Chromosoma
DOI 10.1007/s00412-014-0483-8

RESEARCH ARTICLE

Diversity and evolution of centromere repeats in the maize genome

Paul Bilinski · Kevin Distor · Jose Gutierrez-Lopez · Gabriela Mendoza Mendoza · Jinghua Shi · R. Kelly Dawe · Jeffrey Ross-Ibarra

Received: 12 May 2014 / Revised: 21 July 2014 / Accepted: 11 August 2014

Abstract Centromere repeats are found in most eukaryotes and play a critical role in kinetochore formation. Though centromere repeats exhibit considerable diversity both within and among species, little is understood about the mechanisms that drive centromere repeat evolution. Here, we use maize as a model to investigate how a complex history involving polyploidy, fractionation, and recent domestication has impacted the diversity of the maize centromeric repeat CentC. We first validate the existence of long tandem arrays of repeats in maize and other taxa in the genus *Zea*. Although we find considerable sequence diversity among CentC copies genome-wide, genetic similarity among repeats is highest within these arrays, suggesting that tandem duplications are the primary mechanism for the generation of new copies. Nonetheless, clustering analyses identify similar sequences among distant repeats, and simulations suggest that this pattern may be due to homoplasious mutation. Although the two ancestral subgenomes of maize have contributed nearly equal numbers of centromeres, our analysis shows that the majority of all CentC repeats derive from one of the parental genomes, with an even stronger bias when examining the largest assembled contiguous clusters. Finally, by comparing maize with its wild progenitor teosinte, we find that the abundance of CentC likely decreased after domestication, while the pericentromeric repeat Cent4 has drastically increased.

Introduction

In spite of the rapid growth in the number of sequenced genomes, centromeres remain poorly understood and relatively cryptic due to their highly repetitive content. Centromere repeats are highly diverse across taxa and their turnover appears to be very rapid (Melters et al. 2013). However, little is known about the genetic mechanisms that produce centromere repeat diversity. Domesticated maize (*Zea mays* ssp. *mays*) has a high-quality genome assembly (Schnable et al. 2009) including complete sequence of two centromeres (Wolfgruber et al. 2009), and the breadth of research into maize centromeres makes it one of the best systems to investigate the processes governing centromere repeat evolution.

Maize centromeres are comprised primarily of the 156 bp satellite repeat CentC and the centromeric retrotransposon of maize (CRM) family. Both repeats interact with kinetochore proteins such as CENH3 (Wolfgruber et al. 2009; Zhong et al. 2002) and show variation in abundance across taxa (Albert et al. 2010). While considerable effort has gone to investigating the molecular function of maize centromere repeats (Ananiev et al. 1998; Nagaki et al. 2003; Wolfgruber et al. 2009), we know comparatively little about the evolution responsible for producing the current sequences. CRM elements are better understood, including the age and insertion

Electronic supplementary material The online version of this article (doi:10.1007/s00412-014-0483-8) contains supplementary material, which is available to authorized users.

P. Bilinski · K. Distor · J. Gutierrez-Lopez · G. M. Mendoza · J. Ross-Ibarra
Department of Plant Sciences, University of California Davis,
Davis, CA 95616, USA

J. Gutierrez-Lopez
Department of Natural Resources and the Environment,
University of New Hampshire, Durham, NH 03820, USA

J. Shi · R. K. Dawe
Department of Plant Biology, University of Georgia,
Athens, GA 30602, USA

J. Ross-Ibarra (✉)
The Genome Center and Center for Population Biology,
University of California Davis, Davis, CA 95616, USA
e-mail: rossibarra@ucdavis.edu

Published online: 05 September 2014

preferences of different CRM families (Wolfgruber et al. 2009; Sharma et al. 2008; Sharma and Presting 2014). In contrast, no in-depth characterization of the genetic diversity of centromere repeats in the maize genome exists.

In this paper, we describe the patterns of diversity of centromere repeats across the maize genome. We investigate whether the differential ancestry of maize centromeres (Wang and Bennetzen 2012) has led to chromosome-specific variation of CentC similar to that seen in other species (Kawabe and Nasuda 2005; Hall et al. 2005; Macas et al. 2010) and how genetic relatedness among individual CentC repeats varies spatially across the genome. We find that CentC copies do not form genetic groups consistent with ancient whole genome duplications or chromosome specificity, despite most of the large arrays of CentC originating from only one of the ancestral subgenomes of maize. We show higher genetic similarity of CentC repeats within clusters, indicating the predominance of tandem duplications in the formation of new CentC copies. Lastly, we use low-coverage sequencing and cytological data to show that domesticated maize has less CentC than its wild relatives.

Methods

CentC repeat identification and diversity

We downloaded 218 previously annotated CentC sequences (Ananiev et al. 1998; Nagaki et al. 2003) from GenBank. We then searched the B73 maize reference genome (5b60, www.maizesequence.org) with megaBLAST (McGinnis and Madden 2004) using the 218 annotated CentCs as a reference, keeping the longest hit with a length of over 140 bp and a minimum bit score of 100. In this analysis, we defined CentC copies as tandem if their start locations were within 1,000 bp.

All 12,162 CentC sequences were aligned using seven iterations of muscle (Edgar 2004) with default parameters. A Jukes-Cantor distance matrix of all sequences was calculated with PHYLIP (Felsenstein 1989, http://evolution.genetics.washington.edu/phylip.html), and an unrooted neighbor joining tree was built based on the distance matrix.

We used principle coordinate analysis (PCoA) to cluster CentC variants based on their genetic distances. Eigenvalues from the PCoA were used to determine the number of statistically significant clusters using the Tracy-Widom distribution (Patterson et al. 2006).

We employed the software SPAGeDi (Hardy and Vekemans 2002, http://ebe.ulb.ac.be/ebe/Software.html) to estimate the spatial autocorrelation of sequence similarity of CentC repeats in the completely sequenced centromeres 2 and 5. We calculated Morana's I statistic using Jukes-Cantor genetic distance and measures of physical distance between CentC repeats in base pairs. Confidence intervals for the values of I were estimated by 20,000 random permutations of the physical distances.

Statistical analyses were performed in R with the packages ape (Paradis et al. 2004) and RMTstat (Perry et al. 2009). We compared clusters to chromosome of origin and syntenic maps of maize ancient tetraploidy (Schnable et al. 2011) to determine if the genetic history of maize left a footprint on CentC similarity.

Read mapping and genome size correction

We mapped Illumina reads from a broad panel of *Zea* (Chia et al. 2012; Tenaillon et al. 2011) to a reference consisting of the full complement of 12,162 CentC variants identified in the B73 genome. Reads were mapped with Mosaik v1.0 (https://code.google.com/p/mosaik-aligner/). We first optimized mapping parameters by relaxing mapping stringency and evaluating the number of successfully mapped reads with each combination. Consistent with parameters from previous studies mapping reads to repetitive elements (Tenaillon et al. 2011), we required homology to remain at a minimum of 80 %. For other non-default parameters, we permuted over many values of hash size, alignment candidate threshold, percent of read aligning, and maximum number of hash positions per seed to find a combination that produced believable alignments. We selected an optimum combination of parameters just below the parameters where we observed a large increase in the total number of reads aligning (Figure S1). Our final set of parameters for tandem repeats used an initial hash size of 8, an alignment candidate threshold of 15 bases, 20 % of mismatching bases, a minimum of 30 % overlap to the reference, and stored the top 100 hits for alignment. After reads were mapped, we calculated the percentage of total reads hitting the given reference and multiplied this value by the relative genome size of each accession as reported in (Chia et al. 2012) and (Tenaillon et al. 2011). The total number of reads mapping did not change drastically when using one random copy of CentC versus the full AGPv2 reference, suggesting that our parameters are sufficiently broad to capture genome-wide CentC abundances. Because library preparation has an effect on estimates of repeat abundance (see results), we only used individuals from maize HapMap v2 (Chia et al. 2012) with libraries prepared using identical methods. We also used previously published sequence from whole genome chromatin immunoprecipitation (ChIP) (Wolfgruber et al. 2009; Wang et al. 2013) using CenH3. These reads were mapped with Bowtie2 (Langmead and Salzberg 2012) with the parameter-very sensitive-local.

We used a different set of mapping parameters for long repeats such as transposable elements. Previous studies (Schnable et al. 2009) estimated that approximately 85 % of maize genome derives from transposable elements. Using the short read libraries from (Tenaillon et al. 2011), we selected parameters so that approximately 85 % of the library mapped to the maize transposable element database (www.maizetedb.org) with a minimum homology of 80 %. The final parameters for TEs were a hash size of 10, alignment candidate threshold of 11, 80 % homology excluding non-aligned portions of the read, and a 30 % minimum overlap.

We designed a simulation to estimate the accuracy of our measurements of CentC content (code available at: https://github.com/kddistor/dnasims). In short, our simulations altered the copy number of CentC repeats over a region of fixed length (10 Mb), changing the percentage of the genome deriving from the repeat. Illumina reads were simulated from each of the DNA strings and mapped using our pipeline. These simulations showed that our pipeline captured relative differences in abundance well, but underestimated total abundance of CentC. We found that our pipeline could accurately capture differences of 0.05 % change in CentC abundance, suggesting that larger differences are likely to be biologically real (Figure S2).

Simulation of homoplasious mutations

In order to better understand patterns of diversity at CentC, we performed simulations to test the likelihood of homoplasious mutations (i.e., independent mutations occurring at the same position in two different CentC repeats). Our simulation (code available at: https://github.com/paulbilinski/CentC_Analyses/tree/master/Diversity_sims) assumed that CentC has been evolving for 1 million years since the divergence of maize and *Tripsacum* (Ross-Ibarra et al. 2009), a closely related genus whose centromere repeat shares a large amount of homology (Melters et al. 2013). We assumed a constant copy number, a mutation rate of 3×10^{-8} per generation (Clark et al. 2005), and one generation per year.

PacBio sequencing

Library preparation and sequencing was performed according to the methods described in (Melters et al. 2013). Using those protocols, we sequenced one individual each of maize, *Z. mays* ssp. *mexicana*, *Z. mays* ssp. *parviglumis*, and *Z. luxurians* with Pacific Biosciences (Pacific Biosciences, Menlo Park, CA) technology. Approximately 200 Mb of reads were produced from each cell, and reads with length greater than 600 bp were retained for analysis of tandem CentC content using BLAST (Table S1). CentC copies were considered tandem if the read had four CentC copies each within 300 bp of each other.

FISH

Fluorescent in situ hybridization (FISH) was carried out as described in Kato et al. (2004) for a B73 by *Z. luxurians* hybrid and Shi et al. (2010) for a B73 by *Z. mays* ssp. *parviglumis* hybrid.

Results

Centromere repeats in the maize genome

We found a total of 12,162 CentC copies in the maize reference genome and unassembled BACs. Of these, 8,259 were unique over their full length. While centromeres 2 and 5 are the only chromosomes with high-confidence sequencing of CentC copies, the levels of diversity observed on these two chromosomes is comparable to the rest of the genome (data not shown), suggesting that current assemblies of CentC sequence may estimate sequence variation with some accuracy. No CentC sequence occurred more than 10 times in the genome, and the vast majority (> 75 %, Table S2 of non-unique CentC variants occurred only twice. Of the 2,266 non-unique CentC sequences, only 3 were tandem, identical duplicates. Nearly all of the 10,639 CentC copies on chromosomes 1–10 are found in clusters; only 14 occurred as solo copies. Clusters varied in width from single CentC copies to 84 Kb with a mean of ~7 Kb (~ 45 CentC copies). Chromosomes varied greatly in CentC copy number, although centromere assemblies for all of the chromosomes are not complete. For example, CENH3-ChIP sequence from an oat–maize addition line with one maize chromosome (Kynast et al. 2001) has many reads that map to the unassembled BACs (Table S3). In particular, chromosome 6 has many more reads aligning to the unassembled BACs than it did to its own centromere repeats, suggesting a particularly incomplete assembly. Examining total repeat number, chromosome 7 had the most CentC, with 3,200 copies, while chromosome 6 had the fewest with 32 copies.

We used long-read Pacific Biosciences sequencing to verify that most CentC is in tandem arrays. We sequenced whole genome (~0.1 X) libraries from four *Zea* taxa. In spite of the low coverage, we recovered reads containing CentC sequence from all four taxa (Table S2). In one 6.7 Kb read from the maize reference line B73, for example, we identified approximately 40 independent CentC copies in tandem, and similar arrays were seen in all four *Zea* species analyzed. These results show that overall structure of the repeats has been maintained for approximately 140,000 years since the *luxurians–mays* divergence (Hanson et al. 1996; Ross-Ibarra et al. 2009) and that a majority of CentC is found in tandem arrays (Table S2).

We then identified how many large clusters of CentC were retained from each of the two parental genomes that comprise the extant maize genome, referred to here as subgenome 1 and subgenome 2 (Fig. 1). Previous work identified the parental genome for individual chromosomal segments (Schnable et al.

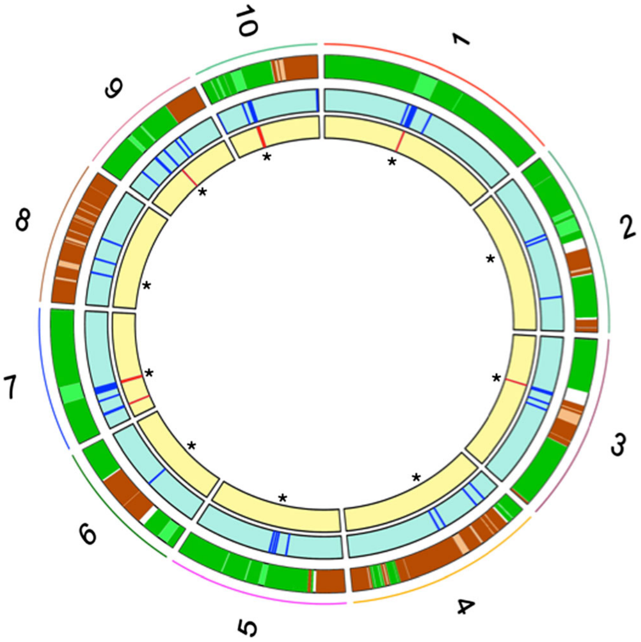

Fig. 1 CentC repeat location in relation to the maize subgenomes. The *outer ring* depicts chromosomal assignment to the two subgenomes, with higher confidence regions in *darker colors*. *Green* corresponds to subgenome 1 and *brown* to subgenome 2. Breakpoints between the subgenomes remain *uncolored* to indicate uncertainty. The *middle ring*, shaded in *blue*, displays the locations of all CentCs across the genome. The *inner ring*, shaded in *yellow*, displays the locations of all CentC clusters greater than 20 Kb in length. *Asterisks* indicate approximate location of the centromere

2011) and centromeres (Wang and Bennetzen 2012), with centromeres 1, 2, 5, 7, 9, and 10 deriving from subgenome 1 and 3, 4, 6, and 8 from subgenome 2. Due to the maize iMap resource (Zhou et al. 2009), large clusters should be less likely to be misplaced within the genome. Therefore, we focused our analyses on the 52 clusters >10 Kb in length (Figure S3). We assign clusters to a subgenome if they are flanked by two regions identified as originating from the same subgenome. Thirty-eight of these clusters could be assigned to subgenome 1 (out of 43 assignable). If we restrict the analysis to clusters > 20 Kb with clear assignment to one subgenome, all 16 clusters were found in subgenome 1. Even correcting for the genome-wide overrepresentation of subgenome 1 (64.7 % of assigned base pairs), these results suggest a strong inequality in the origin of large CentC clusters (Fisher's exact test, $p<0.005$ for total CentC, 10 Kb, and 20 Kb clusters). One cluster > 20 Kb falls within an unassigned region on chromosome 3. This difference between the subgenomes is robust to different criteria of cluster size and distance (Table S4).

Previous studies have also described the Cent4 repeat, a tandem pericentromeric repeat that occurs primarily on chromosome 4 (Page et al. 2001). Available evidence does not point to any centromeric function for Cent4/CenH3 chromatin. Immunoprecipitation data (Wolfgruber et al. 2009; Jin et al. 2004) shows no significant overrepresentation of Cent4 compared to five known non-centromeric TEs, and fiber FISH shows clear separation of Cent4 from centromeric repeats (Jin et al. 2004). Furthermore, Cent4 probes lag behind CentC probes in cell division, suggesting that they are not found in the kinetochore (Jiang et al. 2002; Jin et al. 2004). BLAST analyses of Cent4 sequences from GenBank revealed high homology to the poorly characterized long terminal repeat (LTR) retrotransposon RLX_sela that was previously shown to be associated with heterochromatic knobs (Tenaillon et al. 2011; Chia et al. 2012), but Cent4 lacks any of the protein sequences necessary for autonomous transposition, such as GAG and POL complexes. But while previous work in rice has documented the presence of nonautonomous LTR retrotransposons in or near the centromere (Jiang et al. 2002), RLX_sela also appears to be missing the necessary primer binding sites that would distinguish it as a nonautonomous TE, suggesting that it may be a TE-derived tandem repeat unique to the pericentromere of chromosome 4.

Relatedness of CentC in the maize genome

CentC copies in the maize genome exhibit tremendous diversity: the overall pairwise identity in our alignment was only 65 % and ~ 98 % of sites in the alignment had at least two variants. Such diversity led us to ask whether genetic groups of CentC variants could be distinguished. We performed principle coordinate analyses (PCoA) from a genetic distance matrix estimated from our alignment, and assigned individual repeats to genetic clusters following the approach of (Patterson et al. 2006). We found 58 significant clusters, but observed no pattern of groupings that revealed chromosome specificity of CentCs or the impact of historical tetraploidy (Fig. 2; Table S5).

The tandem nature of CentC suggests it increases in copy number through local duplications that produce initially identical copies. Similar conclusions were found by Ma and Jackson (2006). Tandem duplications can occur through a variety of means, including slippage of the DNA polymerase or recombination that could lead to unequal crossing over or gene conversion. Tandem duplication predicts that clusters of CentC should be more closely related than CentC from different clusters. Comparisons of genetic and physical distance among CentC repeats on chromosomes 2 and 5 shows that average genetic similarity is highest within clusters (Fig. 3), revealing significant spatial autocorrelation of CentC variants over distances up to 10–50 Kb (Figure S4 and S5).

The decreased genetic distance among CentCs in local clusters on chromosomes 2 and 5 suggest that many of the genetic groupings discovered in our genome-wide analysis should correspond to local clusters of repeats. However, repeats within individual clusters are frequently found in different genetic groups as defined by principle coordinate analysis (Fig. 2). We

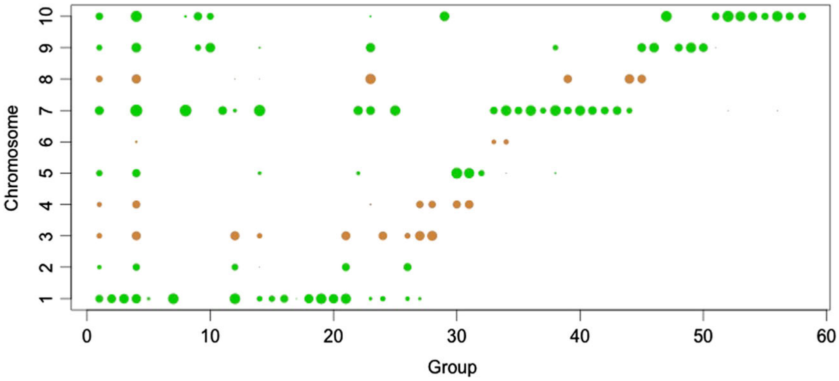

Fig. 2 Presence of CentC in each of the hierarchical groups. The 58 clusters found to be statistically significant in forming genetic groups are represented on the *x*-axis and chromosome of origin on the *y*-axis. The size of each point is proportional to the log number of sequences in that group on that chromosome. CentC counts from chromosomes whose centromeres were derived from subgenome 1 are colored *green* and those from subgenome 2 are colored *brown*

observed multiple pairs of CentC which occur in the same genetic cluster in spite of being separated on the completely sequenced centromeres 2 and 5, suggesting that our result is not simply an artifact of errors in assembly. A comparison of shared mutations across all pairs of CentC sequences reveals a potential explanation. Of the ~74 million possible pairs, approximately 6 million share ≥2 mutations different from the genome-wide consensus, causing CentC copies to group with sequences that share mutations irrespective of their physical distance. Comparing several triplets at random from our alignment confirms that two sequences in one PCoA assignment share greater pairwise identity than two sequences adjacent to one another in different PCoA groups. A simple forward simulation (see "Methods") suggests this pattern could be due entirely to homoplasy rather than long distance movement of CentC repeats. By stochastically applying mutations to an initially homogeneous group of repeat sequences over the time period since divergence from *Tripsacum*, we find that plausible parameter values produce ~10 million pairs of repeats sharing ≥2 mutations.

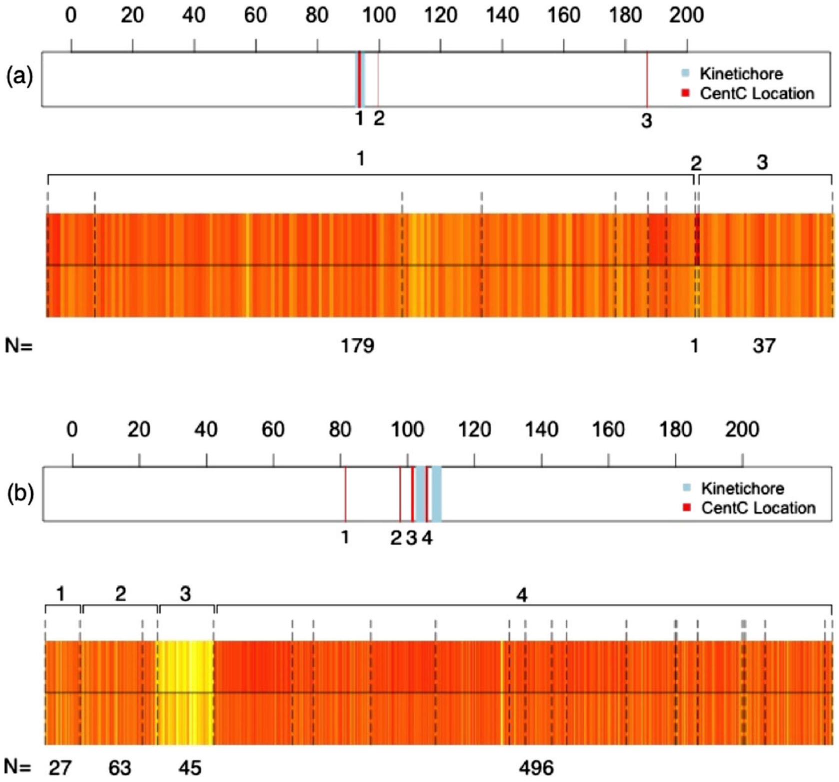

Fig. 3 CentC physical location and genetic relatedness for **a** chromosome 2 and **b** chromosome 5. On the physical map above, *red lines* show locations of numbered CentC clusters and *blue blocks* show the location of the active kinetochores. *Scale bar* is in Mb. Below each physical map is shown a heatmap of genetic relatedness of each CentC to (*top row*) other copies within its island of tandem repeats, delineated by dotted lines, and (*bottom row*) all other copies on the chromosome. *Darker colors* indicate higher relatedness. The total number of CentC in each cluster is shown below the map

Variation of CentC abundance in *Zea*

Shotgun sequence data from the maize HapMap v2 (Chia et al. 2012), reveals a significantly greater abundance of CentC in teosinte than in domesticated maize ($p<0.01$; Fig. 4). Further support for differences between maize and its wild relatives comes from additional sequence from *Z. luxurians* (Tenaillon et al. 2011). Analysis of these data find nearly twice as much CentC in *Z. luxurians* as the maize inbred B73. To corroborate these results, we performed FISH of F1 crosses between inbred maize and teosinte to determine if cytological observations agreed with our sequencing findings. FISH data supports our observation that the teosintes *Z. mays* ssp. *parviglumis* and *Z. luxurians* have more CentC than inbred maize (Fig. 4). Using whole genome shotgun PacBio long reads, we further investigated the overall

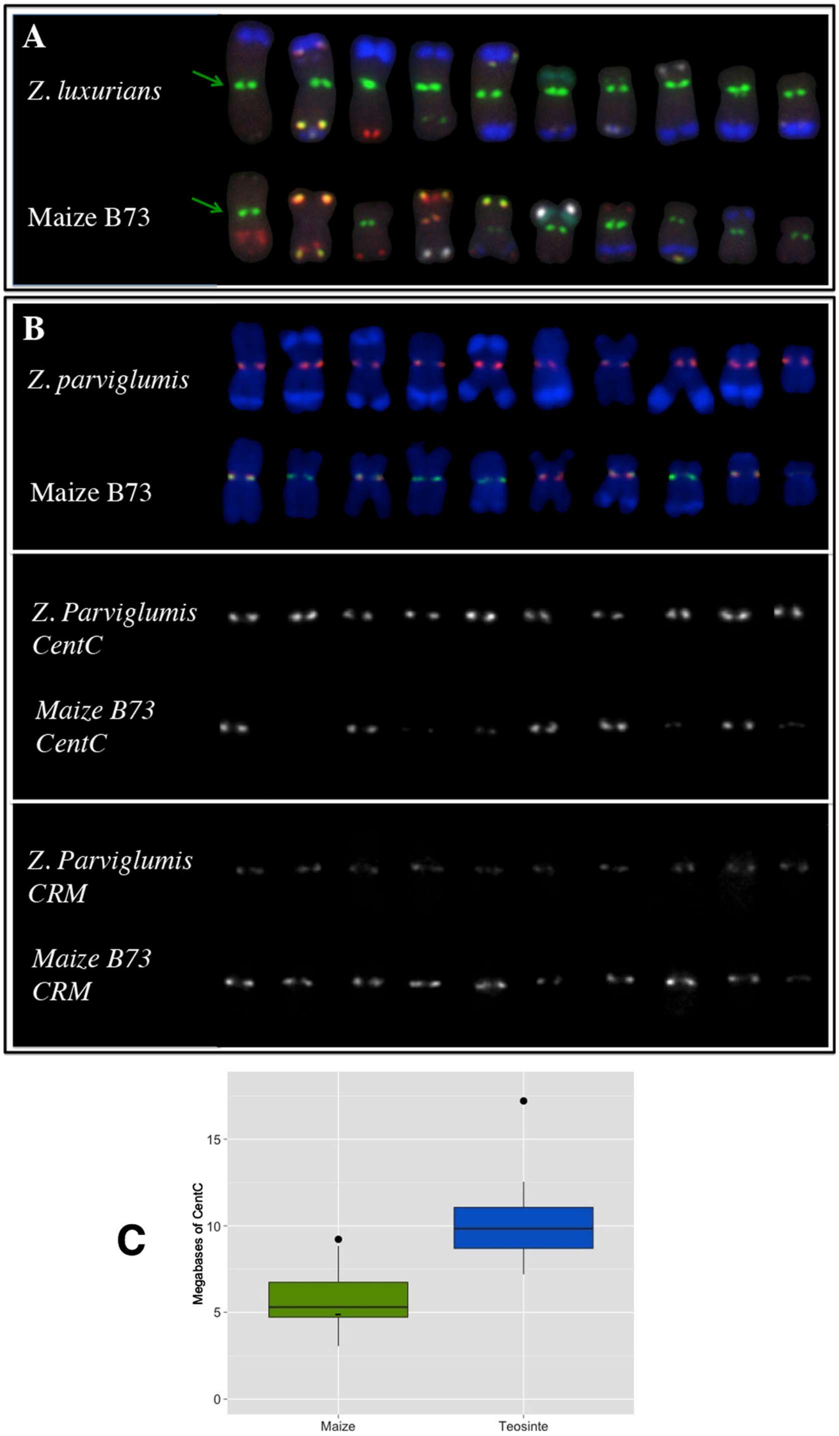

Fig. 4 **a** FISH analysis of a single individual heterozygous for B73 and *Zea luxurians* (GRIN accession PI422162). Chromosomes, ordered from 1 (*left*) to 10 (*right*), were hybridized with the Kato et al. (2004) probe cocktail. Green shows CentC and a 4-12-1 subtelomere repeat, *blue* the 180 bp knob repeat, *red* the abundant TAG microsatellite and another subtelomeric repeat, *white* the TR1 knob repeat, orange the Cent4 repeat, *yellow* the 5S rDNA repeat, and *aqua* shows the Nucleolus Organizer Region. The *green signals* at the primary constrictions (*arrows*) are CentC. Note that *Z. luxurians* have far brighter CentC signals. This image was graciously provided by Patrice Albert. **b** FISH analysis of a single individual heterozygous for B73 and *Z. mays*-ssp. *parviglumis* (GRIN accession PI566687). Chromosomes were hybridized with the (Shi et al. 2010) probe cocktail showing CentC in *red* and CRM2 in *green*. Each separate probe is shown separately below the two-color image, highlighting that CentC is more abundant in *Z. mays* ssp. *parviglumis* and CRM2 is more abundant in maize. **c** Mb of CentC in genomic libraries of maize and teosinte. *Box plots* show data from (Chia et al. 2012). Points show data for maize inbred B73 and the teosinte *Z. luxurians* from (Tenaillon et al. 2011). For comparison, the data point of maize inbred B73 in Chia et al. (2012) is shown with a *tick mark* on the box plot

structure of repeats across the different *Zea* species. Percentages of the libraries showing tandem repeats were also higher in PacBio sequences from the three teosintes compared to B73 (Table S2).

Discussion

Our analysis of centromere repeat diversity across the maize genome identifies thousands of copies exhibiting tremendous diversity. But while we can cluster the repeats into groups of related sequences, these groups have little relation to current or ancient maize chromosomes (Fig. 2). We find no evidence of chromosome specific repeats as observed in *Arabidopsis* species (Kawabe and Nasuda 2005; Pontes et al. 2004), suggesting the presence of a mechanism that homogenizes repeats across centromeres on different chromosomes. Although we believe our results relatively robust to assembly errors, misplaced BACs and collapsed tandem copies almost certainly occur in the maize reference genome and may influence our assessment of chromsome-wide patterns. We further verify that Cent4, once thought to be a chromosome-specific centromere repeat (Page et al. 2001), appears to be a poorly characterized tandem repeat or nonautonomous retroelement, but is not associated with the centromere.

We find that virtually all the large arrays of CentC in the maize reference genome derived from one of the two ancestral genomes present in modern day *Zea* (Fig. 1, Table S5). This biased ancestry mirrors differences in genic expression and deletion seen between the subgenomes (Schnable et al. 2011). Higher deletion rates on subgenome 2 may explain our observation, but the finding of small regulatory RNAs corresponding to centromeric repeats (Reinhart & Bartel 2002) in other taxa may suggest a more active mechanism behind the observed differences.

Our sequence comparison of CentCs also enabled us to explore the relationship between genetic and physical distance among repeats. Using the well-assembled centromeres on chromosomes 2 and 5, we found spatial autocorrelation of relatedness among repeats, but also observed genome-wide that many CentCs within an array fall into the same genetic cluster. We observed no differences in genetic similarity when comparing clusters inside against cluster outside the active kinetochore. Our observations are consistent with the simple idea that most repeats arise due to tandem duplication or related processes with similar outcomes such as small scale gene conversion or unequal crossing over. Long-distance transposition of CentC, while necessary to homogenize repeats across chromosomes (Shi et al. 2010), appears relatively uncommon.

One unusual result from our sequence comparison was the finding that pairs of CentC on different chromosomes share high sequence similarity. Our simulations suggest that, under realistic assumptions about mutation rate and divergence time, such a pattern is possible due to homoplasious mutation alone. Roughly 80 % of the CentC repeats have their closest genetic relative on the same chromosome, as expected under a model of tandem duplication, but only 14 % of closest genetic pairs are found within 10 Kb of each other. Though assembly errors may explain a portion of these relationships, we find several closest genetic pairs separated in the fully sequenced centromeres of chromosomes 2 and 5, suggesting our observation is biologically real. We speculate that the vast majority of CentCs in the genome are thus a result of relatively old tandem duplications, and that sufficient time has occurred since duplication for rearrangements and mutations to break up patterns of identical tandem repeats.

Previous cytogenetic work identified differences in centromere repeat content between domesticated maize and its wild relatives *Z. mays* ssp. *parviglumis*, *Z. mays* ssp. *mexicana*, and *Z. luxurians* but was unable to quantify differences (Albert et al. 2010). Our resequencing results show that while there is little difference in the distribution of CentC in tandem arrays, the absolute abundance of CentC has decreased during domestication, and we verify this with FISH in two maize–teosinte hybrid individuals (Fig. 4).

Variability in observed abundance of transposable elements (Chia et al. 2012) suggests that the decrease seen in CentC is not due to causes common to all repetitive sequences. The maize genome is smaller than its teosinte counterpart, largely due to differences in the abundance of heterochromatic knobs (Poggio et al. 1998). Zhang and Dawe (2012) have postulated an adaptive relationship between centromere size and genome size based on an observed correlation between centromere size and genome size across a number of grass species. A model correlating centromere size to total genome size would propose that the decrease in CentC abundance seen post-domestication is due to selection for smaller active centromeres to complement the smaller overall genome size. While our current data are insufficient to evaluate this conclusion, future work investigating differences in CentC content among maize landraces that vary in genome size (Poggio et al. 1998) may provide an opportunity to further test our hypothesis.

In conclusion, our detailed study of centromere repeats in the B73 maize genome has highlighted differential contribution of subgenome, spatial autocorrelation along a chromosomes, and changes in abundance over the short time scale of domestication.

Since maize lines show vast cytogenetic variation, further work evaluating CentC evolution across multiple populations and multiple taxa may shed additional light on the timing and causes of these changes.

Acknowledgments We wish to thank Pacific Biosciences for sequencing resources. We thank Patrice Albert and James Birchler for providing the high-quality FISH image of the *Z. luxurians* X B73 hybrid. We thank Siddharth Bhadra-Lobo, Vince Buffalo, Anne Lorant, Gernot Presting, Lauren Sagara, Michelle Stitzer, and National Science Foundation summer exchange program interns Cesar Alvarez Mejia, Aurelio Hernandez Bautista for their advice and helpful discussion. This project was funded by National Science Foundation grant IOS-0922703.

References

Albert P, Gao Z, Danilova T, Birchler J (2010) Diversity of chromosomal karyotypes in maize and its relatives. Cytogenet Genome Res 129(1–3):6–16

Ananiev EV, Phillips RL, Rines HW (1998) Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. Proc Natl Acad Sci 95(22):13,073–13,078

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC (2012) Maize hapmap2 identifies extant variation from a genome in flux. Nat Genet 44(7):803–807

Clark RM, Tavaré S, Doebley J (2005) Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Mol Biol Evol 22(11):2304–2312

Edgar RC (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797

Felsenstein J (1989) Phylip—phylogeny inference package (version 3.2). Cladistics 5(2):164–166

Hall SE, Luo S, Hall AE, Preuss D (2005) Differential rates of local and global homogenization in centromere satellites from Arabidopsis relatives. Genetics 170(4):1913–1927

Hanson MA, Gaut BS, Stec AO, Fuerstenberg SI, Goodman MM, Coe EH, Doebley JF (1996) Evolution of anthocyanin biosynthesis in maize kernels: the role of regulatory and enzymatic loci. Genetics 143(3):1395–1407

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2(4):618–620

Jiang N, Bao Z, Temnykh S, Cheng Z, Jiang J, Wing RA, McCouch SR, Wessler SR (2002) Dasheng: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. Genetics 161(3):1293–1305

Jin W, Melo JR, Nagaki K, Talbert PB, Henikoff S, Dawe RK, Jiang J (2004) Maize centromeres: organization and functional adaptation in the genetic background of oat. Plant Cell Online 16(3):571–581

Kato A, Lamb JC, Birchler JA (2004) Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc Natl Acad Sci 101(37):13,554–13,559

Kawabe A, Nasuda S (2005) Structure and genomic organization of centromeric repeats in *Arabidopsis* species. Mol Genet Genomics 272(6):593–602

Kynast RG, Riera-Lizarazu O, Vales MI, Okagaki RJ, Maquieira SB, Chen G, Ananiev EV, Odland WE, Russell CD, Stec AO, Livingston SM, Zaia HA, Rines HW, Phillips RL (2001) A complete set of maize individual chromosome additions to the oat genome. Plant Physiol 125(3):1216–1227

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9(4):357–359

Ma J, Jackson SA (2006) Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. Genome Res 16(2):251–259

Macas J, Neumann P, Novák P, Jiang J (2010) Global sequence characterization of rice centromeric satellite based on oligomer frequency analysis in large-scale sequencing data. Bioinformatics 26(17): 2101–2108

McGinnis S, Madden TL (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32(suppl 2):W20–W25

Melters D, Bradnam K, Young H, Telis N, May M, Ruby J, Sebra R, Peluso P, Eid J, Rank D, Garcia J, DeRisi J, Smith T, Tobias C, Ross-Ibarra J, Korf I, Chan S (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14(1):R10

Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones KM, Dawe RK, Buell CR, Jiang J (2003) Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics 163(2):759–770

Page BT, Wanous MK, Birchler JA (2001) Characterization of a maize chromosome 4 centromeric sequence: evidence for an evolutionary relationship with the b chromosome centromere. Genetics 159(1): 291–302

Paradis E, Claude J, Strimmer K (2004) Ape: analyses of phylogenetics and evolution in R language. Bioinformatics 20(2):289–290

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12):e190

Perry P, Johnstone I, Ma Z, Shahram M (2009) Rmtstat: distributions and statistics from random matrix theory. 2009. R software package version 01

Poggio L, Rosato M, Chiavarino AM, Naranjo CA (1998) Genome size and environmental correlations in maize (*Zea mays* ssp. *mays*, Poaceae). Ann Bot 82(suppl 1):107–115

Pontes O, Neves N, Silva M, Lewis MS, Madlung A, Comai L, Viegas W, Pikaard CS (2004) Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid *Arabidopsis suecica* genome. Proc Natl Acad Sci 101(52):18,240–18,245

Reinhart BJ, Bartel DP (2002) Small RNAs correspond to centromere heterochromatic repeats. Science 297(5588):1831–1831

Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical divergence and gene flow in the genus *Zea*. Genetics 181(4):1399–1413

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956):1112–1115

Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci 108(10):4069–4074

Sharma A, Presting GG (2014) Evolution of centromeric retrotransposons in grasses. Genome Biol Evol. doi:10.1093/gbe/evu096

Sharma A, Schneider KL, Presting GG (2008) Sustained retrotransposition is mediated by nucleotide deletions and interelement recombinations. Proc Natl Acad Sci 105(40):15,470–15,474

Shi J, Wolf SE, Burke JM, Presting GG, Ross-Ibarra J, Dawe RK (2010) Widespread gene conversion in centromere cores. PLoS Biol 8(3): e1000327

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J (2011) Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. Genome Biol Evol 3:219

Wang H, Bennetzen JL (2012) Centromere retention and loss during the descent of maize from a tetraploid ancestor. Proc Natl Acad Sci 109(51):21,004–21,009

Wang K, Wu Y, Zhang W, Dawe RK, Jiang J (2013) Maize centromeres expand and adopt a uniform size in the genetic background of oat. Genome research pp gr–160,887

Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, Shi J, Gao Z, Han F, Lee H, Xu R (2009) Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. PLoS Genet 5(11): e1000743

Zhang H, Dawe RK (2012) Total centromere size and genome size are strongly correlated in ten grass species. Chromosom Res 20(4):403–412

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, Birchler JA, Jiang J, Dawe RK (2002) Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. Plant Cell Online 14(11):2825–2836

Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S et al (2009) A single molecule scaffold for the maize genome. PLoS Genet 5(11):e1000711