Stable patterns of CENH3 occupancy through maize lineages containing genetically similar centromeres

AUTHORS
Jonathan I Gent1*, Kai Wang2*, Jiming Jiang2, and R Kelly Dawe1,3

AUTHOR AFFILIATIONS
1Department of Plant Biology, University of Georgia, Athens, GA 30602
1Department of Horticulture, University of Wisconsin-Madison, Madison, Wisconsin 53706
3Department of Genetics, University of Georgia, Athens, GA 30602
*equal contributors

CORRESPONDING AUTHOR
R. Kelly Dawe
2502 Miller Plant Sciences
University of Georgia, Athens GA 30602
kdawe@uga.edu

RUNNING TITLE:
Epigenetic stability and genetic diversity of maize centromeres
ABSTRACT

While the approximate chromosomal position of centromeres has been identified in many species, little is known about the dynamics and diversity of centromeres positions within species. Multiple lines of evidence indicate that DNA sequence has little or no impact in specifying centromeres in maize and in most multicellular organisms. Given that epigenetically-defined boundaries are expected to be dynamic, we hypothesized that centromere positions would change rapidly over time, which would result in a diversity of centromere positions in isolated populations. To test this hypothesis, we used CENP-A/cenH3 (CENH3 in maize) chromatin immunoprecipitation to define centromeres in breeding pedigrees that included the B73 inbred as a common parent. While we found a diversity of CENH3 profiles for centromeres with divergent sequences that were not inherited from B73, the CENH3 profiles from centromeres that were inherited from B73 were indistinguishable from each other. We propose that specific genetic elements in the region of centromeres favor or inhibit CENH3 accumulation, leading to reproducible patterns of CENH3 occupancy. These data also indicate that dramatic shifts in centromere position normally originate from accumulated or large-scale genetic changes rather than from epigenetic positional drift.
INTRODUCTION

Since the discovery of sequence-specific binding of the lambda phage repressor to operator sites (Gilbert and Müller-Hill 1967; Ptashne 1967), the principle of sequence-specific binding has been a key to understanding gene regulation, DNA replication, genome defense, and other basic genetic processes. It has also become clear that the linear sequence of nucleic acid bases is not the only mediator of protein recognition, but also higher-order structure of DNA and its associated molecules and chemical modifications. In contrast, the centromeres of complex eukaryotes remain a mystery, as they appear to function largely independent of DNA sequence (For review, see Fukagawa and Earnshaw 2014). Theoretically defined, centromeres are the part of chromosomes upon which kinetochores are assembled to guide chromosomes through cell division, but in practice they are defined as the large chromosomal domains occupied by a histone H3 variant called cenH3 (Talbert et al. 2012) or CENP-A (Earnshaw et al. 2013). Experiments done primarily in metazoans have revealed that the presence of cenH3 is key not only for replenishment of itself, but also for recruitment of other conserved centromere proteins such as CENP-C and CENP-T (Kato et al. 2013; Folco et al. 2015). While certain centromere/kinetochore protein domains including the histone fold domain of cenH3 are remarkably conserved across eukaryotes, centromere DNA sequences are in many cases highly divergent (Melters et al. 2013), and sometimes not shared across the chromosomes of an individual species (Gong et al. 2012), or, in an extreme case, not even shared by the same centromeres from the homologous chromosomes (Wang et al. 2014b).
Deposition of cenH3 is replication-independent (FUKAGAWA and EARNSHAW 2014) and the mechanisms that guide cenH3 to the same location at each cell cycle are not known. Aside from single-celled budding yeasts, no eukaryotes are known to have strict, sequence-specific signals for the recruitment of cenH3. Humans are sometimes considered an exception because they (and some other primates) have a repeated binding site for the protein CENP-B in their centromeres that can contribute to cenH3 recruitment and centromeric chromatin structure (OHZEKI et al. 2002; FACHINETTI et al. 2013; HENIKOFF et al. 2015). However CENP-B is not required for viability in mice and apparently acts in concert with other, sequence-independent feedback mechanisms (HUDSON et al. 1998). Other evidence suggests that centromere size is dictated by cellular environment rather than DNA sequence, as indicated by measurements of centromere size in various grass species, including maize centromeres that have been transferred into oat cells (ZHANG and DAVE 2012; WANG et al. 2014a). Similarly, in human cells the number of CENP-A molecules at centromeres is variable and directly proportional to the number of CENP-A proteins in the cell, suggesting that centromeres are assembled by a mass-action mechanism (BODOR et al. 2014). Perhaps the most convincing demonstrations that centromeres are not dependent on particular DNA sequences (or CENP-B in mammals) come from the many examples of neocentromeres, which are large-scale changes in centromere position that have been documented in numerous species (FUKAGAWA and EARNSHAW 2014). These observations also raise the possibility of more subtle and continuous drifting of centromeres relative to their underlying chromosomal positions. One might predict such centromere drift over organismal generations because of zygotic removal and replacement of cenH3 (observed in Arabidopsis (INGOUFF et al. 2010)) and over cellular generations because cenH3 deposition
is replication independent and diluted from centromeres at every cell division (Fukagawa and Earnshaw 2014). Likewise, the fact that cenH3 recruitment is self-propagating makes for a theoretically unstable system in which incremental small variations in cenH3 deposition could build on each other over time to produce large changes. The potential for positional drift is supported by observations of horse chromosome 11, which is unusual in that it lacks tandem repeats. In five samples of cultured horse fibroblast cell lines, centromere 11 occupied a unique position in each one, with locations differing by as much as 200 Kb (Purgato et al. 2014).

In multicellular eukaryotes, the common chromosomal organization involves one centromere per chromosome that consists of up to a Mb-scale region that is enriched for tandem repeats (also called satellites) and retrotransposons (Jin et al. 2004; Ferreri et al. 2011; Neumann et al. 2011; Melters et al. 2013). The repetitive nature of these DNA elements makes assembly of centromeric reference sequences challenging with current sequencing technology. In addition, mapping short sequence reads to the existing reference sequences is complicated by the inability to discern which of multiple possible repetitive loci a read corresponds to. Hence, mapping cenH3 footprints across native centromeres using chromatin immunoprecipitation followed by sequencing (ChIP-seq) has been most easily carried out in fungi with small centromeres (Smith et al. 2011; Yao et al. 2013). However, the technique has also been applied successfully to maize, rice, and nematode centromeres (Yan et al. 2008; Wolfgruber et al. 2009; Steiner and Henikoff 2014). ChIP-seq has also been carried out in several human, chicken, and maize neocentromeres (Fu et al. 2013; Hasson et al. 2013; Shang et al. 2013; Zhang et al. 2013a; Liu et al. 2015). In maize, two of the ten centromeres were selected based on their low content of tandem repeats for
careful sequencing and assembly, and several of the other remaining centromeres have been partially assembled, enabling careful mapping of cenH3 positioning as well as other genomic features in and near centromeres (Wolfgruber et al. 2009; Gent et al. 2012).

Distributions of cenH3 in both native and neocentromeres typically show broad peaks of tens of Kb up to several Mb across (Yan et al. 2008; Wolfgruber et al. 2009; Smith et al. 2011; Gent et al. 2012; Fu et al. 2013; Hasson et al. 2013; Shang et al. 2013; Yao et al. 2013; Zhang et al. 2013a; Purgato et al. 2014; Liu et al. 2015). As these cenH3-ChIP-seq distributions represent averages from millions of cells, it is possible that the cenH3 profiles of individual cells differ from each other considerably. At one extreme, single centromeres could have a block or several blocks of cenH3 at a uniform density (cenH3 molecules per unit of DNA), and the position of these blocks could vary from cell to cell to produce the complex distribution of cenH3 revealed in the population average. This situation would be hard to reconcile with studies on plants and animals cells that reveal substantial non-uniformity in the density of cenH3 across individual centromeres as measured by immunofluorescence (Blower et al. 2002; Ferreri et al. 2011; Ishii et al. 2015), and that genic regions within rice centromeres appear to have less cenH3 than intergenic ones by ChIP-seq (Yan et al. 2008). Alternatively, each centromere could have a complex distribution of cenH3 similar to the population average. We hypothesized that the actual case would be something in between these two extremes, where individual centromeres have a complex distribution of cenH3, but the distribution varies between cells and produces a population average that is different than any individual cell. Although such variation would be difficult to see in a tissue extract representing thousands of somatic cells, in a crossing lineage, each parent contributes a single gamete such that minor
variations could be transmitted. This creates the potential for gradual “centromere drift” to be captured in isolated lineages (Figures 1 and 2).

To test whether diverse centromere locations are indeed present in maize populations and by inference whether such diversity can be attributed to gradual centromere drift, we utilized the high-quality centromere reference sequences as well as pedigrees of closely-related and genetically-characterized stocks of maize (NELSON et al. 2008; VAN HEERWAARDEN et al. 2012). Consistent with our expectations, we found great diversity in centromeres, as defined by location of maize cenH3 (CENH3). But contrary to our expectations, diversity in CENH3 location corresponded to genetic changes in centromeres rather than centromere drift. Furthermore, we found that the complex patterns of CENH3 within centromeres correlated with the presence of specific genetic elements, where genes were associated with low levels of CENH3 and centromeric repeats with high levels. These results suggest that despite the potential for abrupt, large-scale epigenetic shifts in centromere size or position, the normal situation involves reproducible stability, stability that is contributed to by centromeric genetic elements.

MATERIALS & METHODS

The sequence data produced for this study are available at the NCBI sequence read archive, accession number SRP049952.

CENH3 chromatin immunoprecipitation (ChIP)
We collected seedling leaves from 30 to 70 seedlings per inbred stock when the
seedlings were ~12 cm in height. We obtained all the B73-related inbreds stocks as well as
one of our B73 stocks (B73 “distant kin” in Table 2) from GRIN National Genetic Resources
Program (http://www.ars-grin.gov). The other B73 samples (B73 “reference” and “selfed
siblings” in Table 2) were from a B73 lineage that had been separated from the GRIN B73
for at least eight generations. We carried out native ChIP (NAGAKI et al. 2003) on these
samples using anti-CENH3 (maize) antibodies (ZHONG et al. 2002). Briefly, we extracted
chromatin from finely ground, frozen tissue, digested with Micrococcal nuclease (MNase)
to reduce the chromatin to mainly mononucleosomes, before immunoprecipitating to
select for CENH3 nucleosomes. We used Protein A sepharose to purify the
antibody/nucleosome/DNA complexes. Immediately after pre-cleaning and before
incubating with antibodies, a portion of the B73 reference sample was removed and set
aside as an input sample.

Preparation of MNase and Fragmentase control libraries

The library of total genomic nucleosomes used as a control for measuring CENH3
enrichment for all of the 14 CENH3 ChIP samples was prepared by MNase digestion of
isolated chromatin. Details of library preparation as well as raw reads are available at the
NCBI Sequence Read Archive, experiment SRX708840 (GENT et al. 2014). The Fragmentase
library used as a control for CENH3 enrichment was prepared using NEBNext dsDNA
Fragmentase (New England Biolabs, #M0348S) of naked DNA. Details of library
preparation as well as raw reads are available from the NCBI Sequence Read Archive,
experiment SRX708865 (GENT et al. 2014).
Preparation of ChIP-seq, MNase-seq, and Fragmentase-seq libraries

For all samples, we started with 100-300 ng of double-stranded, fragmented DNA and end-repaired and 5'-phosphorylated using the End Repair Module (New England Biolabs #E6050L), A-tailed with Klenow exo- (New England Biolabs #M0212S), and ligated to adapters (8.3 nM) with a Quick Ligase Kit (New England Biolabs, #M2200S. We PCR amplified using Phusion 2X HF master mix (New England Biolabs #M0531S) in a 50-ul total volume containing 200 nM of each primer. Thermocycler conditions were as follows: An initial denaturation at 98°C for 30 seconds; no more than 12 cycles of 98°C for 30 seconds, 65°C for 30 seconds, and 72°C for 30 seconds; and a final extension at 72°C for 5 minutes. We purified DNA between each enzymatic treatment using Agencourt AMPure XP at a ratio of 1.8:1, except after PCR, where the ratio was 0.9:1. Prior to sequencing, we selected amplicons corresponding ~150-bp inserts by gel electrophoresis on a 2.5% agarose gel and purified using a QIAquick gel extraction kit (Qiagen #28704). We used indexed PCR primers from the ScriptSeq Index PCR primers (Epicentre, #RSBC10948 and #SSIP1202) at a concentration of 200 nM. We used a Y adapter of the following two oligos at a concentration of 8.3 nM each: /5Phos/GATCGGAAGAGCACACGTCT and ACACTCTTTCCCTACACGACGCTCTTCCGATCT.

Analysis of Illumina reads

We aligned all reads to the maize genome version 3 (B73 RefGen v3; obtained from www.maizesequence.org) using the Burrows-Wheeler Aligner BWA-MEM (Li and Durbin 2009), default parameters. We used a combination of BEDTools (Quinlan and Hall 2010),
SAMtools (Li et al. 2009), and custom python3 scripts for converting between sequence read formats, calculating coverage, and calculating overlaps between genomic loci. In all cases, we defined uniquely-mapping reads as those with MAPQ scores of at least 20. In identifying reads that matched to specific genetic elements, we required at least a 50% overlap between each read’s genomic alignment and the annotated location of the element.

To find peaks of CENH3 enrichment across the genome and to measure the enrichment for each peak, we used HOMER findPeaks (default parameters accept the following: -region -size 5000 -minDist 40000 -F 10 -L 0 -C 0). We used the B73 (“reference”) CENH3 ChIP and total nucleosome control to define CENH3 enrichment. We required both ChIP and control reads to produce a MAPQ value of at least 20 to be included in this analysis.

**Identifying genomic locations of centromeric repeats**

To identify loci corresponding to CRM1-4 we used sequences provided by Gernot Presting, the same as used in our earlier nucleosome positioning work (GENT et al. 2011), and used Blastall to identify homologous regions in the genome (-e 1e-20). In cases where the alignments from multiple CRMs overlapped, we selected the one with the lowest E value. Similarly, to identify loci corresponding to CentC, we aligned a CentC consensus sequence (courtesy of Paul Bilinski and Jeffrey Ross-Ibarra) to the genome using the same method.

**Determining SNP identify between B73-related inbreds and B73**
Data on SNPs in B73 and related inbreds (van Heerwaarden et al. 2012) are available at http://figshare.com/articles/van_Heerwaarden_et_al_2012/757738 (van Heerwaarden et al. 2012). To transform the version 2 reference genome positions of the SNPs to reference version 3 positions, we took the ~100 bp centered on the SNP and aligned it to the genome using BWA-MEM (default parameters). We excluded SNP sequences that failed to yield an alignment or produced multiple/chimeric alignments. We divided the genome into 2.5-Mb bins and calculated the percent of SNPs in each bin that matched B73 for each inbred. We excluded SNPs in inbreds that yielded a “missing data” genotype (designated as “-”-”) from the calculations. When B73 itself had missing data for a SNP, we excluded the SNP from the calculation for all genotypes.

RESULTS

Differences in CENH3 distributions correlate with differences in centromere DNA sequence

Consistent with maize naming nomenclature and prior work, we refer to the maize cenH3/CENP-A protein as CENH3 (Zhong et al. 2002). To allow us to distinguish between centromere diversity that could be attributed to epigenetic rather than genetic change, we carried out ChIP-seq on B73 (the inbred stock whose genome provides the standard reference sequence (Russell 1972; Schnable et al. 2009)) and a set of ten inbred stocks that are closely related to B73: DJ7, LH74, LH119, LH132, LH149, LP5, NS701, PHG86, PHW17, and F42 (Nelson et al. 2008). These stocks were developed over many years in the 1970s and 1980s by six different seed companies. Nine of these stocks were produced using a
method that involved crossing B73 to other stocks, in some cases backcrossing to B73 multiple times, then selfing for sufficient numbers of generations to insure uniformity and by increasing (propagating) for a sufficient number of generations to produce commercial seed. F42 was an exception in that it was derived from a mutagenesis of B73 with nitrosoguanidine before selfing and increasing. The number of generations involved in creating the inbred stocks varied from five to twelve. Assuming a minimum of three generation to the last common B73 ancestor before development of the stocks and three generations of maintenance by the seed company and/or the national seed storage laboratory (GRIN) afterwards yields an average of 15 generations to the last common ancestor for each centromere. Thus the ten inbred stocks plus B73 encompass 165 (15 x 11) organismal generations of potential change (Figure 2). More relevantly, the number of cell divisions that are required to produce the complete maize lineage from a single-cell embryo to an egg or sperm has been estimated to be 50 (Otto and Walbot 1990). Thus the number of cellular generations encompassed in the ten inbred stocks plus B73 is on the order of 8000.

By analysis of SNP frequencies in these inbred stocks (van Heerwaarden et al. 2012), we were able to determine the location of recombination breakpoints within the chromosomes and which centromeres had been contributed by B73 during their production. We focused our analysis on centromeres 2 and 5 because these two have the highest quality assemblies (Wolfgruber et al. 2009). Five of the inbreds contained either a centromere 2 or a centromere 5 that had not been contributed by B73 (Figure 3A,B). We measured CENH3 fold enrichment by normalizing the number of uniquely-mapping reads per 200-KB locus by reads from a control library prepared from total nucleosomes. To our
surprise, the only large-scale variation in peaks of CENH3 enrichment we observed occurred in centromeres that were not derived from the B73 parent. In the case of centromere 2, the centromeres were all closely positioned, but two of the inbreds, PHW17 and PHG86, had very little enrichment for CENH3 compared to the others, correlating with their large differences in SNP frequency with B73 (Figure 3A). In the case of centromere 5, two alternative centromere positions were present, as indicated by the presence of two peaks of CENH3 enrichment. B73 centromeres produced only the left peak, while three of the four non-B73 centromeres produced only the right peak. One inbred, LP5, was of particular interest in that a large, central portion of chromosome 5 was not contributed by B73, but its centromere still had high SNP similarity to B73, particularly in the area of the left peak. Like B73, LP5 produced only a centromere 5 left peak. This result suggests that genetic similarity to B73 favors a B73-like distribution of CENH3, even though some SNPs may have accumulated.

Although centromeres 2 and 5 have the highest quality assemblies, the approximate location of multiple other centromeres is discernable on the other chromosomes as well. Examples of apparent deviations from the B73 pattern are illustrated by PHW17 and PHG86 (Figure 3C). For example, centromeres 8 and 10 of PHG86 are so highly diverged from B73 that they produce almost no CENH3 peak relative to the B73 reference assembly. We note, however, that SNP densities mapped to the B73 reference genome can give a very misleading picture of the real genetic differences in these inbreds. Maize is notorious for the amount of structural variation between populations, including in centromeres, and it is entirely possible that regions the size of centromeres can be lost, inverted, duplicated, or otherwise rearranged—none of which would be directly visible by SNPs (ALBERT et al.)
These data support the view that changes in centromere sequence can lead to substantial changes in centromere position (TOPP et al. 2009; BURRACK and BERMAN 2012; LIU et al. 2015).

Large-scale stability of centromeres and their locations on the B73 RefGen v3 genome assembly

In order to facilitate genetic and genomic analyses that require coordinates of centromeres on the physical map, we systematically defined the boundaries of each centromere in B73 using HOMER peak-finding software (HEINZ et al. 2010) (Table 1). As depicted in Figure 3C, several centromeres are either poorly assembled in the version 3 reference genome (B73 RefGen v3) or lack sufficient sequence complexity for unique read mapping such that they produced apparently minute centromeres (centromeres 1 and 6 being the worst cases). Furthermore, chromosome 10 has a misassembled extra centromere on the end of its short arm. Limiting the precision of this analysis was the complex patterns of CENH3 density across individual centromeres, particularly their lack of sharp boundaries (Figure 4A-D). Rather than a clear demarcation between the pericentromere and centromere, centromeres were bounded by regions of low (less than 20-fold) but reproducible CENH3 enrichment, similar to what has been observed in chicken neocentromeres (SHANG et al. 2013). Given the variability in ChIP efficiency between samples and the small numbers of reads in centromere tails, we hypothesized that the center of mass of CENH3 enrichment would provide a more robust metric than centromere boundaries for detection of potential small-scale changes in centromere positions. We found that the basic shape and location of centromeres is stably maintained, albeit with
some apparent shifting in the center of mass left or right (Table 2). Given that the center of mass was also influenced by IP efficiency (Table 2 and Figure 4E), we cannot evaluate the significance of these minor shifts. It is particularly noteworthy that centromere 5 of LP5 diverged from B73 long enough ago that it has accumulated SNPs, yet its CENH3 distribution (Figure 4C) and center of mass (Table 2) are within the range of the B73 centromere 5 in inbred stocks which had obtained a B73 centromere directly by interbreeding.

Patterns of CENH3 occupancy correlate with genetic elements

The high-resolution ChIP-seq maps for centromeres 2 and 5 (20-Kb loci, Figure 4) revealed a highly non-uniform distribution of CENH3 within centromeres, where peaks indicate regions of higher CENH3 occupancy and dips indicate regions of lower CENH3 occupancy. Both centromere 2 and 5 contained multiple large peaks and dips, but to a lesser magnitude in centromere 5. We wondered whether the non-uniformity of CENH3 enrichment could be a consequence of genetic elements that either favor or inhibit CENH3 accumulation. To examine this possibility, we divided the entire set of centromere-overlapping 20-kb loci into two categories based on the measured enrichment for CENH3, where the one-half of the loci with greater than the median enrichment are “high CENH3 loci” and the one half of the loci with less than the median enrichment are “low CENH3 loci”. We then measured overlap between centromeric loci and protein-coding genes, the 156-bp tandem repeat CentC (BILINSKI et al. 2015), and four centromeric retrotransposons CRM1-CRM4 (SHARMA and PRESTING 2014). Since the measured value for CENH3 enrichment is determined not only by ChIP reads but also by the control reads used for normalization,
we used two control datasets, one derived from total nucleosomes and another from randomly fragmented DNA. In addition, we also measured CENH3 enrichment using a normalization-free approach, simply by comparing the raw number of CENH3 reads at each locus with the expected number based on a uniform distribution across the genome. For this normalization-free approach we included non-uniquely-mapping reads to avoid a misleading depletion of reads from repetitive elements. All three methods consistently indicated a negative correlation of CENH3 with genes and a positive correlation with CentC and CRM1 (Figure 5A). CRM2 and CRM3 were strongly enriched in the centromere as a whole (Figure 5B), but it was unclear whether their enrichment correlated with CENH3 levels within centromeres (Figure 5A).

The relatively low CENH3 enrichment in genes in centromeres is not surprising because similar observations have been made in nematodes, rice and oat-maize addition lines (Yan et al. 2008; Gassmann et al. 2012; Wang et al. 2014a). While CENH3 was clearly depleted within these genes relative to the centromere as a whole, we wondered whether CENH3 was enriched in these genes relative to the genome as a whole. The 40 genes we identified within maize centromeres are listed in Table 3, though this list should be interpreted with caution as it may include pseudogenes, missannotations, or genes that were incorrectly placed in the reference genome assembly. We compared the number of CENH3 ChIP reads from the B73 sample to the number of reads from the B73 input for each of these genes. The coverage of six of these genes was too low to measure enrichment with statistical significance, but the other 34 were all enriched for CENH3 relative to the genome average (one-tailed p-value < 0.01). Data from our collaborators suggest that those
genes with the highest expression have the lowest CENH3 enrichment (Zhao, H., Zhu X., Wang K., Gent JI, Zhang, W., Zhang, T., Dawe, R.K., and Jiang, J., submitted).

DISCUSSION

A classical feature of epigenetic phenomena is their instability, variability, and low penetrance, exemplified by phenomena such as position effect variegation (Ptashne 1986), genomic imprinting (Kermicle 1970; Morgan et al. 1999), paramutation (Brink 1956), transposon silencing (Singh et al. 2008), and other epialleles (Schmitz et al. 2011). The fact that centromeres are determined by the CENH3 histone variant and are subject to large-scale movement events strongly suggests that centromeres are epigenetically defined (Heun et al. 2006; Smith et al. 2011; Purgato et al. 2014; Wang et al. 2014a). In the most extreme case, centromeres could be entirely unconstrained by sequence, such that observed distributions of cenH3 across centromeres reflect population averages, with each individual having its own unique cenH3 signature (Figure 1). A recent detailed study of nucleosome number and occupancy in human cells provides good support for such an interpretation (Bodor et al. 2014). Among other remarkably variably characteristics, the Bodor et al study revealed that only 4% of the centromeric nucleosomes were occupied cenH3 at any given time, that cenH3 loading appeared to follow a mass-action mechanism, and that cenH3 was randomly segregated to sister chromatids at mitosis. This quantitative view combined with a self-propagating system for cenH3 recruitment predicts that centromere sizes and locations vary among cells, individuals, and ultimately lineages and populations. Here we tested this prediction using maize lineages that had been separated
on the order of 15 organismal generations (Figure 2), which correspond to approximately 750 cellular generations (Otto and Walbot 1990). (The number of cellular generations is the more pertinent number in terms of cycles of cenH3 dilution and replenishment). Nonetheless, we did not find evidence for significant shifting of centromere position over this time scale: the size, shape, and position of maize centromeres on identical sequences passed down through different lineages were remarkably conserved. The only variation in maize CENH3 distributions we could confidently detect was connected with differences in DNA sequence (Figures 3 and 4).

It is difficult to envision a purely epigenetic mechanism stable enough for such long-term precision in cenH3 distributions. The existence of limiting components could enforce a stable overall centromere size (Zhang and Dawe 2012; Wang et al. 2014a), but it is not clear how they could maintain precise positioning of cenH3 domains. Instead, the complex yet stable patterns of maize CENH3 density on centromeres suggest genetic constraints on CENH3 positioning (Figures 4 and 5). There are a number of reasons to be skeptical about the role of genetics, as native centromere sequences have been shown to be insufficient to nucleate centromere formation (Phan et al. 2007) and multiple cases of normal centromere DNA lacking cenH3 have been reported, including in maize (Han et al. 2009; Liu et al. 2015). These results suggest that DNA sequences cannot be entirely sufficient for centromere formation, but leave open the possibility that particular DNA sequences may reinforce the stability of centromeres. One type of sequence in particular, sequence that encodes genes, could negatively enforce centromere positions, that is by inhibiting cenH3. Genes could restrict the accumulation of cenH3 either because their sequences do not interact stably with cenH3 nucleosomes, because they are associated with euchromatin, or
because of their transcriptional activity. (As cenH3 lacks key residues that are post-
translationally modified on canonical H3 (Talbert et al. 2012), it could also interfere with
gene regulation). While there may be an antagonistic relationship between cenH3 and
genes, transcription in non-genic contexts is thought to promote multiple aspects of
centromere chromatin regulation (for review, see Gent and Dawe 2012). Studies in fission
yeast suggests that transcription through certain DNA sequences that facilitate pausing of
RNA polymerase contributes toward cenH3 recruitment (Catania et al. 2015).

It is likely that the abundant tandem repeats found in many eukaryotic centromeres
contribute to centromere stability. Tandem repeats in both maize and rice have a sequence
composition that favors strong interactions between DNA and nucleosomes (Gent et al.
2011; Zhang et al. 2013b). Similarly, centromeric retrotransposons might not only target
cenH3-containing chromatin (Neumann et al. 2011; Bircher and Presting 2012), but could
also contribute to a genetically favorable environment for cenH3. One centromeric
retrotransposon, CRM2 is an interesting candidate in that it has the capability of phasing
maize CENH3-containing nucleosomes relative to its long terminal repeats (Gent et al.
2011). However, while the tandem repeat CentC and three centromeric retrotransposons—
CRM1, CRM2, and CRM3—were highly enriched in centromeres, only CentC and CRM1 were
clearly overrepresented in high-CENH3-regions compared to low-CENH3- regions of
centromeres. In maize, neocentromeres occupy smaller regions than native centromeres
(~300 Kb vs ~2000 Kb, which could reflect a reduced efficiency of CENH3 recruitment or
maintenance as a result of non-optimal sequence features such as lack of tandem repeats or
centromeric retrotransposons (Fu et al. 2013; Liu et al. 2015). Similarly, some human
neocentromeres are associated with defects in structure and transmission (ALONSO et al. 2010; BASSETT et al. 2010), which might be explained by non-optimal sequence features.

Within a single centromere, the influences of both favorable genetic elements such as repeats and inhibitory elements such as genes could produce a complex centromere shape such as maize centromere 2 with areas of greater and lesser cenH3 enrichment. The subtle reinforcing influences of DNA sequence could serve to stabilize an otherwise dynamic epigenetic replication process and allow centromeres to be maintained with a stability that nearly mimics a purely genetic form of inheritance. Major genetic changes such as deletions or duplications within the centromere core would naturally cause shifts in centromeres and promote the sudden loss and movement of centromeres to new locations. In fact, most documented neocentromeres can be traced to genetic events that affected the sequence of an existing centromere (BURRACK and Berman 2012). In conclusion, while it is clear that centromeres can undergo changes in size, shape, and position independently of DNA sequence, these findings indicate that the diversity of cenH3 distributions present in maize lines is predominantly a consequence of genetic diversity rather than positional drift.

ACKNOWLEDGEMENTS

We thank Alex Harkess and Nathanael Ellis for help with bioinformatics software, Paul Bilinksi and Jeffrey Ross-Ibarra for a CentC consensus sequence and for advice on SNP analysis, and Jennifer Monson-Miller for an Illumina sequencing library preparation
protocol. This study was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia’s Office of the Vice President for Research and Office of the Vice President for Information Technology.
REFERENCES

Albert, P. S., Z. Gao, T. V. Danilova and J. A. Birchler, 2010 Diversity of chromosomal karyotypes in maize and its relatives. Cytogenet Genome Res 129: 6-16.

Alonso, A., D. Hasson, F. Cheung and P. E. Warburton, 2010 A paucity of heterochromatin at functional human neocentromeres. Epigenetics Chromatin 3: 6.

Bassett, E. A., S. Wood, K. J. Salimian, S. Ajith, D. R. Foltz et al., 2010 Epigenetic centromere specification directs aurora B accumulation but is insufficient to efficiently correct mitotic errors. J Cell Biol 190: 177-185.

Bilinski, P., K. Distor, J. Gutierrez-Lopez, G. M. Mendoza, J. Shi et al., 2015 Diversity and evolution of centromere repeats in the maize genome. Chromosoma 124: 57-65.

Birchler, J. A., and G. G. Presting, 2012 Retrotransposon insertion targeting: a mechanism for homogenization of centromere sequences on nonhomologous chromosomes. Genes Dev 26: 638-640.

Blower, M. D., B. A. Sullivan and G. H. Karpen, 2002 Conserved organization of centromeric chromatin in flies and humans. Dev Cell 2: 319-330.

Bodor, D. L., J. F. Mata, M. Sergeev, A. F. David, K. J. Salimian et al., 2014 The quantitative architecture of centromeric chromatin. Elife 3: e02137.

Brink, R. A., 1956 A Genetic Change Associated with the R Locus in Maize Which Is Directed and Potentially Reversible. Genetics 41: 872-889.

Burrack, L. S., and J. Berman, 2012 Neocentromeres and epigenetically inherited features of centromeres. Chromosome Res 20: 607-619.

Catania, S., A. L. Pidoux and R. C. Allshire, 2015 Sequence Features and Transcriptional Stalling within Centromere DNA Promote Establishment of CENP-A Chromatin. PLoS Genet 11: e1004986.

Earnshaw, W. C., R. C. Allshire, B. E. Black, K. Bloom, B. R. Brinkley et al., 2013 Esperanto for histones: CENP-A, not CenH3, is the centromeric histone H3 variant. Chromosome Res 21: 101-106.

Eichten, S. R., J. M. Foerster, N. de Leon, Y. Kai, C. T. Yeh et al., 2011 B73-Mo17 near-isogenic lines demonstrate dispersed structural variation in maize. Plant Physiol 156: 1679-1690.

Fachinetti, D., H. Diego Folco, Y. Nechemia-Arbely, L. P. Valente, K. Nguyen et al., 2013 A two-step mechanism for epigenetic specification of centromere identity and function. Nat Cell Biol 15: 1056-1066.

Fang, Z., T. Pyhäjärvi, A. L. Weber, R. K. Dawe, J. C. Glaubitz et al., 2012 Megabase-scale inversion polymorphism in the wild ancestor of maize. Genetics 191: 883-894.

Ferreri, G. C., J. D. Brown, C. Obergfell, N. Jue, C. E. Finn et al., 2011 Recent amplification of the kangaroo endogenous retrovirus, KERV, limited to the centromere. J Virol 85: 4761-4771.

Folco, H. D., C. S. Campbell, K. M. May, C. A. Espinoza, K. Oegema et al., 2015 The CENP-A N-tail confers epigenetic stability to centromeres via the CENP-T branch of the CCAN in fission yeast. Curr Biol 25: 348-356.

Fu, S., Z. Lv, Z. Gao, H. Wu, J. Pang et al., 2013 De novo centromere formation on a chromosome fragment in maize. Proc Natl Acad Sci U S A 110: 6033-6036.
Fukagawa, T., and W. C. Earnshaw, 2014 The centromere: chromatin foundation for the kinetochore machinery. Dev Cell 30: 496-508.

Gassmann, R., A. Rechtsteiner, K. W. Yuen, A. Muroyama, T. Egelhofer et al., 2012 An inverse relationship to germline transcription defines centromeric chromatin in C. elegans. Nature 484: 534-537.

Gent, J. I., and R. K. Dawe, 2012 RNA as a structural and regulatory component of the centromere. Annu Rev Genet 46: 443-453.

Gent, J. I., Y. Dong, J. Jiang and R. K. Dawe, 2012 Strong epigenetic similarity between maize centromeric and pericentromeric regions at the level of small RNAs, DNA methylation and H3 chromatin modifications. Nucleic Acids Res 40: 1550-1560.

Gent, J. I., T. F. Madzima, R. Bader, M. R. Kent, X. Zhang et al., 2014 Accessible DNA and Relative Depletion of H3K9me2 at Maize Loci Undergoing RNA-Directed DNA Methylation. Plant Cell 26: 4903-4917.

Gent, J. I., K. L. Schneider, C. N. Topp, C. Rodriguez, G. G. Presting et al., 2011 Distinct influences of tandem repeats and retrotransposons on CENH3 nucleosome positioning. Epigenetics Chromatin 4: 3.

Gilbert, W., and B. Müller-Hill, 1967 The lac operator is DNA. Proc Natl Acad Sci U S A 58: 2415-2421.

Gong, Z., Y. Wu, A. Koblízková, G. A. Torres, K. Wang et al., 2012 Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell 24: 3559-3574.

Han, F., Z. Gao and J. A. Birchler, 2009 Reactivation of an inactive centromere reveals epigenetic and structural components for centromere specification in maize. Plant Cell 21: 1929-1939.

Hasson, D., T. Panchenko, K. J. Salimian, M. U. Salman, N. Sekulic et al., 2013 The octamer is the major form of CENP-A nucleosomes at human centromeres. Nat Struct Mol Biol 20: 687-695.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin et al., 2010 Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38: 576-589.

Henikoff, J. G., J. Thakur, S. Kasinathan and S. Henikoff, 2015 A unique chromatin complex occupies young α-satellite arrays of human centromeres. Science Advances 1.

Heun, P., S. Erhardt, M. D. Blower, S. Weiss, A. D. Skora et al., 2006 Mislocalization of the Drosophila centromere-specific histone CID promotes formation of functional ectopic kinetochores. Dev Cell 10: 303-315.

Hudson, D. F., K. J. Fowler, E. Earle, R. Saffery, P. Kalitsis et al., 1998 Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. J Cell Biol 141: 309-319.

Ingouff, M., S. Rademacher, S. Holec, L. Soljić, N. Xin et al., 2010 Zygotic resetting of the HISTONE 3 variant repertoire participates in epigenetic reprogramming in Arabidopsis. Curr Biol 20: 2137-2143.

Ishii, T., R. Karimi-Ashtiyani, A. M. Banaei-Moghaddam, V. Schubert, J. Fuchs et al., 2015 The differential loading of two barley CENH3 variants into distinct centromeric substructures is cell type- and development-specific. Chromosome Res.
Jin, W., J. R. Melo, K. Nagaki, P. B. Talbert, S. Henikoff et al., 2004 Maize centromeres: organization and functional adaptation in the genetic background of oat. Plant Cell 16: 571-581.

Kato, H., J. Jiang, B. R. Zhou, M. Rozendaal, H. Feng et al., 2013 A conserved mechanism for centromeric nucleosome recognition by centromere protein CENP-C. Science 340: 1110-1113.

Kermicle, J. L., 1970 Dependence of the R-mottled aleurone phenotype in maize on mode of sexual transmission. Genetics 66: 69-85.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

Liu, Y., H. Su, J. Pang, Z. Gao, X. J. Wang et al., 2015 Sequential de novo centromere formation and inactivation on a chromosomal fragment in maize. Proc Natl Acad Sci U S A 112: E1263-1271.

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

Morgan, H. D., H. G. Sutherland, D. I. Martin and E. Whitelaw, 1999 Epigenetic inheritance at the agouti locus in the mouse. Nat Genet 23: 314-318.

Nagaki, K., P. B. Talbert, C. X. Zhong, R. K. Dawe, S. Henikoff et al., 2003 Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of Arabidopsis thaliana centromeres. Genetics 163: 1221-1225.

Nelson, P. T., N. D. Coles, J. B. Holland, D. M. Bubeck, S. Smith et al., 2008 Molecular Characterization of Maize Inbreds with Expired U.S. Plant Variety Protection.

Neumann, P., A. Navrátilová, A. Kobližková, E. Kejnovský, E. Hřibová et al., 2011 Plant centromeric retroposons: a structural and cytogenetic perspective. Mob DNA 2: 4.

Ohzeki, J., M. Nakano, T. Okada and H. Masumoto, 2002 CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. J Cell Biol 159: 765-775.

Otto, S. P., and V. Walbot, 1990 DNA methylation in eukaryotes: kinetics of demethylation and de novo methylation during the life cycle. Genetics 124: 429-437.

Phan, B. H., W. Jin, C. N. Topp, C. X. Zhong, J. Jiang et al., 2007 Transformation of rice with long DNA-segments consisting of random genomic DNA or centromere-specific DNA. Transgenic Res 16: 341-351.

Ptashne, M., 1967 Specific binding of the lambda phage repressor to lambda DNA. Nature 214: 232-234.

Ptashne, M., 1986 Gene regulation by proteins acting nearby and at a distance. Nature 322: 697-701.

Purgato, S., E. Belloni, F. M. Piras, M. Zoli, C. Badiale et al., 2014 Centromere sliding on a mammalian chromosome. Chromosoma.

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

Russell, W. A., 1972 Registration of B70 and B73 parental lines of maize. Crop Science 12.
Schmitz, R. J., M. D. Schultz, M. G. Lewsey, R. C. O’Malley, M. A. Urich et al., 2011 Transgenerational epigenetic instability is a source of novel methylation variants. Science 334: 369-373.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei et al., 2009 The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112-1115.

Shang, W. H., T. Hori, N. M. Martins, A. Toyoda, S. Misu et al., 2013 Chromosome engineering allows the efficient isolation of vertebrate neocentromeres. Dev Cell 24: 635-648.

Sharma, A., and G. G. Presting, 2014 Evolution of centromeric retrotransposons in grasses. Genome Biol Evol 6: 1335-1352.

Singh, J., M. Freeling and D. Lisch, 2008 A position effect on the heritability of epigenetic silencing. PLoS Genet 4: e1000216.

Smith, K. M., P. A. Phatale, C. M. Sullivan, K. R. Pomraning and M. Freitag, 2011 Heterochromatin is required for normal distribution of Neurospora crassa CenH3. Mol Cell Biol 31: 2528-2542.

Steiner, F. A., and S. Henikoff, 2014 Holocentromeres are dispersed point centromeres localized at transcription factor hotspots. Elife 3: e02025.

Talbert, P. B., K. Ahmad, G. Almouzni, J. Ausió, F. Berger et al., 2012 A unified phylogeny-based nomenclature for histone variants. Epigenetics Chromatin 5: 7.

Topp, C. N., R. J. Okagaki, J. R. Melo, R. G. Kynast, R. L. Phillips et al., 2009 Identification of a maize neocentromere in an oat-maize addition line. Cytogenet Genome Res 124: 228-238.

van Heerwaarden, J., M. B. Hufford and J. Ross-Ibarra, 2012 Historical genomics of North American maize. Proc Natl Acad Sci U S A 109: 12420-12425.

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

Wang, L., Z. Zeng, W. Zhang and J. Jiang, 2014b Three potato centromeres are associated with distinct haplotypes with or without megabase-sized satellite repeat arrays. Genetics 196: 397-401.

Wolfgruber, T. K., A. Sharma, K. L. Schneider, P. S. Albert, D. H. Koo 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.

Yan, H., P. B. Talbert, H. R. Lee, J. Jett, S. Henikoff et al., 2008 Intergenic locations of rice centromeric chromatin. PLoS Biol 6: e286.

Yao, J., X. Liu, T. Sakuno, W. Li, Y. Xi et al., 2013 Plasticity and epigenetic inheritance of centromere-specific histone H3 (CENP-A)-containing nucleosome positioning in the fission yeast. J Biol Chem 288: 19184-19196.

Zhang, B., Z. Lv, J. Pang, Y. Liu, X. Guo et al., 2013a Formation of a functional maize centromere after loss of centromeric sequences and gain of ectopic sequences. Plant Cell 25: 1979-1989.

Zhang, H., and R. K. Dawe, 2012 Total centromere size and genome size are strongly correlated in ten grass species. Chromosome Res 20: 403-412.

Zhang, T., P. B. Talbert, W. Zhang, Y. Wu, Z. Yang et al., 2013b The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. Proc Natl Acad Sci U S A 110: E4875-4883.
Zhong, C. X., J. B. Marshall, C. Topp, R. Mroczek, A. Kato et al., 2002 Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. Plant Cell 14: 2825-2836.
**FIGURE LEGENDS**

**Figure 1: Conceptual models for stability of centromeres**

In the centromere drift model, centromere diversity can arise because of accumulated small changes in cenH3 recruitment or maintenance and independently of changes in centromere DNA. In the centromere constraint model, centromere diversity would arise because of changes in the constraining factors, e.g., structural changes in centromere DNA or disruption of cenH3 recruitment/maintenance conditions.

**Figure 2: Lineages of B73 centromeres in inbred stocks used for ChIP**

Assuming averages of three generations to last common B73 ancestor, nine generations in development of the inbred stocks, and a further three generations of maintenance in the subsequent three decades allow us to conservatively estimate that fifteen generations separate each inbred stock. Further details on the development of the inbred stocks can be found in their Plant Variety Protection (PVP) Certificates available through the GRIN website (www.ars-

**Figure 3: Differences in apparent centromere location correlate with genetic change**

**(A)** Chromosome 2 CENH3 enrichment and SNP frequency in B73 and seven partial-B73 inbred stocks. “CENH3 fold enrichment” is the normalized ratio of CENH3 ChIP reads to total nucleosome control reads per 200-kb locus across the genome. Only uniquely-mapping reads were included in this analysis. Three of the inbreds that had only B73-like centromeres are omitted from this plot because of graphical limitations, but they are
included in Figure 4. The B73 used for this analysis is the Dawe lab “reference” (see Figure 2).

(B) Chromosome 5 CENH3 enrichment and SNP frequency in B73 and seven partial-B73 inbred stocks, as in (A).

(C) CENH3 enrichment and SNP frequency for all chromosomes, as in (A).

Figure 4: Centromere location and shape are maintained across lineages

(A) High-resolution view of centromere 2 CENH3 enrichment in B73 and seven partial-B73 inbred stocks. “CENH3 fold enrichment” is the normalized ratio of CENH3 ChIP reads to total nucleosome control reads per 20-kb locus across the genome. Only uniquely-mapping reads were included in this analysis. Gaps in the plot are due to exclusion of loci with fewer than 50 control reads from the analysis. The B73 used for this analysis is the Dawe lab “reference” (see Figure 2).

(B) High-resolution view of centromere 2 CENH3 enrichment in purebred B73 stocks, as in (A).

(C) High-resolution view of centromere 5 CENH3 enrichment in B73 and seven partial-B73 inbred stocks, as in (A).

(D) High-resolution view of centromere 5 CENH3 enrichment in purebred B73 stocks, as in (A).

(E) Apparent shift in CENH3 center of mass is related to ChIP efficiency. The center of mass shifts in Table 2 were plotted against CENH3 ChIP enrichment, and linear regression lines were plotted for centromere 2 and 5 separately.
Figure 5: Genes and repetitive DNA in centromeres

(A) Overlap between CENH3 enrichment and genetic elements. Within the centromere, we assigned each 20-kb locus into one of two categories, “low CENH3 loci” and “high CENH3 loci” based on whether it exhibited less than or greater than the median CENH3 enrichment. Results are shown for three methods of measuring CENH3 enrichment. The first two are relative to a total nucleosome control produced by MNase digestion of total chromatin and relative to a randomly sheared naked DNA control produced by NEBNext dsDNA Fragmentase (Gent et al. 2014). In both cases, only uniquely mapping reads were considered. In the third method, no control was used, and enrichment was defined by raw read counts, including non-uniquely-mapping reads. Errors bars are standard errors of the means for each set of loci. (B) Comparison of genetic elements in centromeric 20-kb loci and whole genome 20-kb loci, but without regard to relative CENH3 enrichment within centromeres
Table 1: Centromere positions on the B73 RefGen v3 assembly

All regions of CENH3 enrichment over 100-Kb are listed, as is the single region on chromosome 1 over 10-Kb in size. Positions and CENH3 fold enrichment were produced by HOMER peak finding software (Heinz et al. 2010).

| Chromosome | Start (Mb) | End (Mb) | Region size (Kb) | CENH3 fold enrichment |
|------------|------------|----------|------------------|------------------------|
| 1          | 134.57     | 134.58   | 10               | 11                     |
| 2          | 93.50      | 95.37    | 1869             | 56                     |
| 3          | 99.79      | 100.84   | 1045             | 52                     |
| 4          | 103.61     | 103.98   | 368              | 67                     |
| 4          | 105.36     | 105.73   | 370              | 97                     |
| 4          | 105.79     | 106.21   | 419              | 56                     |
| 5          | 102.06     | 104.16   | 2103             | 75                     |
| 6          | 39.15      | 39.31    | 160              | 97                     |
| 7          | 22.73      | 23.05    | 322              | 114                    |
| 8          | 49.18      | 49.66    | 481              | 46                     |
| 8          | 49.79      | 49.91    | 123              | 128                    |
| 8          | 49.97      | 50.23    | 264              | 103                    |
| 8          | 50.44      | 50.98    | 549              | 74                     |
| 9          | 52.42      | 53.68    | 1262             | 105                    |
| 9          | 53.76      | 53.90    | 135              | 12                     |
| 9          | 54.06      | 54.21    | 150              | 18                     |
| 10         | 0.00       | 0.49     | 486              | 97                     |
| 10         | 50.17      | 51.73    | 1565             | 80                     |
Table 2: Center of CENH3 mass positions in centromeres 2 and 5 of B73 and related inbreds

For the centromeres that were similar to B73 among the inbreds, we measured the center of mass of CENH3 enrichment by treating each centromere as a series of discreet 20-Kb loci and multiplying the value for CENH3 enrichment by the chromosomal position of each locus. We defined shift to be the difference between the center of mass for a particular sample and the average center of mass for the whole group. “Est. % B73” indicates the total amount of genomic DNA contributed by B73 during the production of each inbred (NELSON et al. 2008). In addition to the B73 and nine of the B73-related inbreds already discussed (PHW17 excluded because neither its centromere 2 nor centromere 5 are B73-like), we included three other B73 sample populations. Two of the B73 samples had been separated from the first B73 (“reference”) by two generations of self crosses (“selfed sibling 1” and “selfed sibling 2”), and one B73 had been separated from the first by an unknown number of generations (“distant kin”; the last common ancestor is documented as greater than 6 generations earlier). For a high-resolution view of CENH3 distributions across centromeres, see Figure 4.

|            | Centromere 2 | Centromere 5 |
|------------|--------------|--------------|

| Inbred                          | Est. % B73 | Center of mass (Mb) | Shift relative to average (Mb) | CENH3 fold enrichment | Center of mass (Mb) | Shift relative to average (Mb) | CENH3 fold enrichment |
|--------------------------------|------------|---------------------|-------------------------------|------------------------|---------------------|-------------------------------|------------------------|
| B73 (reference)                | 100        | 94.44               | -0.01                         | 56                     | 103.02              | 0.05                          | 75                     |
| B73 (selfed sibling 1)         | 100        | 94.44               | -0.01                         | 63                     | 103.01              | 0.07                          | 83                     |
| B73 (selfed sibling 2)         | 100        | 94.42               | 0.01                          | 56                     | 103.05              | 0.02                          | 75                     |
| B73 (distant kin)              | 100        | 94.45               | -0.01                         | 45                     | 103.02              | 0.05                          | 50                     |
| F42                            | 100        | 94.44               | -0.01                         | 54                     | 103.09              | -0.01                        | 65                     |
| DJ7                            | 100        | 94.46               | -0.03                         | 44                     | 103.05              | 0.02                          | 59                     |
| LH119                          | 80         | 94.42               | 0.02                          | 51                     | 103.09              | -0.01                        | 66                     |
| LH132                          | 80         | 94.45               | -0.01                         | 36                     | 103.16              | -0.09                        | 42                     |
| LH149                          | 80         | 94.40               | 0.03                          | 48                     | 103.08              | -0.01                        | 56                     |
| LP5                            | 80         | 94.34               | 0.10                          | 48                     | 103.15              | -0.08                        | 50                     |
| NS701                          | 60         | 94.54               | -0.10                         | 34                     |                     |                               |                        |
| PHG86                          | 60         |                     |                               | 103.08                 |                     | -0.01                        | 62                     |
| LH74                           | 50         | 94.42               | 0.01                          | 52                     |                     |                               |                        |

Table 3: CENH3 enrichment in putative centromere genes
All annotated protein-coding genes from the Zea_mays.AGPv3.21 reference list that overlap with a region of CENH3 enrichment in Table 1 are listed here. “CENH3 fold enrichment” is the normalized ratio of input read count to reference B73 CENH3 ChIP read count for each gene (uniquely-mapping reads only). A one-tailed p-value was calculated for each gene based on the null hypothesis that the proportion of reads mapping to the gene was smaller in the CENH3 ChIP library than in the input library.

| Chromosome | Start position (bp) | End position (bp) | Length (bp) | Gene ID | Input read count | CENH3 ChIP read count | CENH3 fold enrichment | p-value |
|------------|---------------------|-------------------|-------------|---------|-----------------|-----------------------|-----------------------|---------|
| 2          | 93503688            | 93507160          | 3472        | GRMZM2G1545685 | 164             | 53                    | 6.1                   | 0.000   |
| 2          | 93627363            | 93630442          | 3079        | GRMZM2G326177 | 32              | 49                    | 28.8                  | 0.000   |
| 2          | 93759614            | 93761482          | 1868        | GRMZM2G083935 | 28              | 21                    | 14.1                  | 0.000   |
| 2          | 93810893            | 93815199          | 706         | GRMZM5G89541 | 14              | 45                    | 65.0                  | 0.000   |
| 2          | 94266077            | 94267223          | 1146        | GRMZM2G170586 | 25              | 8                     | 6.0                   | 0.009   |
| 2          | 94436882            | 94436805          | 203         | GRMZM2G421160 | 6               | 13                    | 40.8                  | 0.000   |
| 2          | 94558679            | 94559149          | 470         | GRMZM2G038916 | 5               | 10                    | 37.7                  | 0.001   |
| 2          | 94566238            | 94567668          | 530         | AC200178.2_FG003 | 3             | 0                     | 0.0                   | 0.958   |
| 2          | 95190274            | 95198729          | 8455        | GRMZM2G099580 | 209             | 145                   | 13.1                  | 0.000   |
| 3          | 99979951            | 99985771          | 6820        | AC233900.1_FG002 | 28        | 146                   | 98.2                  | 0.000   |
| 3          | 100262600           | 100302280         | 37688       | GRMZM2G409893 | 767             | 243                   | 6.0                   | 0.000   |
| 3          | 100817511           | 100821080         | 3569        | GRMZM2G139472 | 69              | 28                    | 7.6                   | 0.000   |
| 4          | 105880601           | 105881362         | 761         | GRMZM2G071042 | 13              | 3                     | 4.3                   | 0.093   |
| 5          | 102196478           | 102196962         | 484         | GRMZM2G000411 | 32              | 40                    | 23.5                  | 0.000   |
| 5          | 102484601           | 102485399         | 798         | GRMZM2G175425 | 4               | 13                    | 61.2                  | 0.000   |
| 5          | 102549951           | 102556475         | 6524        | GRMZM2G426140 | 74              | 270                   | 68.7                  | 0.000   |
| 5          | 102570932           | 102575325         | 4393        | GRMZM2G429781 | 11              | 43                    | 73.6                  | 0.000   |
| 5          | 102581188           | 102581854         | 666         | GRMZM2G115127 | 2               | 13                    | 122.4                 | 0.000   |
| 5          | 103419867           | 103420154         | 287         | GRMZM2G701752 | 20              | 37                    | 34.8                  | 0.000   |
| 5          | 103750694           | 103752458         | 1764        | GRMZM2G135228 | 19              | 31                    | 30.7                  | 0.000   |
| 7          | 22742037            | 22742852          | 815         | GRMZM2G700267 | 2               | 17                    | 160.1                 | 0.000   |
| 8          | 49642905            | 49643534          | 2629        | AC234517.1_FG005 | 0          | 0                     | 0.0                   | 0.000   |
| 8          | 49868596            | 49870710          | 2114        | GRMZM2G321072 | 63              | 234                   | 70.0                  | 0.000   |
| 8          | 49872505            | 49872997          | 492         | AC198230.4_FG000 | 21         | 105                   | 94.2                  | 0.000   |
| 8          | 50174904            | 50175264          | 450         | GRMZM2G426753 | 0               | 1                    | 0.0                   | 0.000   |
| 8          | 50459010            | 50464400          | 5390        | GRMZM2G345194 | 85              | 361                   | 80.0                  | 0.000   |
| 8          | 50755590            | 50758353          | 245         | GRMZM2G100662 | 13              | 47                    | 68.1                  | 0.000   |
| 9          | 52423318            | 52424569          | 1251        | GRMZM6G474234 | 39              | 4                     | 1.9                   | 0.171   |
| 9          | 52961497            | 52964749          | 3252        | GRMZM2G061764 | 40              | 98                    | 46.1                  | 0.000   |
| 9          | 53764682            | 53766013          | 1331        | GRMZM2G143473 | 35              | 22                    | 11.8                  | 0.000   |
| 9          | 53828063            | 53828852          | 789         | GRMZM2G528479 | 30              | 9                     | 5.6                   | 0.007   |
| 10         | 300326              | 310785            | 10459       | GRMZM2G040843 | 187             | 68                    | 6.8                   | 0.000   |
| 10         | 50308455            | 50309427          | 972         | GRMZM2G069264 | 8               | 10                    | 23.5                  | 0.001   |
| 10         | 50310327            | 50315148          | 4821        | GRMZM2G368486 | 134             | 278                   | 39.1                  | 0.000   |
| 10         | 50588819            | 50590827          | 2008        | GRMZM2G145909 | 51              | 83                    | 30.6                  | 0.000   |
| 10         | 50616674            | 50621840          | 5166        | AC196714.3_FG004 | 242       | 911                   | 70.9                  | 0.000   |
| 10         | 51010484            | 51019069          | 8585        | GRMZM2G137715 | 219             | 63                    | 5.4                   | 0.000   |
| 10         | 51456481            | 51456996          | 2515        | GRMZM2G361718 | 114             | 36                    | 5.9                   | 0.000   |
Centromere drift model:
Each individual has a unique cenH3 distribution

Centromere constraint model:
cenH3 distributions are conserved across individuals

passage through single-cell developmental lineages and organismal generations
Lineages of B73 centromeres in inbred stocks used for ChIP

B73
(released in 1972)

Unknown numbers of generations to last common B73 ancestor

Unknown number of generations between Dawe lab B73 and GRIN B73, but greater than 6

Dawe lab B73
("reference", "selfed sibling 1", "selfed sibling 2")

GRIN B73
("distant kin", unknown relation to reference and selfed sibling B73)

F42

DJ7, LH74, LH119, LH132, LH149, LP5, NS701, PHG86, PHW17

Unknown number of generations before deposition of stock with GRIN and of maintainence by GRIN before receipt by the Dawe lab in 2009

5 to 12 generations following last cross between B73 and non-B73 or following mutagenesis of B73 (F42 only) in 1970s and 1980s
