Dynamic chromatin changes associated with de novo centromere formation in maize euchromatin

Handong Su¹,²,†, Yalin Liu¹,²,†, Yong-Xin Liu¹, Zhenling Lv¹, Hongyao Li³, Shaojun Xie³, Zhi Gao⁴, Junling Pang²,⁵, Xiu-Jie Wang⁵, Jinsheng Lai³, James A. Birchler⁴ and Fangpu Han¹

¹State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, 100101, China

²University of Chinese Academy of Sciences, Beijing, 100049, China

³Chinese Agriculture University, Beijing, 100193, China

⁴Division of Biological Sciences, University of Missouri-Columbia, Columbia, MO, 65211-7400, United States of America

⁵State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, 100101, China

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tpj.13305

This article is protected by copyright. All rights reserved.
†. Equal contributors

Corresponding authors James A Birchler and Fangpu Han

Dr. Fangpu Han
Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, 100101, China
Email: fphan@genetics.ac.cn
Phone: +86-10-64807926
Fax: +86-10-64854467

Dr. James A. Birchler
Division of Biological Sciences, University of Missouri-Columbia, Columbia, MO, 65211-7400, United States of America
Email: birchlerj@missouri.edu
Phone: 573-882-4905
Fax: 573-882-0123

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors is: Dr. Fangpu Han (fphan@genetics.ac.cn).

Running title: De novo centromeric chromatin in maize

This article is protected by copyright. All rights reserved.
Keywords: centromeric chromatin; DNA methylation; de novo centromere; maize

Accession numbers: GSE64641; GSE76533.

Summary

The inheritance and function of centromere are not strictly dependent on any specific DNA sequence, but involve an epigenetic component in most species. CENH3, a centromere histone H3 variant, is one of the best described epigenetic factors in centromere identity. However, the required chromatin features during centromere formation have not been revealed yet. We previously identified two de novo centromeres on maize minichromosomes derived from euchromatic sites with high-density gene distributions but low-density transposons distributions. The distribution of gene location and gene expression in these sites indicates that transcriptionally active regions can initiate de novo centromere formation and CENH3 seeding shows preference for gene-free regions or regions with no gene expression.

The locations of expressed genes detected were at relatively hypomethylated loci, and the altered gene expression was due to the de novo centromere formation but not to the additional copy of minichromosome. The initial overall DNA methylation level of the two de novo regions was at a low level, but increased substantially to that of native centromeres after centromere formation. These results illustrate the dynamic chromatin changes during euchromatin-originated de novo centromere formation, which provides insight into the mechanism of de novo centromere formation and regulation of subsequent consequences.

This article is protected by copyright. All rights reserved.
Introduction

Centromeres are specialized chromosomal regions that mediate the assembly of kinetochores for the attachment of spindle microtubules during mitosis and meiosis. Compared to the evolutionarily conserved function of centromeres, the organization and underlying DNA sequence of centromeres vary largely in different species, which has been coined the centromere paradox (Henikoff et al. 2001, Steiner and Henikoff 2015). The previously described functionally inactivated centromere on dicentric chromosomes in animals and plants (Earnshaw and Migeon 1985, Han et al. 2006) reveals an epigenetic mechanism for centromere identity. The centromere-specific histone H3 variant, CENH3 for plants (Talbert et al. 2002, Zhong 2002) or CENP-A for humans (Earnshaw and Migeon 1985, Palmer and Margolis 1985), is one of the most characterized epigenetic elements involved in establishing centromeric chromatin for centromere identity and function (Black et al. 2010, Guse et al. 2011, Mendiburo et al. 2011). However, the roles of epigenetic factors and the repetitive DNA sequence found in most centromeres for centromere function have not been fully elucidated.

Neocentromeres are generated on a new site following inactivation or deletion of an original centromere and are found in various species (Fu et al. 2013, Ketel et al. 2009, Maggert and Karpen 2001, Marshall et al. 2008, Nasuda et al. 2005, Olszak et al. 2011). Because of the non-repetitive nature of the underlying DNA sequence, de novo centromeres with well assembled sequences have been excellent tools for the detailed analysis of CENH3-binding domains and of the mechanism of centromere formation and maintenance. Human de novo centromeres preferentially form on AT-rich regions (Barry et al. 1999). Long
Interspersed Nuclear Element (LINE) retrotransposons show significant enrichment in human

de novo centromeres and the transcripts are essential epigenetic components for core
neocentromeric chromatin (Chueh et al. 2009). These results prompted us to hypothesize that
certain kinds of “markers” show preference for CENH3 loading during the initiation of de
novo centromeres. However, the required DNA sequence and chromatin structure during
centromeres formation is still a key understudied issue.

Accumulating evidence suggests that DNA methylation plays important roles in the
establishment and maintenance of centromeric chromatin (Gopalakrishnan et al. 2009, Yan et
al. 2010, Zhang et al. 2008). However, several ambiguous results were reported on the
changes of DNA methylation level following centromeric chromatin variation. DNA
fiber-based results indicate that the normal maize B centromere exhibits hypomethylation,
and the epigenetic state was modified into hypermethylation when the B centromere becomes
inactive (Koo et al. 2011). The DNA methylation level of a 723-kb region was not
significantly changed when it was involved in neocentromere formation in maize (Zhang et
al. 2013). Meanwhile, actively expressed genes in centromeric domains were reported in
DNA methylation, chromatin domain distribution and transcriptional competence has been
observed in a functional human de novo centromere (Wong et al. 2006). Thus, more data is
needed on the study of DNA methylation associated with neocentromere formation or
centromere inactivation, and the role of DNA methylation on the regulation of centromeric
transcription in centromeric chromatin.

This article is protected by copyright. All rights reserved.
We previously found two maize de novo centromeres on Dp3a (Duplication 3a) and Derivative 3-3 that were both located over euchromatic, protein-encoding regions of the genome (Fu et al. 2013, Liu et al. 2015). Here we investigate the features of chromosomal regions that form these centromeres and the dynamic changes of chromatin environments underlying these regions with respect to transcriptional competence and DNA methylation before and after de novo centromere formation using next generation sequencing technology.

The LTR/Gypsy retrotransposons show preference for CENH3 binding in the two de novo centromeric regions. Long genes or highly expressed genes combined with reduced tendency for DNA methylation at the boundary of de novo centromere suggest multi-layer mechanism for the regulation of centromere size. De novo centromeres are initiated by loading CENH3 within normally hypermethylated regions. BisChIP-seq (Bisulfite sequencing of chromatin immunoprecipitated DNA) results reveal a substantially increased DNA methylation level on these chromosomal sites after de novo centromere formation, reaching the level of native centromeres. The dynamic changes of DNA methylation patterns indicate that CENH3 seeding was affected by intrinsic DNA methylation patterns before neocentromere formation and the CENH3 loadings can also shape the DNA methylation patterns after de novo centromere formation. These results of de novo centromeres provide detailed insight into the epigenetic centromeric chromatin and the cis acting DNA sequences for centromere formation and maintenance.
Results

Two maize de novo centromeres form in euchromatic sites of chromosome arms.

Centromeres are often embedded in heterochromatin-rich and transcriptionally inert regions (Cardone et al. 2006, Lomiento et al. 2008, Ventura et al. 2007). De novo centromeres have been reported in gene-poor regions in different species (Ketel et al. 2009, Maggert and Karpen 2001, Nasuda et al. 2005, Olszak et al. 2011, Wang et al. 2014). Our previous study identified two de novo centromeres (Dp3a and Derivative 3-3) in maize (Fu et al. 2013, Liu et al. 2015). The immunostaining with kinetochore proteins revealed that Dp3a and Derivative 3-3 apparently assembles functional kinetochore structures similar to native centromeres. The Dp3a fragment was derived by UV irradiation. Derivative 3-3 minichromosome was a product of misdivision of the B centromere. A chromosome in the original maize lines (2N = 20) was split into two fragments, one of which gave rise to a novel centromere. The origin of the progenitor chromosome is unknown, which has been induced decades ago. The materials contain 20 normal originally chromosomes and one fragment with neocentromere. The normal chromosomes in these stocks are from a tester for maintenance of the material. We use offspring of the selfed materials (2N = 20) that do not containing the fragment as controls to characterizes the “pre-novel centromere” state.

The previous chromatin immunoprecipitation sequencing (ChIP-seq) results found a 350-kb and 288-kb CENH3 binding domain in maize chromosomes 3 and 9 for Dp3a and Derivative 3-3 de novo centromeres, respectively (Fu et al. 2013, Liu et al. 2015). The two regions were both present over euchromatic sites. The 350-kb region (3: 218,917,000 -
219,267,000) for the Dp3a de novo centromere has 12 non-TE genes (Figure 1b) annotated in maize B73 version AGPv3.22 reference, which is one of the highest gene-rich and TE-deficient regions on chromosome 3 (Figure 1a). The 288-kb region (9: 4,300,000 - 4,588,000) for Derivative 3-3 de novo centromere contains 7 non-TE genes (Figure 1d), which is also a gene-rich and TE-deficient region along chromosome 9 (Figure 1c). Genome wide, the gene density is 1 gene per 52 kb on average, and for the 10 native centromeres it changes from 1 gene per 112 kb to 362 kb. The gene density in the 350-kb region was 1 gene per 29.2 kb, which is higher than the density of 1 gene per 55.6 kb on average for chromosome 3. The gene density in the 288-kb region was 1 gene per 41.1 kb, which is higher than the level of 1 gene per 52.5 kb on average for chromosome 9. These results indicate that in addition to the traditionally gene poor regions in which de novo centromeres have been documented, the gene rich regions are also suitable for de novo centromere formation. Thus, we investigated the characteristics of the chromosomal regions including the DNA sequence preference, gene distribution and chromatin state.

The DNA sequence composition of CENH3 binding domains in the two de novo centromeres

The de novo centromeres are composed of unique sequences with an uneven distribution profile of CENH3 ChIP-seq reads due to the intermingled CENH3-occupied and CENH3-depleted subdomains (Figure 1b, d) as reported in other species (Gong et al. 2012, Yan et al. 2008). The MACS software (Feng et al. 2012) was used to call peaks as
CENH3-occupied subdomains within the \textit{de novo} centromeres. The CENH3-depleted subdomains were determined from the remaining CENH3-binding domains compared with control samples (see Methods). We identified a total of 279,475 nucleotides sequence in CENH3-occupied subdomains within the Dp3a \textit{de novo} centromere, which composed 79.85\% of the whole 350-kb region, and a total of 152,464 nucleotides sequence in CENH3-occupied subdomains within the Derivative 3-3 \textit{de novo} centromere, which occupies 52.94\% of the 288-kb region.

To further characterize the \textit{de novo} centromeres, we analyzed the proportion of DNA sequence compositions of CENH3 binding domains between the two regions using the Repeatmasker tool (http://www.repeatmasker.org; version open-4.0.5). No preference for GC content level was detected between the \textit{de novo} centromeric regions and other genomic regions (Figure 1a, c, Figure S1 and Table 1). Long terminal repeat (LTR) retrotransposons have been shown to constitute more than ~70\% of the maize genome with 25.79\% LTR/Copia and 47.44\% LTR/Gypsy (Table 1) (Schnable \textit{et al.} 2009). The native centromeres show a concentration for LTR/Gypsy retrotransposons (~70\%) and depletion for LTR/Copia (~9\%) compared with the genome-wide and chromosome-wide proportions (Table 1). The overall sequence compositions of Dp3a and Derivative 3-3 \textit{de novo} centromeric regions were with ~29\% LTR/Copia and ~35\% LTR/Gypsy, which was quite different from the compositions of native centromeres (Table 1). The LTR/Gypsy retrotransposons show a higher level in CENH3-occupied subdomains than in CENH3-depleted subdomains, and the simple repeats are concentrated only in CENH3-depleted subdomains both in the two regions (Figure S1). However, other DNA compositions, including the LTR/Copia, LINEs, and DNA
transposons, are quite different for the two regions (Figure S1). These results suggest that particular types of DNA sequences may be present in *de novo* centromeric regions with CENH3 binding.

**Gene distribution and expression of the two chromosomal regions before centromere formation**

We compared the gene distributions in the two chromosomal regions. The genes located at the centromere edge (less than 3 kb to the ends of centromere boundaries) were not included as they may represent pericentromeric regions. Six genes in Dp3a and no gene in Derivative 3-3 were located or partially located within the CENH3-occupied subdomains (Table S1 and Table S2). To investigate the gene expression level of the two chromosomal regions before *de novo* centromere formation, we conducted RNA-seq using seedling and young leaf tissues in plants without minichromosomes (Tables S3). Generally, the transcription level of the two chromosomal regions exhibited a higher or a moderate level compared to the average expression level of 1000 randomly selected regions (350-kb: Figure 2a for seedling tissues; Figure S2a for young leaf tissues; 288-kb: Figure 2b for seedling tissues; Figure S2b for young leaf tissues). For the 350-kb region, the genes with no expression were located in the CENH3-occupied subdomains (Table S1). However, most genes which displayed expression were located in the CENH3-depleted subdomains both for Dp3a and Derivative 3-3 (Table S1 and Table S2, defined as RPKM >= 1 in control samples at least for one tissue; GRMZM2G045318, GRMZM2G045275 in the 350-kb region;
GRMZM2G465046, GRMZM2G164072; GRMZM2G465053; GRMZM2G094586; GRMZM2G081239 in the 288-kb region). These results indicate that transcriptional actively regions can initiate de novo centromere formation and CENH3 seeding shows a preference for gene-free regions or regions with no gene expression.

There are two genes in both side boundaries of the Dp3a de novo centromere (GRMZM2G074097 and GRMZM2G081816, 1 and 12 designations in Figure 1b) and one long gene in the left boundary of Derivative 3-3 de novo centromere (GRMZM2G435294, 10.042 kb; asterisk in Figure 1d). Genes located at the edge of the 350-kb region show very high expression level before centromere formation (Figure 2c). Long or highly expressed genes located in the boundary of centromeres restrict the expansion of centromere size (Wang et al. 2014). Other mechanisms that regulate the centromere size may exist because neither long genes nor highly expressed genes are located in the right border of the 288-kb region (Figure 2d).

**Original DNA methylation of the two chromosomal regions before centromere formation**

We sought to analyze the DNA methylation of the two regions to characterize the original chromatin environment before centromere formation. Here we used the genome-wide cytosine methylation profile of B73 (Regulski et al. 2013) and another two maize inbred lines Ye 478 and Zheng 58 (Zhang et al. 2014) to characterize the methylation status before de novo centromere formation. Firstly, we divided the genome-wide DNA methylation profile
into 100-kb windows and counted the DNA methylation levels of each window. The original total DNA methylation level of the two regions on chromosome 3 and 9 was lower than the average distributed methylation level throughout the whole chromosome 3 and 9 (Figure 3a, e and Figure S3). Separate pathways exist to maintain DNA methylation depending on the sequence contexts (CG, CHG, and CHH; where H is A, C or T). The original methylation of the two regions on CG, CHG contexts exhibited the same tendency except on the CHH context for B73 (Figure 3a, e). The original methylation of the two regions on CG, CHG and CHH contexts exhibited the lower level than the average distribution level in Ye 478 and Zheng 58 (Figure S3).

To further examine the DNA methylation pattern within the two chromosomal regions, we divided the 350-kb, 288-kb region and 50-kb of flanking sequence on both sides into 3-kb windows and counted the DNA methylation level with each window from the B73 genome-wide DNA methylation profile. We compared the patterns between original DNA methylation level profile and CENH3-ChIP-seq reads distribution profile along the two regions. The total DNA methylation level along the original two regions show no correlation with CENH3 enrichment (Dp3a: $R^2 = 0.048$; Derivative 3-3: $R^2 = 0.002$) (Figure 3b, f). The methylation levels on different sequence contexts also show no correlation with CENH3 binding (Dp3a: CG $R^2 = 0.078$, CHG $R^2 = 0.128$, CHH $R^2 = 0.009$; Derivative 3-3: CG $R^2 = 0.006$, CHG $R^2 = 0.0116$, CHH $R^2 = 0.0004$) (Figure 3c, g). These results indicated the original fluctuation of DNA hypermethylated with hypomethylated state in each context was not completely correlated with de novo centromeric chromatin states. A decreased tendency of methylation level was observed near the boundary of centromeric regions (Figure 4b, c, f, g)}
and g). The DNA hypomethylation level on the original chromatin may also be one of mechanisms that restrict centromere size.

**Transcription of genes in the regions after de novo centromere formations**

To investigate whether the genes within the two *de novo* centromeric regions were expressed or not, we conducted RNA-seq using seedling and young leaf tissues in plants with *de novo* centromeres, respectively (Tables S3).

In seedling and young leaf tissues, the overall distribution of gene expression in Dp3a was highly coincident with the distribution in the control (Figure S2c. left panels, seedling tissues with and without Dp3a, $R^2 = 0.9158$, $p < 2.2e$-16; right panels, young leaf tissues with and without Dp3a, $R^2 = 0.9471$, $p < 2.2e$-16). This was the same for the gene expression profile in samples with and without Derivative 3-3 (Figure S2d. left panels, seedling tissues with and without Derivative 3-3, $R^2 = 0.9335$, $p < 2.2e$-16; right panels, young leaf tissues with and without Derivative 3-3, $R^2 = 0.8932$, $p < 2.2e$-16). The high correlation of gene transcription patterns between samples with and without *de novo* centromere indicates that the emergence of *de novo* centromere formation does not result in large-scale genome wide changes in gene expression (Figure S2e, f).

The transcription of the 350-kb region in Dp3a exhibited a slightly higher than average expression level of randomly selected 1000 350-kb regions in control and Dp3a (Figure 2a). The transcription of the 288-kb region in Derivative 3-3 exhibited a slightly lower than average expression level of randomly selected 1000 288-kb regions in control and Derivative.
To examine the expression level of the specific genes within the de novo centromeres, we conducted allele counts of each genotype distinguished by SNP sites instead of the RPKM value of each gene. We identified 33 SNPs sites from seedling RNA-seq data set between the control and Dp3a (Table S4), which belong to four genes expressed in the 350-kb region (Figure 2c), and 7 SNPs sites from the leaf RNA-seq data set between the control and Derivative 3-3 (Table S4) were detected, which belong to three genes that were expressed in the 288-kb region (Figure 2d). These results indicate that the genes located in the de novo centromeric regions were with a weak expression. (There may be transcripts in other tissues of other genes in the de novo centromeric regions).

The breakpoints of Dp3a minichromosome were with known genetic markers SH2 (GRMZM2G429899) and ET1 (GRMZM2G157574) (Figure S4a, b). To determine whether the expression level was due to the de novo centromere formation or the additional copy of minichromosome, we compared the expression level of 246 genes in Dp3a and control samples within a 7.33-Mb region located between the two loci. The distribution of 112 expressed genes (FPKM >= 1) in Dp3a showed a moderate increase compared with the control (Figure S4c). Only one gene (GRMZM2G000593) had a differential expression (Figure S4d). The ratiodistribution analysis of Dp3a/control of all expressed endogenous genes was performed (Sun et al. 2013). These ratios were plotted in a distribution with bins of 0.2 increments. The trends of gene expression on different chromosomes could be illustrated from this type of analysis with gene numbers present in the major bins. A major peak around a ratio of 1.2, with a minor peak around a ratio of 1.6 (Figure 4; top) was found with genes on Dp3a-minichromosome. A normal distribution around a ratio of 1.0 was found.
for genes on chromosome 3 not included in Dp3a-minichromosome (Figure 4; middle) and other chromosomes (Figure 4; bottom) as a control. These results indicate the expression level was due to the de novo centromere formation.

**Increased DNA methylation level of the euchromatic chromosomal regions after de novo centromere formation**

DNA methylation is an epigenetic mark for transcriptional silencing, which is vital for the maintenance of chromatin structure and genomic stability. We treated the CENH3-enriched DNA from plants containing Dp3a and Derivative 3-3 separately with sodium bisulfite and then performed Illumina sequencing to determine the DNA methylation level of centromeric regions. Genome-wide-bisulfite sequencing data from B73 and another two maize inbred lines Ye 478 and Zheng 58 were used as controls to determine the original DNA methylation level of these two regions. The maize native centromeric regions were determined by the CENH3-ChIP-seq data.

The DNA methylation level of native centromeres has no significant change between control (B73) and samples with neocentromeres (Figure 5a, e), indicating the method by genome-wide-bisulfite-seq and BisChIP-seq to determine DNA methylation levels of centromeric regions works effectively. The DNA methylation of the original chromosomal regions in the control were with a lower level compared to the native centromeric regions, which show no difference from random genomic regions (Figure 5a,e). The DNA methylation level of the Dp3a de novo centromeric region increased substantially compared
to the same region in the control, reaching a level similar to the native centromeric regions (P-value < 0.01, Figure 5a). There is also a significant increase of the DNA methylation level in the Derivative 3-3 de novo centromere compared to the same region in the control (P-value < 0.01, Figure 5e). The methylation level in the 288-kb region also reached the level of native centromeric regions (Figure 5e). The result was the same when we used Ye 478 and Zheng 58 as the controls (Figure S5), suggesting that the change of the DNA methylation level after de novo centromere formation unlikely result from the background or polymorphisms in different maize lines. It seems that the low methylation levels are insufficient to sustain the de novo centromeric chromatin.

We measured methylation values in the different DNA contexts, which shows the same increasing tendency in Dp3a compared to the control with significantly higher CG, CHG and CHH methylation level (P-value < 0.01, T-test, Figure 5b-d). The CG, CHG and CHH methylation level in Derivative 3-3 has the same tendency compared to the control (P-value < 0.01, T-test, Figure 5f-h). These results were also similar when compared to the Ye 478 and Zheng 58 controls (Figure S5), both of which indicat that centromeric regions require relatively high methylation status, and the DNA methylation level of the chromosomal regions will rise after de novo centromere formation if the original level was low.
The dynamic change of methylation levels before and after de novo centromere formation in the CENH3-depleted and CENH3-occupied subdomains

The transcription of genes detected in the Dp3a and Derivative 3-3 de novo centromeres (Table S4) suggests that the methylation level is not constant with a hypermethylated state across the entire centromeric regions (Figure 3b, f). It is interesting to investigate the relationship between DNA methylation and gene transcription in the de novo centromeric regions.

To further examine the DNA methylation pattern of the two regions after de novo centromere formations, we counted the DNA methylation level determined by BisChIP-seq in 3-kb windows. The general hypermethylation of these two regions was observed after de novo centromere formation (Figure 3b, f). We compared the DNA methylation level and CENH3 distribution profile along the two regions after de novo centromere formation. The methylation levels on different sequence contexts also show no correlation with CENH3 binding after de novo centromere formation (Dp3a: total $R^2 = 0.003$, CG $R^2 = 0.051$, CHG $R^2 = 0.036$, CHH $R^2 = 0.034$; Derivative 3-3: total $R^2 = 0.024$, CG $R^2 = 0.168$, CHG $R^2 = 0.164$, CHH $R^2 = 0.0002$). The patterns of DNA methylation level on different contexts before and after de novo centromere formation in these two regions were not maintained (Figure 3. Dp3a: total $R^2 = 0.485$, CG $R^2 = 0.378$, CHG $R^2 = 0.370$, CHH $R^2 = 0.065$; Derivative 3-3: total $R^2 = 0.333$, CG $R^2 = 0.195$, CHG $R^2 = 0.229$, CHH $R^2 = 0.001$), indicating the DNA methylation distribution patterns altered after CENH3 seeding, and CHH methylation pattern changed substantially.
The DNA methylation levels through the two de novo centromeres show an overall hypermethylation with hypomethylation in some specific sites (Figure 3b, d and Figure 4f, h). The locations of the four expressed genes detected in Dp3a and three expressed genes detected in Derivative 3-3 (Table S2) were at the hypomethylated loci (Figure 4d, h and Figure S6). These results illustrate the intricate nature of centromeric chromatin as it uses the hypermethylation level to maintain a tight chromatin structure but allows transcription to occur in the hypomethylated pockets (Wong et al. 2006).

We counted the average methylation level within CENH3-occupied and CENH3-depleted subdomains in the two regions, respectively. The total DNA methylation level, and the level on CG, CHG contexts of CENH3-occupied subdomains were significant higher than the level of CENH3-depleted subdomains before Dp3a and Derivative 3-3 de novo centromere formation (P-value <0.05, T-test. Figure 6), indicating that de novo centromeres are initiated by loading CENH3 within normally hypermethylated regions. However, after de novo centromere formation, the DNA methylation levels on CENH3-occupied and CENH3-depleted subdomains show totally different trends for Dp3a and Derivative 3-3 on different contexts (Figure 6). This may be the result of different cases between Dp3a and Derivative 3-3 de novo centromere. The dynamic change of methylation levels before and after de novo centromere formation on different sequence contexts indicates that the original chromatin environment can affect the CENH3 loading, such as to restrict centromere expansion and gene expression, and the DNA methylation patterns can also be shaped after CENH3 loading.
Discussion

The distinct chromatin organization, epigenetic factors, centromere-associated proteins and histone modifications as well as histone variants that determine centromere identity have been deciphered in recent decades in various species (Allshire and Karpen 2008, Black and Cleveland 2011, Fukagawa and Earnshaw 2014, Marshall et al. 2008, Verdaasdonk and Bloom 2011). However, the mechanism that establishes and maintains the centromeric chromatin is largely unknown due to the nature of the highly repetitive DNA underlying it. The dynamic alteration of chromatin status before and after de novo centromere formation caused by CENH3 seeding can contribute to the changes of centromeric chromatin state with respect to DNA methylation level and gene transcription. Here we show that the de novo centromeric chromatin requires a relatively high methylation level to maintain the centromeric chromatin environment. This suggests that the methylation level of de novo centromeric regions adjusts if the original methylation level is too low (Figure 5 and Figure S5); if the primordial methylation level of de novo centromeric region is high enough to sustain the requirement of centromeric chromatin, the methylation level will hardly change, as illustrated by the case of the sDic15 de novo centromere (Zhang et al. 2013). The observed increase in DNA methylation at the de novo centromere could provide the basis for centromeric chromatin states that is essential for the epigenetic identity and maintenance of the centromere (Schueler and Sullivan 2006).

De novo centromeres are found in various species such as fungi, animals, and plants. However, where and how de novo centromeres form is still a key underdetermined issue. Some de novo centromeres can form near the native centromeres in Candida albicans (Ketel et al. 2011).
et al. 2009) and Drosophila melanogaster (Maggert and Karpen 2001); some occur at the heterochromatin boundaries (Heun et al. 2006, Olszak et al. 2011); and others occur in the telomeric heterochromatin (Ishii et al. 2008). Collectively, our studies showed that de novo centromeres in maize can occur in gene poor regions (Wang et al. 2014) or gene rich regions, of which have also been documented for de novo centromeres in human and chicken cells (Marshall et al. 2008, Shang et al. 2013). The difference between the two kinds of regions was consistent with the change of DNA methylation level before and after de novo centromere formation. We found that CENH3 seedling has a preference for gene-free regions or on regions with very low gene expression (Table S1, S2). Newly formed CENP-A domains are established in a pattern inverse to regions that are transcribed in the germline and early embryo in Caenorhabditis elegans (Gassmann et al. 2012). This may be the reason that de novo centromere formation in plants, D. melanogaster and fungi has a preference for heterochromatic regions (Ishii et al. 2008, Ketel et al. 2009, Murphy and Karpen 1995, Nasuda et al. 2005, Topp et al. 2009, Wang et al. 2014, Zhang et al. 2013).

The centromere size varies in different species and even in different chromosomes in one species, but the mechanism regulating centromere boundaries is largely unknown. The analysis of centromere size in ten grass species reveals a strong correlation with genome size (Zhang and Dawe 2012). A particularly active chromatin state with gene transcription may be a barrier for CENH3 nucleosome seeding. Long or highly expressed genes may restrict centromere size (Wang et al. 2014). A decreased tendency of methylation was observed at the boundary of pericentromeric and centromeric regions (Figure 3) and no genes were expressed

This article is protected by copyright. All rights reserved.
on the right side of Derivative 3-3 (Figure 1c), indicating that DNA methylation may also function to regulate centromere size.

Our results show that the *de novo* centromeric chromatin is composed of hypermethylated interspersed with hypomethylated regions in which transcription occurs in the hypomethylated pockets (Figure 3d, h and Figure S6). Transcription from centromeric genes and repeat sequences has been reported in many species (Chan and Wong 2012, Chueh *et al.* 2009, Nagaki *et al.* 2004, Saffery *et al.* 2003, Topp *et al.* 2004), and emerging evidence shows a role for centromeric transcription in CENH3 deposition, kinetochore assembly and cell division (Chen *et al.* 2015, Quenet and Dalal 2014, Rosic *et al.* 2014). Here we showed that expressed genes exist in the boundary of the *de novo* centromeres, but there is low transcription inside the *de novo* centromeres (Figure 2). The documented transcripts within the *de novo* centromeric chromatin (Figure 2) indicate that the transcriptional machinery operate in the centromere. The questions remaining unanswered are the roles of the transcription in centromere formation and why some genes are expressed and others are not.

The gene functions underlying native centromeres are mostly unknown. The expressed genes in the two *de novo* centromeres are annotated with different functions or as uncharacterized proteins. These results suggest that the role of gene transcriptions in centromeres may be not for the production of transcription, but just for the process of transcription, and the transcripts are only the byproducts of the chromatin reorganization during CENH3 incorporation.

DNA methylation state is an epigenetic mark for ‘open’ or ‘closed’ chromatin that is essential for chromatin maintenance and genomic stability. An overall hypermethylated state surrounds the native centromeres but the detailed pattern of methylation level has remained.
poorly understood, mainly because of the repetitive nature of the centromere. Several observations suggest a role for positioned nucleosomes in localizing DNA methylation (Huff and Zilberman 2014, Pennings et al. 2005, Perez et al. 2012). However, no correlation was observed between the methylation levels on different contexts with CENH3 enrichment subdomains (Figure 3). The dynamic changes of DNA methylation patterns indicate that CENH3 seeding was affected by intrinsic DNA methylation patterns before neocentromere formation and the CENH3 loadings can also shape the DNA methylation patterns after de novo centromere formation. Our work reveals the roles of chromatin and DNA during de novo centromere formation and the dynamic changes of the DNA methylation and transcription before and after de novo centromere formation, which may promote further analysis in specific chromatin character in centromere formation.

**Experimental procedures**

**Plant materials**

Dp3a kernels were kindly supplied by Pat Schnable (Iowa State University) and the Maize Genetics Cooperation Stock Center (University of Illinois). All the materials used for cytogenetic analysis and next generation sequencing experiments were planted in the greenhouse and the field. The seedling and young leaf tissues were used for RNA-seq experiments. The DNA probe preparation and fluorescence in situ hybridization (FISH) on root tip cells were performed according to our previously works (Zhang et al. 2013).
RNA-seq and RNA-seq data analysis

Total RNA was extracted using Trizol reagent (Invitrogen). RNA libraries were prepared according to the standard Illumina TruSeq RNA library kits. The selected polyA RNA was subjected to the HiSeq2000 platform to generate paired-end 101-nt reads. Seedling and young leaf tissues harvested from plant materials with and without minichromosome Dp3a and Derivative 3-3 were used for RNA-seq analysis.

The raw reads from RNA-seq data were analyzed for quality control using FastQC (Andrews 2010). The adapter and low quality sequences of the paired end reads were treated using Trimmmomatic (v0.30, Bolger et al. 2014). The cleaned reads were mapped to the B73 RefGen_v3 genome reference using TopHat2 (v2.0.12) with the default parameters (Trapnell et al. 2012). The following procedures and steps were combined with cufflinks and cuffdiff for differential expression analysis on different samples (Trapnell et al. 2012). The SNPs were detected from the RNA-seq data. Data processing and figures were produced using Perl and R scripts.

Identification of CENH3-occupied and CENH3-depleted subdomains in the de novo centromeric regions

To identify the CENH3-occupied subdomains in the de novo centromeric regions, the MACS (Feng et al. 2012) were used to call CENH3-binding peaks following parameters with P value \(< 1.00e-05\) and genome size, \(g = 2.3\). The materials without minichromosomes were used as a control for calling CENH3-occupied peaks. The method to identify CENH3-depleted subdomains was modified from Zhao et al (Zhao et al. 2015). The
centromere edges (less than 3 kb to the ends of centromere boundaries) were not included as it may represent pericentromeric regions. The remaining regions of the de novo centromere were the candidates for CENH3-depleted subdomains. The uniquely-mapped read counts from ChIP were tested against input read counts for each subdomain. Subdomains with p-value less than 0.01 were regarded as CENH3-depleted subdomains.

**Analysis of DNA methylation read alignment and methylation calling**

Low quality bases were filtered according to previously described methods. The cleaned bisulfite sequencing reads were aligned to the B73 RefGen_v3 genome reference by BS-Seeker2 (Guo et al. 2013) with parameters “-m 3 --aligner=bowtie2 --bt2”, and methylation was called. Cytosine sites covered by more than three reads were used for subsequent analyses. The mean DNA methylation level of selected regions was calculated separately for “CG”, “CHG” and “CHH”.

**Accession numbers**

The BisChIP-seq data sets and RNA-seq data sets have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession number GSE64641, GSE76533. The accession numbers of anti-CENH3 ChIP data are GSM1057274, GSM1057275 and GSE59124. The input data for anti-CENH3 ChIP-seq was from SRR1652619. The control data of
genome-wide DNA methylation profile was from the whole-genome bisulfite sequencing of different maize inbred lines (SRP011933, SRP011991).

**Acknowledgments**

We thank Dr. Gernot Presting (University of Hawaii) for critically reading the manuscript, and providing helpful suggestions and comments. This work was supported by the National Natural Science Foundation of China (31320103912) and US National Science Foundation (ISO1445144).

The authors declare no conflict of interest.

**Supporting Information Legends**

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

**Figure S1.** Comparison between CENH3-depleted and CENH3-occupied subdomains within de novo centromeric regions.

**Figure S2.** RNA-seq analyses of Dp3a and Derivative 3-3 mini-chromosomes.

**Figure S3.** Original chromatin environment of the de novo centromeric regions before centromere formation.

**Figure S4.** Gene expression analysis on the Dp3a minichromosome after de novo centromere formation.

This article is protected by copyright. All rights reserved.
**Figure S5.** Comparison of DNA methylation levels between the *de novo* centromeric and other genomic regions on *de novo* centromeres and controls (inbred lines Zheng 58 and Ye478).

**Figure S6.** DNA methylation level on the expressed genes after *de novo* centromere formation.

**Table S1.** Expression level of 12 genes annotated within the 350-kb *de novo* centromeric region between samples with and without Dp3a.

**Table S2.** Expression level of 7 genes annotated within the 288-kb *de novo* centromeric region between samples with and without Derivate 3-3.

**Table S3.** Statistics of the mapping results of RNA-Seq.

**Table S4.** SNPs information detected from RNA-seq data within two *de novo* centromeric regions on chromosome 3 and chromosome 9.

**Reference**


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


Table 1. DNA sequence composition of de novo centromeres, native centromeres, chromosomes, and genome-wide

<table>
<thead>
<tr>
<th></th>
<th>Dp3a</th>
<th>Derivate 3-3</th>
<th>Cen2</th>
<th>Cen5</th>
<th>Chr3</th>
<th>Chr9</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC level</td>
<td>47.49%</td>
<td>44.97%</td>
<td>45.92%</td>
<td>45.21%</td>
<td>46.88%</td>
<td>46.98%</td>
<td>46.89%</td>
</tr>
<tr>
<td>Retroelements</td>
<td>69.88%</td>
<td>63.52%</td>
<td>84.73%</td>
<td>83.80%</td>
<td>75.00%</td>
<td>73.58%</td>
<td>74.19%</td>
</tr>
<tr>
<td>SINEs</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.02%</td>
<td>0.02%</td>
</tr>
<tr>
<td>LINES</td>
<td>2.85%</td>
<td>0.99%</td>
<td>0.48%</td>
<td>0.49%</td>
<td>0.90%</td>
<td>0.82%</td>
<td>0.86%</td>
</tr>
<tr>
<td>RTE/Bov-B</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.07%</td>
<td>0.00%</td>
<td>0.11%</td>
<td>0.11%</td>
<td>0.10%</td>
</tr>
<tr>
<td>L1/CIN4</td>
<td>2.85%</td>
<td>0.99%</td>
<td>0.41%</td>
<td>0.49%</td>
<td>0.79%</td>
<td>0.71%</td>
<td>0.76%</td>
</tr>
<tr>
<td>LTR elements</td>
<td>67.03%</td>
<td>62.53%</td>
<td>84.25%</td>
<td>83.30%</td>
<td>74.09%</td>
<td>72.75%</td>
<td>73.31%</td>
</tr>
<tr>
<td>Ty1/Copia</td>
<td>29.91%</td>
<td>28.52%</td>
<td>9.03%</td>
<td>9.07%</td>
<td>25.77%</td>
<td>25.09%</td>
<td>25.79%</td>
</tr>
<tr>
<td>Gypsy/DIRS1</td>
<td>37.12%</td>
<td>33.51%</td>
<td>75.14%</td>
<td>74.10%</td>
<td>48.24%</td>
<td>47.57%</td>
<td>47.44%</td>
</tr>
<tr>
<td>DNA transposon</td>
<td>2.72%</td>
<td>6.92%</td>
<td>1.18%</td>
<td>3.65%</td>
<td>4.07%</td>
<td>4.35%</td>
<td>4.22%</td>
</tr>
<tr>
<td>hobo-Activator</td>
<td>0.44%</td>
<td>0.44%</td>
<td>0.04%</td>
<td>0.13%</td>
<td>0.78%</td>
<td>0.73%</td>
<td>0.73%</td>
</tr>
<tr>
<td>Tc1-IS630-Pogo</td>
<td>0.09%</td>
<td>0.23%</td>
<td>0.02%</td>
<td>0.00%</td>
<td>0.07%</td>
<td>0.07%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Tourist/Harbinger</td>
<td>0.29%</td>
<td>0.38%</td>
<td>0.10%</td>
<td>0.04%</td>
<td>0.49%</td>
<td>0.50%</td>
<td>0.48%</td>
</tr>
<tr>
<td>Small RNA</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.10%</td>
<td>0.10%</td>
<td>0.15%</td>
</tr>
<tr>
<td>Satellites</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.08%</td>
<td>1.78%</td>
<td>0.10%</td>
<td>0.30%</td>
<td>0.25%</td>
</tr>
<tr>
<td>Simple repeats</td>
<td>0.30%</td>
<td>0.34%</td>
<td>0.20%</td>
<td>0.17%</td>
<td>0.31%</td>
<td>0.31%</td>
<td>0.31%</td>
</tr>
<tr>
<td>Low complexity</td>
<td>0.02%</td>
<td>0.02%</td>
<td>0.02%</td>
<td>0.06%</td>
<td>0.06%</td>
<td>0.05%</td>
<td>0.06%</td>
</tr>
</tbody>
</table>

RepeatMasker version open-4.0.5, default mode.
(http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker).

This article is protected by copyright. All rights reserved.
Figure legends

Figure 1. Two de novo centromeres form in 350-kb and 288-kb regions on the distal arms of chromosomes 3 and 9 with high-density gene distribution.

(a) Annotated gene density (non-TE gene: blue; TE-related gene: red) and GC content (green) distribution of every 100-kb window along maize chromosome 3. The left Y-axis: gene density. The right Y-axis: GC level. The purple border represents the native centromere of chromosome 3. The magenta border represents the 350-kb Dp3a de novo centromere (218,917, 000-219,267,000). (b) Mapped reads from ChIP-seq reads with CENH3 antibodies along the 350-kb de novo centromere between Dp3a and control. Each vertical blue bar in the first track represents the log₂ (Dp3a_{RPM}/Control_{RPM}) ratio (RPM: reads per million) in 0.1-kb windows of the 350-kb region. The second track illustrates the 12 genes annotated along the region (Table S1). The magenta and green arrows show the direction of gene transcription. The third track shows the distribution of different repeat elements along the 350-kb region.

(c) Gene density and GC content distribution along maize chromosome 9. The purple border represents the native centromere of chromosome 9. The magenta border represents the Derivative 3-3 de novo centromere (4,300,000-4,588,000). (d) Mapped reads from ChIP-seq reads with CENH3 antibodies along 288-kb de novo centromere between Derivative 3-3 and control. The second track illustrates the 7 genes annotated along the region (Table S2). The asterisk represents the gene GRMZM2G435294. The third track shows the distribution of different repeat elements within the 288-kb region.
**Figure 2. Analysis of gene expression within the 350-kb and 288-kb regions before and after de novo centromere formation.**

(a) Boxplot shows the comparison of the average gene expression level in the 350-kb region and randomly-selected 1000 contiguous regions between seedlings of Dp3a and control. The Y-axis represents the average gene expression level using log$_2$(RPKM) value (RPKM: reads per kilobase per million reads). The X-axis represents Rand_Ctr (Random Control), Rand_Dp3a (Random Dp3a), Ctrl and Dp3a in 350-kb region, respectively. (b) Boxplot shows the comparison of the average gene expression level in the 288-kb region and randomly-selected 1000 contiguous regions between seedlings of Derivative 3-3 and control. The X-axis represents Rand_Ctr (Random Control), Rand_3-3 (Random Derivative 3-3), Ctrl and 3-3 (Derivative 3-3) in the 288-kb region, respectively. (c) The detailed expression levels of 12 genes with about 1 Mb of flanking genes on both sides in the control and Dp3a. The region between the two dash lines shows the 350-kb region. (d) The detailed expression levels of 7 genes with 2 flanking genes on both sides in control and Derivative 3-3. The region between the two dash lines shows the 288-kb region.

**Figure 3. The distribution pattern of DNA methylation level before and after de novo centromere formations.**

(a,e) The boxplot displays the distribution of original DNA methylation level (B73) on different contents along chromosome 3 (a) and chromosome 9 (e). The blue line shows the DNA methylation level of the 350-kb Dp3a (a) and the 288-kb Derivative 3-3 de novo centromeric region (e).
(b,f) The distribution of total DNA methylation pattern with CENH3-ChIP-seq read profile before and after Dp3a (b) and Derivative 3-3 (f) de novo centromere formation along the two regions with 50 kb flanking on both sides. The red line represents the “pre-novel centromere” state. The blue line represents “post-novel centromere” state. The green chart shows the CENH3-enriched profile along the regions. The left-Y-axis represents the DNA methylation level. The right-Y-axis represents the ChIP-seq reads counts.

(c-d) The distribution of methylation pattern on different DNA contexts for control (c) and Dp3a (d) along the 350-kb region. The 4 solid blue boxes represent the expressed genes detected in Table S4. The left-Y-axis represents the CG and CHG methylation level. The right-Y-axis represents the CHH methylation level.

(g-h) The distribution of methylation pattern on different DNA contexts for control (g) and Derivative 3-3 (h) along the 288-kb region. The 3 solid blue boxes represent the expressed genes detected in Table S4.

**Figure 4. Distributions of Dp3a/control ratios of global endogenous gene expressions.**

The global expression patterns were obtained by RNA-seq and analyzed by generating the ratio distributions from Dp3a/control of gene quantities on Dp3a minichromosome (top), chr3 not including Dp3a minichromosome (middle) and other chromosomes (below). The ratio 1.00 means no change and is designated with a red line.
Figure 5. Comparison of DNA methylation levels between the de novo centromeric and other genomic regions on de novo centromeres and control (B73).

(a-d) The average DNA methylation level on Total (a) CG (b), CHG (c) and CHH (d) in different regions between Dp3a and control. The Y-axis represents the DNA methylation level. Color bars represent the different genomic regions, including the de novo 350-kb centromere on Dp3a (Dp3a-350 kb) and 350-kb on control (Control-350 kb), randomly selected 1000 genomic regions (Control-Random) and 10 native centromeric regions (Control-Cen, Dp3a-Cen).

(e-h) The average DNA methylation level on Total (e) CG (f), CHG (g) and CHH (h) in different regions between Derivative 3-3 and control. Color bars represent the different regions, including the de novo 288-kb centromere on Derivative 3-3 (3-3-288 kb) and 288-kb on control (Control-288 kb), randomly selected 1000 genomic regions (Control-Random) and 10 native centromeric regions (Control-Cen, 3-3-Cen).

Datum shows means ± standard errors (SE). T-test was performed for the comparison. ** P-value <0.01.

Figure 6. Dynamic change of DNA methylation levels before and after de novo centromere formations on different sequence contexts.

(a) The dynamic change of methylation levels before and after Dp3a de novo centromere formation on different sequence contexts for CENH3-occupied and CENH3-depleted subdomains. Color bars represent the DNA methylation levels of CENH3-occupied (blue) and CENH3-depleted (red) regions from the control, and the DNA methylation levels of
CENH3-occupied (green) and CENH3-depleted (purple) regions from Dp3a. (b) The dynamic change of methylation levels before and after Derivative 3-3 de novo centromere formations on different sequence contexts for CENH3-occupied and CENH3-depleted subdomains. Datum shows means ± standard errors (SE). T-test was performed for the comparison. * P-value <0.05. ** P-value <0.01.