anti-sheep FITC (Roche) for 1 h at 37°C.

Immunostaining was carried out according to a previously described protocol (Yang *et al.*, 2010). In brief, root tips pre-treated with saturated α-bromonaphthalene were fixed in 4% (w/v) paraformaldehyde for 40 min at 4°C, and then digested with a mixture of 2.5% pectolyase and 2.5% cellulose dissolved in 1 × PBS for 0.5–1.0 h at 37°C. The root tips were squeezed onto slides and stored at −80°C before use. The prepared slides were blocked with 1% bovine serum albumin in PBS-T (1% Triton X-100 in 1 × PBS) for 1 h at 37°C. Subsequently, the slides were incubated with anti-NnCenH3 antibody (1 : 100 dilution) at 4°C overnight. For secondary antibody incubation, the slides were incubated with FITC-conjugated Affinipure goat anti-rabbit IgG (1:50 dilution; Proteintech, http://www.ptglab.com) for 1 h at 37°C.

The chromosomes used for the FISH and immunostaining assays were counterstained by adding 15 μl of 4′,6′-diamidino-2-phenylindole (DAPI, 10 μg ml^{-1}) in anti-fade Vectashield agent H1000 (Vector Laboratories, http://vectorlabs.com). Images of the chromosomes were acquired with an Olympus DP80 2CCD color/monochrome camera system mounted on an Olympus BX60 microscope.

ChIP, ChIP-seq and ChIP-qPCR

ChIP assays were performed using NnCenH3 antibodies as previously described (O'Neill and Turner, 2003), with modifications. Tender leaves (20 g) were ground to a fine, dry powder in liquid nitrogen and 2 g of polyvinyl-polypyrrolidone (PVPP). Chromatin was extracted from these tissues in isolation buffer and fragmented using micrococcal nuclease (New England Biolabs, http://www.neb.com). The digested chromatin was used for the ChIP assay. ChIP-ed DNA (10 ng) was submitted to the Beijing Genomics Institute (BGI, http://www.genomics.cn) for ChIP-seq library construction and sequencing on the Illumina Hiseq 2000 platform. Over 20 million clean reads were aligned with the sacred lotus genome from LOTUS-DB 1.0 (http://lotus-db.wbgcas.cn; Wang *et al.*, 2015) using the SOAP2 aligner. To identify the NnCenH3 domains, ChIP-seq read-enriched regions were analyzed by SICER 1.1, using the parameters described by Gong *et al.* (2012). Gaps within the full length of NnCBS sequences were not counted. Quantitative PCR (qPCR) analysis of ChIP samples was performed using the ABI Step One Plus Real-Time PCR system (Applied Biosystems, now ThermoFisher Scientific, http://www.thermofisher.com) and FastStart Universal SYBR Green Master (ROX; Roche). Each reaction was carried out in triplicate. The calculated relative fold enrichment (RFE) for each active gene in NnCBS was normalized based on the $\Delta\Delta C_t$ method using the SuperArray ChIP-qPCR Data Analysis Template (Chakrabarti *et al.*, 2002). The primers used in the ChIP-qPCR experiments are listed in Table S2.

Expression of NnCBS-associated genes

The annotated gene expression (FPKM) was obtained from the transcript file of the reference sacred lotus genome, LOTUS-DB 1.0 (http://lotus-db.wbgcas.cn; Wang *et al.*, 2015). The transcription levels of 16 NnCenH3-associated genes were verified by RT-PCR. Total RNA was isolated from the leaves of sacred lotus

and treated with DNase I (ThermoFisher Scientific) to remove template DNA. Double-distilled water (ddH_2O) was used as a negative control, and the centromere-irrelevant *NnEF1a* gene (GenBank accession number AB491177.1), which is highly expressed, served as a positive control for comparison with the expression of NnCenH3-associated genes. gDNA was used as the template to prove the existence of the genes that failed to amplify (*NNU_16348*, *NNU_16349*, *NNU_21723*, *NNU_07763*, *NNU_03501*, *NNU_11507* and *NNU_15700*) with the cDNA template. For RT-PCR, the number of amplification cycles was initially limited to 33. The cycle number was subsequently increased to 40 to clarify the transcription of low-expression genes. PCR assays were performed with the following parameters: 2 min at 94°C, followed by 33 or 40 cycles of 30 sec at 94°C, 30 sec at 65°C and 6 sec at 72°C. The reactions were then heated to 70°C for 2 min. The PCR products were analyzed by electrophoresis on 3% agarose gels.

Identification and characterization of centromeric repeats

To determine the centromere-associated repetitive sequence, the similarities of the sacred lotus NnCBS, rice *Cen8* and potato centromeres to the repeat database were assessed using RepeatMasker (Smit *et al.*, 2013), with default parameters. The Lotus Repeat Database was kindly provided by Professor Ray Ming (Ming *et al.*, 2013). Rice *Cen8*, obtained from the Rice Genome Research Program (RGP; Wu *et al.*, 2004), and the rice repeat database, obtained from the Rice Annotation Project Database (RAP-DB; Ouyang and Buell, 2004), were used to analyze the repetitive sequence in rice *Cen8*. The sequences and repeat database in potato centromeres were downloaded from http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml (Potato Genome Sequencing Consortium *et al.*, 2011). perl scripts and r scripts were used to parse and visualize the distribution of centromere-associated genes and LTR retrotransposons. The raw sequence data of the sacred lotus genome from the GenBank Short Read Archive raw data (GenBank accession number SRR2131192) were used to reconstruct genomic repeats using a graph-based clustering analysis (Novak *et al.*, 2010, 2013). The clusters obtained were mapped using 10 million ChIP-seq reads and PatMaN (Prüfer *et al.*, 2008), using parameters described by Gong *et al.* (2012). The primers of ChIP read-associated clusters (Table S2) were designed to construct recombinant clones for the FISH assay.

Data access

The ChIP-seq data generated in this study are available from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/) under accession number SRR2970623.

ACKNOWLEDGEMENTS

We thank Professor Ray Ming (Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, The Chinese Academy of Sciences; Department of Plant Biology, University of Illinois, USA) for kindly providing the Lotus Repeat Database. This work is supported by National Natural Science Foundation of China (31271310).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Coding sequences of *NnCenH3-A*, *NnCenH3-B* and *NnH3.3* identified by homology based-cloning in this study.

Figure S2. Expression of NnCenH3 *in vitro* (a and b) and *in vivo* (c).

Figure S3. Alignment of ChIP-seq reads with the reference genome.

Figure S4. Distribution of ChIP-ed reads, genes and retrotransposons in NnCBS.

Figure S5. Similarities of NnCenH3 subdomain-associated genes and their homologs.

Figure S6. Relative fold enrichment (RFE) of genes in NnCenH3 subdomains.

Figure S7. Accumulations of mapped clusters in ChIP-seq reads.

Figure S8. Chromosomal localization of the CL6 and ChIP-ed DNA probes.

Figure S9. Graph layouts of read similarities in centromere-associated repeat clusters.

Table S1. Transcript data of CBS-associated genes in sacred lotus

Table S2. Primers used in the study

Table S3. Abundance and RepeatMasker analysis of clusters in the sacred lotus genome

REFERENCES

Alkan, C., Cardone, M.F., Catacchio, C.R. *et al.* (2011) Genome-wide characterization of centromeric satellites from multiple mammalian genomes. *Genome Res.* **21**, 137–145.

Amborella Genome, P. (2013) The Amborella genome and the evolution of flowering plants. *Science*, **342**, 1241089.

Billia, F. and de Boni, U. (1991) Localization of centromeric satellite and telomeric DNA sequences in dorsal root ganglion neurons, *in vitro*. *J. Cell Sci.* **100** (Pt 1), 219–226.

Black, B.E., Foltz, D.R., Chakravarthy, S., Luger, K., Woods, V.L. and Cleveland, D.W. (2004) Structural determinants for generating centromeric chromatin. *Nature*, **430**, 578–582.

Black, B.E., Jansen, L.E., Maddox, P.S., Foltz, D.R., Desai, A.B., Shah, J.V. and Cleveland, D.W. (2007) Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. *Mol. Cell*, **25**, 309–322.

Brandes, A., Heslop-Harrison, J.S., Kamm, A., Kubis, S., Doudrick, R.L. and Schmidt, T. (1997) Comparative analysis of the chromosomal and genomic organization of Ty1-copia-like retrotransposons in pteridophytes, gymnosperms and angiosperms. *Plant Mol. Biol.* **33**, 11–21.

Chakrabarti, S.K., James, J.C. and Mirmira, R.G. (2002) Quantitative assessment of gene targeting *in vitro* and *in vivo* by the pancreatic transcription factor, Pdx1. Importance of chromatin structure in directing promoter binding. *J. Biol. Chem.* **277**, 13286–13293.

Chueh, A.C., Wong, L.H., Wong, N. and Choo, K.H. (2005) Variable and hierarchical size distribution of L1-retroelement-enriched CENP-A clusters within a functional human neocentromere. *Hum. Mol. Genet.* **14**, 85–93.

Dong, F. and Jiang, J. (1998) Non-Rabl patterns of centromere and telomere distribution in the interphase nuclei of plant cells. *Chromosome Res.* **6**, 551–558.

Emadzade, K., Jang, T.S., Macas, J., Kovařík, A., Novák, P., Parker, J. and Weiss-Schneeweiss, H. (2014) Differential amplification of satellite PaB6 in chromosomally hypervariable Prospero autumnale complex (Hyacinthaceae). *Ann. Bot.* **114**, 1597–1608.

Fachinetti, D., Folco, H.D., Nechemia-Arbely, Y. *et al.* (2013) A two-step mechanism for epigenetic specification of centromere identity and function. *Nat. Cell Biol.* **15**, 1056–1066.

Foltz, D.R., Jansen, L.E., Bailey, A.O., Yates, J.R., Bassett, E.A., Wood, S., Black, B.E. and Cleveland, D.W. (2009) Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. *Cell*, **137**, 472–484.

Fu, S., Lv, Z., Gao, Z. *et al.* (2013) *De novo* centromere formation on a chromosome fragment in maize. *Proc. Natl Acad. Sci. USA*, **110**, 6033–6036.

Garrido-Ramos, M.A. (2015) Satellite DNA in plants: more than just rubbish. *Cytogenet. Genome Res.* **146**, 153–170.

Gisselsson, D., Pettersson, L., Höglund, M., Heidenblad, M., Gorunova, L., Wiegant, J., Mertens, F., Dal Cin, P., Mitelman, F. and Mandahl, N. (2000) Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proc. Natl Acad. Sci. USA*, **97**, 5357–5362.

Gong, Z., Wu, Y., Koblizkova, A. *et al.* (2012) Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell*, **24**, 3559–3574.

He, L., Liu, J., Torres, G.A., Zhang, H., Jiang, J. and Xie, C. (2013) Interstitial telomeric repeats are enriched in the centromeres of chromosomes in Solanum species. *Chromosoma Res.* **21**, 5–13.

Henikoff, S., Ahmad, K. and Malik, H.S. (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, **293**, 1098–1102.

Hirsch, C.D., Wu, Y., Yan, H. and Jiang, J. (2009) Lineage-specific adaptive evolution of the centromeric protein CENH3 in diploid and allotetraploid Oryza species. *Mol. Biol. Evol.* **26**, 2877–2885.

Hřibová, E., Neumann, P., Matsumoto, T., Roux, N., Macas, J. and Doležel, J. (2010) Repetitive part of the banana (Musa acuminata) genome investigated by low-depth 454 sequencing. *BMC Plant Biol.* **10**, 204.

Jin, W., Melo, J.R., Nagaki, K., Talbert, P.B., Henikoff, S., Dawe, R.K. and Jiang, J. (2004) Maize centromeres: organization and functional adaptation in the genetic background of oat. *Plant Cell*, **16**, 571–581.

Kawabe, A., Nasuda, S. and Charlesworth, D. (2006) Duplication of centromeric histone H3 (HTR12) gene in Arabidopsis halleri and A. lyrata, plant species with multiple centromeric satellite sequences. *Genetics*, **174**, 2021–2032.

Kuppu, S., Tan, E.H., Nguyen, H., Rodgers, A., Comai, L., Chan, S.W.L. and Britt, A.B. (2015) Point mutations in centromeric histone induce post-zygotic incompatibility and uniparental inheritance. *PLoS Genet.* **11**, e1005494.

Lefrancois, P., Auerbach, R.K., Yellman, C.M., Roeder, G.S. and Snyder, M. (2013) Centromere-like regions in the budding yeast genome. *PLoS Genet.* **9**, e1003209.

Li, L.J., Arumuganathan, K., Rines, H.W., Phillips, R.L., Riera-Lizarazu, O., Sandhu, D., Zhou, Y. and Gill, K.S. (2001) Flow cytometric sorting of maize chromosome 9 from an oat-maize chromosome addition line. *Theor. Appl. Genet.* **102**, 658–663.

Li, G.R., Liu, C., Wei, P., Song, X.J. and Yang, Z.J. (2012) Chromosomal distribution of a new centromeric Ty3-gypsy retrotransposon sequence in Dasypyrum and related Triticeae species. *J. Genet.* **91**, 343–348.

Li, B., Choulet, F., Heng, Y., Hao, W., Paux, E., Liu, Z., Yue, W., Jin, W., Feuillet, C. and Zhang, X. (2013) Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. *Plant J.* **73**, 952–965.

Liu, Z., Yue, W., Li, D., Wang, R.R.C., Kong, X., Lu, K., Wang, G., Dong, Y., Jin, W. and Zhang, X. (2008) Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. *Chromosoma*, **117**, 445–456.

Logsdon, G.A., Barrey, E.J., Bassett, E.A., DeNizio, J.E., Guo, L.Y., Panchenko, T., Dawicki-McKenna, J.M., Heun, P. and Black, B.E. (2015) Both tails and the centromere targeting domain of CENP-A are required for centromere establishment. *J. Cell Biol.* **208**, 521–531.

Lomiento, M., Jiang, Z., D'Addabbo, P., Eichler, E.E. and Rocchi, M. (2008) Evolutionary-new centromeres preferentially emerge within gene deserts. *Genome Biol.*, **9**, R173.

Luo, S., Mach, J., Abramson, B., Ramirez, R., Schurr, R., Barone, P., Copenhaver, G. and Folkerts, O. (2012) The cotton centromere contains a Ty3-gypsy-like LTR retroelement. *PLoS ONE*, **7**, e35261.

Malik, H.S. and Henikoff, S. (2009) Major evolutionary transitions in centromere complexity. *Cell*, **138**, 1067–1082.

Mammalian Gene Collection (MGC) Program Team. (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.

Mehta, G.D., Agarwal, M.P. and Ghosh, S.K. (2010) Centromere identity: a challenge to be faced. *Mol. Genet. Genomics*, **284**, 75–94.

Melters, D.P., Bradnam, K.R., Young, H.A. *et al.* (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10.

Miller, J.T., Dong, F., Jackson, S.A., Song, J. and Jiang, J. (1998) Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. *Genetics*, **150**, 1615–1623.

Ming, R., VanBuren, R., Liu, Y. *et al.* (2013) Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). *Genome Biol.* **14**, R41.

Moraes, I.C., Lermontova, I. and Schubert, I. (2011) Recognition of A. thaliana centromeres by heterologous CENH3 requires high similarity to the endogenous protein. *Plant Mol. Biol.* **75**, 253–261.

Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C.R., Cheng, Z. and Jiang, J. (2005) Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Mol. Biol. Evol.* **22**, 845–855.

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J. and Macas, J. (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob. DNA*', **2**, 4.

Novak, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J. and Macas, J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

O'Neill, L.P. and Turner, B.M. (2003) Immunoprecipitation of native chromatin: NChIP. *Methods*, **31**, 76–82.

Ouyang, S. and Buell, C.R. (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.* **32**, D360–D363.

Pluta, A.F., Mackay, A.M., Ainsztein, A.M., Goldberg, I.G. and Earnshaw, W.C. (1995) The centromere: hub of chromosomal activities. *Science*, **270**, 1591–1594.

Potato Genome Sequencing Consortium; Xu, X., Pan, S. *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.

Presting, G.G., Malysheva, L., Fuchs, J. and Schubert, I. (1998) A Ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant J.* **16**, 721–728.

Prüfer, K., Stenzel, U., Dannemann, M., Green, R.E., Lachmann, M. and Kelso, J. (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, **24**, 1530–1531.

Ravi, M., Kwong, P.N., Menorca, R.M., Valencia, J.T., Ramahi, J.S., Stewart, J.L., Tran, R.K., Sundaresan, V., Comai, L. and Chan, S.W.L. (2010) The rapidly evolving centromere-specific histone has stringent functional requirements in Arabidopsis thaliana. *Genetics*, **186**, 461–471.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.* **37**, 501–506.

Shimizu, N., Shingaki, K., Kaneko-Sasaguri, Y., Hashizume, T. and Kanda, T. (2005) When, where and how the bridge breaks: anaphase bridge breakage plays a crucial role in gene amplification and HSR generation. *Exp. Cell Res.* **302**, 233–243.

Slotkin, R.K. and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285.

Smit, A.F. and Riggs, A.D. (1996) Tiggers and DNA transposon fossils in the human genome. *Proc. Natl Acad. Sci. USA*, **93**, 1443–1448.

Smit, A., Hubley, R. and Green, P. (2013) *RepeatMasker 4.0*. Institute for Systems Biology: Seattle, WA.

Sullivan, L.L., Boivin, C.D., Mravinac, B., Song, I.Y. and Sullivan, B.A. (2011) Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells. *Chromosome Res.* **19**, 457–470.

Tachiwana, H., Kagawa, W., Shiga, T. *et al.* (2011) Crystal structure of the human centromeric nucleosome containing CENP-A. *Nature*, **476**, 232–235.

Talbert, P.B., Masuelli, R., Tyagi, A.P., Comai, L. and Henikoff, S. (2002) Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. *Plant Cell*, **14**, 1053–1066.

Tek, A.L. and Jiang, J. (2004) The centromeric regions of potato chromosomes contain megabase-sized tandem arrays of telomere-similar sequence. *Chromosoma*, **113**, 77–83.

Topp, C.N., Zhong, C.X. and Dawe, R.K. (2004) Centromere-encoded RNAs are integral components of the maize kinetochore. *Proc. Natl Acad. Sci. USA*, **101**, 15986–15991.

Torres, G.A., Gong, Z., Iovene, M., Hirsch, C.D., Buell, C.R., Bryan, G.J., Novák, P., Macasand, J. and Jiang, J. (2011) Organization and evolution of subtelomeric satellite repeats in the potato genome. *G3*, **1**, 85–92.

Tudor, M., Lobocka, M., Goodell, M., Pettitt, J. and O'Hare, K. (1992) The pogo transposable element family of Drosophila melanogaster. *Mol. Gen. Genet.* **232**, 126–134.

Verdaasdonk, J.S. and Bloom, K. (2011) Centromeres: unique chromatin structures that drive chromosome segregation. *Nat. Rev. Mol. Cell Biol.* **12**, 320–332.

Vermaak, D., Hayden, H.S. and Henikoff, S. (2002) Centromere targeting element within the histone fold domain of Cid. *Mol. Cell. Biol.* **22**, 7553–7561.

Villasante, A., Abad, J.P. and Mendez-Lago, M. (2007) Centromeres were derived from telomeres during the evolution of the eukaryotic chromosome. *Proc. Natl Acad. Sci. USA*, **104**, 10542–10547.

Wang, G., He, Q., Liu, F., Cheng, Z., Talbert, P.B. and Jin, W. (2011) Characterization of CENH3 proteins and centromere-associated DNA sequences in diploid and allotetraploid Brassica species. *Chromosoma*, **120**, 353–365.

Wang, Y., Fan, G., Liu, Y. *et al.* (2013) The sacred lotus genome provides insights into the evolution of flowering plants. *Plant J.* **76**, 557–567.

Wang, K., Wu, Y., Zhang, W., Dawe, R.K. and Jiang, J. (2014) Maize centromeres expand and adopt a uniform size in the genetic background of oat. *Genome Res.* **24**, 107–116.

Wang, K., Deng, J., Damaris, R.N., Yang, M., Xu, L. and Yang, P. (2015) LOTUS-DB: an integrative and interactive database for Nelumbo nucifera study. *Database (Oxford)*, **2015**, bav023.

Weber, B. and Schmidt, T. (2009) Nested Ty3-gypsy retrotransposons of a single Beta procumbens centromere contain a putative chromodomain. *Chromosome Res.* **17**, 379–396.

Westhorpe, F.G. and Straight, A.F. (2015) The centromere: epigenetic control of chromosome segregation during mitosis. *Cold Spring Harb. Perspect. Biol.* **7**, a015818.

Wolfgruber, T.K., Sharma, A., Schneider, K.L. *et al.* (2009) Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons. *PLoS Genet.* **5**, e1000743.

Wu, J., Yamagata, H., Hayashi-Tsugane, M. *et al.* (2004) Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell*, **16**, 967–976.

Wu, Z., Gui, S., Quan, Z., Pan, L., Wang, S., Ke, W., Liang, D. and Ding, Y. (2014a) A precise chloroplast genome of Nelumbo nucifera (Nelumbonaceae) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. *BMC Plant Biol.* **14**, 289.

Wu, Z., Gui, S., Wang, S. and Ding, Y. (2014b) Molecular evolution and functional characterisation of an ancient phenylalanine ammonia-lyase gene (NnPAL1) from Nelumbo nucifera: novel insight into the evolution of the PAL family in angiosperms. *BMC Evol. Biol.* **14**, 100.

Yan, H., Jin, W., Nagaki, K., Tian, S., Ouyang, S., Buell, C.R., Talbert, P.B., Henikoff, S. and Jiang, J. (2005) Transcription and histone modifications in the recombination-free region spanning a rice centromere. *Plant Cell*, **17**, 3227–3238.

Yan, H., Ito, H., Nobuta, K. *et al.* (2006) Genomic and genetic characterization of rice Cen3 reveals extensive transcription and evolutionary implications of a complex centromere. *Plant Cell*, **18**, 2123–2133.

Yan, H., Talbert, P.B., Lee, H.R., Jett, J.H., Henikoff, S., Chen, F. and Jiang, J. (2008) Intergenic locations of rice centromeric chromatin. *PLoS Biol.* **6**, e286.

Yang, F., Zhang, L., Li, J., Huang, J., Wen, R., Ma, L., Zhou, D. and Li, L. (2010) Trichostatin A and 5-azacytidine both cause an increase in global histone H4 acetylation and a decrease in global DNA and H3K9 methylation during mitosis in maize. *BMC Plant Biol.* **10**, 178.

Yuan, J., Guo, X., Hu, J., Lv, Z. and Han, F. (2015) Characterization of two CENH3 genes and their roles in wheat evolution. *New Phytol.* **206**, 839–851.

Zalensky, A.O., Allen, M.F., Kobayashi, A., Zalenskaya, I.A., Balhorn, R. and Bradbury, E.M. (1995) Well-defined genome architecture in the human sperm nucleus. *Chromosoma*, **103**, 577–590.