

Diversity and evolution of centromere repeats in the maize genome

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

Received: date / Accepted: date

Abstract Centromere repeats are found in most eukaryotes and play a critical role in kinetochore formation. Though they 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 CentC repeat. We first validate the existence of long tandem arrays of repeats in maize and other taxa in *Zea*. We show that genetic similarity among CentC copies is highest within these arrays, suggesting that tandem duplications are the primary mechanism for the generation of new copies. In spite of this, we see little evidence that CentC variants form distinct genetic groups, instead finding that homoplasious mutations have likely homogenized CentC diversity. We identify Cent4, a repeat initially thought to be specific to the chromosome 4 centromere, as an LTR retrotransposon that has increased drastically in frequency through domestication. Although the two ancestral subgenomes of maize have contributed nearly equal numbers of centromeres, our analysis shows that a vast

Paul Bilinski : Kevin Distor : Jose Gutierrez-Lopez : Gabriela Mendoza Mendoza : Jeffrey Ross-Ibarra
Department of Plant Sciences, UC Davis
Davis, California, 95616

Jeffrey Ross-Ibarra
The Genome Center and Center for Population Biology, University of California
Davis, California, 95616
E-mail: rossibarra@ucdavis.edu

Jinghua Shi : R. Kelly Dawe
Department of Plant Biology, University of Georgia
Athens, Georgia, 30602

R. Kelly Dawe
Department of Genetics, University of Georgia
Athens, Georgia, 30602

majority of all CentC repeats derive from only one of the parental genomes. Finally, by comparing maize with its wild progenitor teosinte, we find that the abundance of CentC has decreased through domestication.

I deleted the phrase talking about rates because we dont really do/say anything about rate do we?

Keywords First keyword · Second keyword · More

Introduction

In spite of the rapid growth of sequenced plant genomes, plant centromeres remain poorly understood and relatively cryptic, due largely to their highly repetitive content. Centromere repeats are highly diverse across taxa and their turnover appears to be very rapid (?). 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 (?) including complete sequence of two centromeres (**cite wolfgruber here?? or is schnable enough?**), and the breadth of research into maize centromeres makes it one of the best systems in which to investigate the processes governing centromere repeat evolution. **The line about old nonfunctional repeats is a good point, but we don't look for these or really say much about function.**

Maize centromeres are comprised primarily of the 156bp satellite repeat CentC and a family of retrotransposons (CRM), both of which interact with kinetochore proteins such as CENH3 (**please do citations as shown above. just put the bibtex version of the citation from bookends in the refcentc.bib file.** Wolfgruber et al 2009, Zhong et al 2002), and both repeats show considerable variation in local abundance across taxa (Albert et al 2010). But while there is considerable effort in investigating the molecular function of maize centromere repeats (Presting Chapter and references therein? **here. citing and references therein is sort of weak scholarship. yes, i have been guilty of it myself.**), we know comparatively little about the evolution of the sequences themselves. CRM elements are better understood, including the age and insertion preferences of different CRM families (Wolfgruber et al 2009) **shouldn't we cite Anupma's PNAS paper here too?**. In contrast, studies **what studies? reference?** have only examined the flanking sequences to CentC islands despite their known association with functional centromeres. To date, there is no in-depth characterization of the genetic diversity of centromere repeats in the maize genome.

In this paper, we describe the patterns of diversity of centromere repeats across the maize genome. We investigate the genetic diversity of these repeats, including whether the differential ancestry of maize centromeres (Wang and Bennetzen 2012) has led to chromosome-specific repeats as found in other species (Heslop-Harrison 1999, Hirsch et al 2009) and how genetic relatedness among individual repeats varies spatially along chromosomes and across the genome. We then compare centromere abundance across a number of maize lines, including to its wild relatives the teosintes. We find that CentC copies do not form genetic groups consistent with ancient whole genome duplications

or chromosome specificity. Instead, we show higher genetic similarity within clusters, potely indicating the predominance of tandem duplications in the formation of new CentC copies. Lastly, we use low coverage sequencing and cytological analyses datshow that domesticated maize has less CentC and reli its wild relative teosinte line

Methods

CentC Repeat Identification and Diversity

We downloaded 218 previously annotated CentC sequences (Annaniev et al 1998, Nagaki et al 2003) from Genbank. We then searched the maize genome (5b60, www.maizesequence.org) with megaBLAST (McGinnis and Madden 2004) using the 218 annotated CentCs as a reference. We kept hits with a length of over 140bp and a minimum bit score of 100. After meeting the bit score threshold, the longest hit was retained. We defined CentCs as being in tandem if the CentCs start location was within 1000bp of the start location of another CentC.

All 12,162 CentC sequences were aligned using 7 iterations of Muscle (Edgar 2004) with default parameters. A Jukes-Cantor distance matrix of all sequences was calculated with PHYLIP (Felsenstein 2005, <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 (c.f. Patterson et al 2006).

We employed the software SpaGeDi (<http://ebe.ulb.ac.be/ebe/Software.html>, Hardy and Vekemans 2002) to estimate the spatial autocorrelation of sequence similarity of CENTC repeats in the completely sequenced centromeres 2 and 5. We calculated Morans 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, Claude, and Strimmer 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 species (??) to a reference consisting of the full complement of 12,162 CentC variants identified

in the B73 genome. We also used previously published whole genome chromatin immunoprecipitation (ChIP) (Wolfgruber et al 2009, Wang et al 2013) with CenH3. Reads were mapped using 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 (?), 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 (Supplementary Figure X [please start filling these in](#)). The parameter boundary at which percentage of reads mapping increased disproportionately was identified, and parameters just below this threshold were selected. Our final set of parameters for tandem repeats used an initial hash size of 8, an alignment candidate threshold of 15 bases, 20% percent 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 ? and ?. [note difference here in effect of citep and citet commands](#) 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. Because library preparation has an effect on estimates of repeat abundance (see results), we only used individuals from maize HapMap v2 (?) with libraries prepared using identical methods.

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 (maizetdb.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.

CentC simulations

[needs some basic setup. we start with alignment \(?\) and then do what to it? what was the code written in? is the code available? if so cite github url.](#) Our simulation assumed that CentC had been evolving since the divergence of maize and *Tripsacum* (10 million years, Ross-Ibarra et al 2009)[that's not the date i cite in my paper! should be 1 million.](#), a closely related genus whose cen-

centromere repeat shares a large degree of homology (?). We assumed a constant copy number, a mutation rate of 3×10^{-8} , and one generation per year.

Sequencing

Library preparation and sequencing was performed according to the methods cited in Melters et al 2012. Using those protocols, we sequenced one individual from *mays*, *mexicana*, *parviglumis*, and *Z. luxurians* with Pacific Biosciences ([cite](#)) technology. Approximately 200Mbp of reads were produced from each cell, and reads with length greater than 600bp were retained for analysis of tandem CentC content using BLAST (Supplementary Table 1 [check out how tables are done and referenced with this example](#)). CentC copies were considered in tandem if the read had 4 CentC copies within 300bp of each other.

FISH

Flourescent in situ hybridization (FISH) was used to compare abundance of CENTC repeats in chromosomes of a maize x *parviglumis* F1 hybrid and a maize x *Z. luxurians* F1 hybrid. FISH protocols closely followed those of X and Y. [Need Kelly to fill in](#)

Results

Centromere repeats in the maize genome

We found a total of 12,162 CentC variants in the maize reference genome and the unassembled BACs. Of these 12,162 copies, 9,896 were unique over their full length. No CentC occurred more than 10 times in the genome, and the vast majority ($>X\%$ [if its easy, this number would be nice.](#)) of non-unique CentC variants occurred only twice. Of the 2,266 non-unique CentC sequences, only 3 were tandem duplicates. Genome-wide CentC locations also show that nearly all of the 10,639 CentC copies are found in one of the 248 clusters identified on chromosomes 1-10; only 14 occurred as solo copies. Cluster size varied from single CENTC copies to 84KB, with a mean of 7kb (approximately 45 CentC copies). Chromosomes varied greatly in copies of CentC. Chromosome 7, with 3,200 copies, had the most CentC, while chromosome 6 had the fewest with 32 copies. The presence of 1,523 copies of CentC on the unassembled BACs and so few copies on chromosome 6 suggest that Centromeric regions still require further assembly.

We used long-read Pacific Biosciences sequencing to verify that most CentC is in tandem arrays. We sequenced whole genome ($\approx 0.1X$) libraries from 4 Zea species. In spite of the low coverage, we recovered reads containing CentC sequence from all four taxa (Supplementary Table [add](#)). In one 6.7kb 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 the 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 (add table ref). this last sentence is a nice addition

seems like in this paragraph we should cite the circos figure, no? We then identified how many large clusters of CentC were retained from each of the two parental genomes that comprise the extant maize genome. Previous work identified the parental genome for individual chromosomal segments (Schnable et al 2011) and centromeres (Wang and Bennetzen 2012). Because large clusters are less likely to be misassembled, we focused our analyses on the 52 clusters >10KB in length (Supplementary Fig Histogram add). While Schnable et al. (2010) do not assign subgenome status to repetitive regions, 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); of the 17 clusters >20KB, 16 were unambiguously assigned to subgenome 1. Even correcting for the genome-wide overrepresentation of subgenome 1 (62.5% of assigned base pairs), these results suggest a strong inequality in the origin of large CentC clusters (binomial test, $p < 0.005$ for both 10kb and 20kb clusters). isn't this a fisher's exact?

In addition to CentC, previous studies have described a second high copy number centromere repeat specific to maize chromosome 4 (Page et al 2001). BLAST analyses of Cent4 sequences, however, revealed that all 15 had high homology to the poorly characterized LTR retrotransposon RLX_sela, which was previously shown to be associated with heterochromatic knobs (Tenaillon et al 2011, Chia et al 2012), suggesting that Cent4 is actually a pericentromeric retrotransposon.

Whole-genome cenH3 chromatin immuno-precipitation data from Wolfgruber (et al 2009) shows no significant over-representation of cent4 compared to five known non-centromeric TEs, suggesting the cent4 repeat is not involved in kinetichore formation on chromosome 4. the jin result is very cool and i wasn't aware of. i moved to discussion though.

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 2 variants . Such diversity led us to ask whether genetic groups of CentC variants could be distinguished. We performed principle coordinate analyses 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

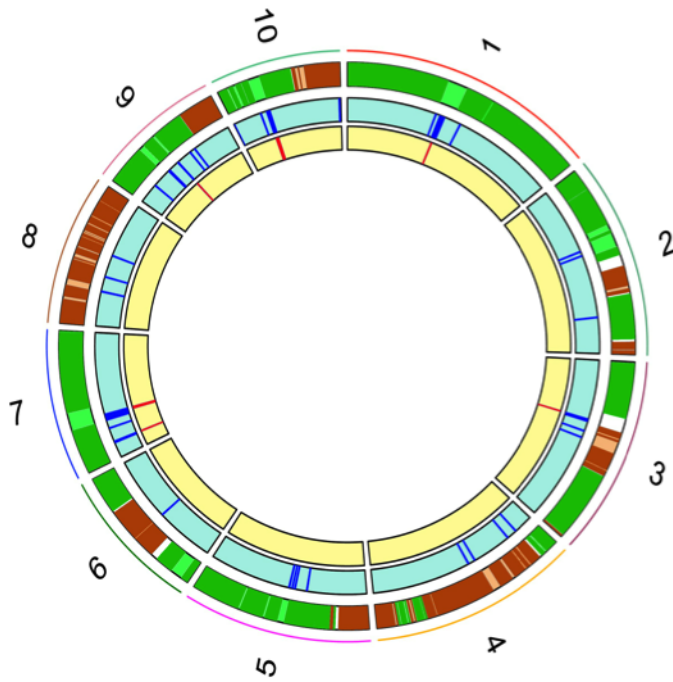


Fig. 1 CentC repeat location in relation to the maize subgenomes. High confidence regions are colored with darker colors while low confidence regions are colored with lighter colors. Breakpoints between the subgenomes remain uncolored to indicate uncertainty.

pattern of groupings that revealed chromosome specificity of CentCs or the impact of historical tetraploidy (Supplemental Table X [add](#)).

The tandem nature of CentC suggests it increases in copy number through local duplications that produce initially identical copies. This predicts that local clusters of CentC should be more closely related than CentC from different clusters. Analysis of genetic and physical distance among CentC repeats on chromosomes 2 and 5 indeed shows this pattern (Figure 2), revealing significant spatial autocorrelation of CentC variants over distances up to 10 KB (Supplementary Figure SPAGED1 [add](#)). [need to standardize usage of kb. also i thought autocorrelation was significant up to larger distances?](#)

The decreased genetic distance among CentCs in local clusters on chromosome 2 and 5 suggest that many of the genetic groupings discovered in our genome-wide analysis should correspond to local clusters of repeats. We see no evidence of this, however, as repeats within individual clusters are frequently found in different genetic groups as defined by PCoA (Supplemental Table X [add](#)). A comparison of all pairs of CentC reveals a likely explanation: of the ≈ 74 million possible pairs, approximately 6 million [this said 10, isn't it 6 observed 10 simulated??](#) share ≥ 2 mutations different from the genome-wide consensus, likely grouping in genetic clusters despite their physical distance.

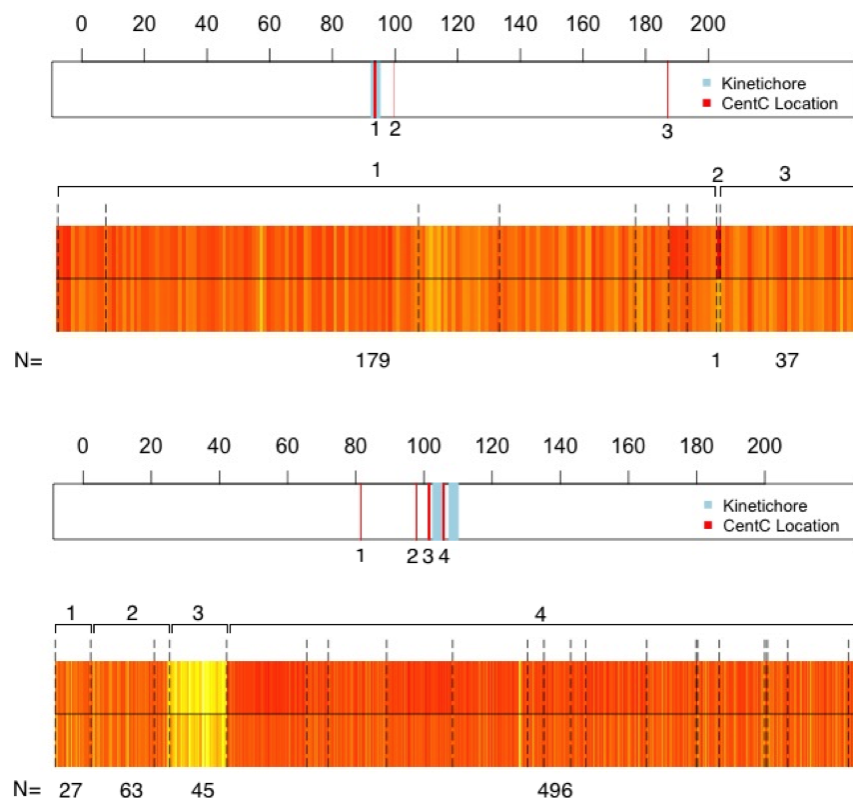


Fig. 2 CentC physical location and genetic relatedness for (a) chromosome 2 and (b) chromosome 5. On the physical map, red lines show locations of numbered CentC clusters and blue blocks show the location of the active kinetichores. Scale bar is in Mb. Below each physical map is shown a heatmap of genetic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

have you verified any of these? that is, take some centcs in a single genetic group, find one that isnt from the same physical cluster, and show that its in that genetic group because it shares some non-consensus mutations that are different from those repeats in its own physical cluster? A simple forward simulation (see) 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, we find that plausible parameter values produce ≈ 10 million pairs of repeats sharing ≥ 2 mutations.

Variation in CentC abundance in Zea

Shotgun sequence data from the maize HapMap v2 ??, reveals a significantly greater abundance of CentC in teosinte than in domesticated maize ($p < 0.01$; Fig. boxplot). Teosintes have more CentC than inbred maize. Further support for differences between maize and its wild relatives comes from shotgun sequence data 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 fluorescent in-situ hybridization of F1 crosses between inbred maize and teosinte to determine if cytological observations agreed with our sequencing findings. Cytology needed to be performed on an F1 cross so that we could compare relative probe fluorescence of the chromosomes within a single individual. FISH data supports our observation that the teosintes *parviglumis* and *Z. luxurians* have more CentC than inbred maize (Figure 3) [add in this figure, and better quality images of first two](#).

Discussion

Our study traces the changes in centromere repeat genetic relatedness across maize chromosomes to show how ancient tetraploidy and subsequent evolution has impacted centromeres. Most interestingly, we show that most large arrays of CentC in the maize reference genome derive from maize1 subgenome, which is known to have lower gene loss and higher average expression of its genes than maize subgenome 2. Regulatory microRNAs are known to correspond to centromeric repeats (Reinhart and Bartel 2002). Therefore, it may be worthwhile to further investigate the potential correlation between centromere retention and gene expression as the phenomenon of higher expression of genes from one ancestor may be very common across plants (A. thaliana Cheng et al 2012 and Gossypium Kapp and Wendel unpublished). [Two comments about this: first, its a terrible sentence so help make this better, because i think the point is really nice. second, the best citations i found for cotton was an unpublished ref that is still not published from what i can gather. advice? e.g. <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0036442> also check out papers that cite Schabs: \[http://scholar.google.com/scholar?cites=5313714612108479532&as_sdt=2005&sciodt=0,5&hl=en\]\(http://scholar.google.com/scholar?cites=5313714612108479532&as_sdt=2005&sciodt=0,5&hl=en\)](#) There are several possible explanations for the differential contribution of subgenome 1 to the diversity of CentC in the extant maize genome. We put forth several hypotheses that may explain the strong bias for CentC to be located within a block from subgenome 1. Subgenome 2 may have had severely reduced quantities of CentC and therefore never contributed equally. The CentC from subgenome 2 may have decayed rapidly and thus less of it remains. Lastly, the CentC from subgenome 1 may have expanded in copy number since the allopolyploidy event. Though we are unable to conclusively identify the reason, the lack of a large number of identical tandem duplicates and lack of close proximity highest genetic relatedness pairs suggests that any large scale expansion of CentC has not been recent. Also, CentCs in subgenome

2 cluster do not appear to have accumulated excess mutation, suggesting that they are not decaying more rapidly than their subgenome 1 counterparts. We therefore believe that it is most likely that the subgenomes had an unequal contribution of CentC at the formation of the allopolyploid, though more research about the ancient parents is required.

We also explored how genetic relatedness between CentCs correlated with location along the physical map. According to our PCoA analyses, CentC copies across the genome do not form distinct genetic groups correlating with past origin. The lack of subgenome or chromosome grouping in modern CentC may be a result of homogenization of the repeats since the polyploidy event or a lack of differentiation between the repeats of the ancestors. **the fact that repeats in Trip are the same suggests no differentiation in ancestors right? if they had homogenized (read: changed) since ancestors, why would trip == maize? unless trip and maize both went with the CentC most like one of the 2 ancestors** Given the lack of a genome wide pattern, we also wanted to investigate genetic relatedness on individual chromosomes. We hypothesized that most copies of CentC arose from local duplication rather than transposition, and therefore the genetic distance between CentCs would be lowest across CentCs within a large tandem array. **This part is ugly, and needs help** From the reference genome, we chose to investigate chromosomes 2 and 5, since they have been sequenced from end to end, allowing for inferences about the relationship between diversity and physical proximity (Wolfgruber et al 2009). Furthermore, the role of the repeat arrays on chromosomes 2 and 5 appear very different, as the largest array on chromosome 2 interacts with the kinetochore, while no array on chromosome 5 does [cite](#). We find that repeats on chromosomes 2 and 5 are most highly genetically related to neighboring copies (Figure Heatmap). This relationship was recapitulated in analyses using SpaGeDi, where CentCs within approximately 10Kb of one another were more genetically similar (Supplemental Data). Though we did not investigate the relationship between CentC location and genetic similarity on the other chromosomes due to incomplete sequencing of centromeres, we observe that many CentCs within an array fall into the same significant grouping in our Tracy-Widom clustering analysis, suggesting that high local similarity of CentCs is a genome-wide phenomenon. This local relatedness suggests that most CentC is evolving through tandem duplication and not a long distance mechanism such as transposition that had been previously suggested (Shi et al 2010). Concerning chromosomes 2 and 5, the higher local relatedness of all CentC clusters regardless of presence within the kinetochore suggests that CentC interaction with kinetochore proteins does not have a detectable change on their rates of evolution.

When investigating why a pair of chromosomes on two different chromosomes are each other closest genetic relative, we revealed that homoplasy in mutations is common across CentC variants. We suggest that copies of CentC are sufficiently old within the genome that homoplasious mutations cause physically distant CentCs to be highly genetically related. Importantly, we also do not see PCoA group capturing full clusters across chromosomes in a way that

would be consistent with retrotransposition. **So that was a cool point, but I dont know how/if we can test it statisticall? kinda a cool observation.** We speculate that a vast majority of the CentCs exist as a result of very old duplications meaning that mutations have had a long time to accumulate, as only 3 instances of identical tandem copies exist. Roughly 80% of the CentC repeats have their closest genetic relative on the same chromosome, on observation we would expect if CentCs on a chromosome share ancestry. However, only 14% of closest genetic pairs are found within 10KB of each other, suggesting that their tandem duplication is old and that most CentCs have persisted within the genome for a long time. **This is an important sentence, and I cant seem to get it to communicate my point properly. perhaps because im making too many points**

When studying CentC changes through domestication, our findings show that the repeat has experienced a decrease in copy number over time without a noticeable change in the structure of the repeat arrays. Using PacBio long read sequencing, we confirm genome-wide the observation we see on chromosomes 2 and 5, that most copies of CentC exist in large tandem arrays in all Zea taxa. **i think the discussion should largely mirror the results. this doesnt go in the domestication part really. the pacbio is no where in the results, just methods. Ok here in discussion?** From short read sequencing data, we show that modern maize has reduced genomic abundance of CentC when compared to the teosintes, a finding confirmed through FISH (Figure3). Both teosintes within the species mays and its sister species luxurians have elevated levels of CentC. Lower levels of centromere repeats in inbred maize contrasts previous studies that characterized the changing abundance of other common repeats. For example, the abundance of most transposable element families increased after domestication (?). Knowing that TE abundance increases in domesticated maize, we might have expected CentC content to increase as well if centromere size had to expand alongside most other repetitive content. Alternatively, due to their structural role in kinetichore formation, we might have also expected centromere repeats to be largely excluded from the genome wide fluctuations in repetitive content assuming that selection exists to maintain centromere size. Instead, the decrease in CentC content may indicate an active process of removing CentC. We hypothesize that the removal of CentC content may correlate with genome size, as the sharp decrease in knob content through domestication actually led to an overall smaller genome size in inbred maize. The correlation between functional centromere size and genome size has been observed in grass species (Zhang and Dawe 2012). **the problem here is kelly shows functional cent. size, not repeat array size. i dont think this changes the point... CentC, whether in or out the kinetichore, seems to maintain the same relatedness/diversity** Further investigation would be required to discover whether only those copies of CentC outside of the active kinetichore **dude you spell kinetochore wrong throughout whole ms!** are being deleted.

I dont like wrapping up with this... but it makes the most sense. It also does have some cool findings grounded in other works. We also identify the cent4 repeat as the LTR retrotransposon RLX_sela. Characterization of the

TE consensus sequences shows that it contains a 232bp LTR on both ends but lacks any of the protein sequences for autonomous transposition, such as GAG and POL complexes. Furthermore, RLX_sela has [three] **why brackets? is this uncertain?** regions showing > 60bp of homology to the 180bp knob repeat, one of which is approximately 200bp in length **how does a 200bp piece show homology to a 180bp repeat? (I guess it could have indels or repeats) Is this is correct?**. The lack of binding between CenH3 and RLX_sela seen in the ChIP data further suggests that this repeat is likely pericentromeric rather and not involved in centromere function, a result echoed by Jin et al (2004), who showed that Cent4 probes lagging behind CentC probes in cell division. **great find. I didnt know this part.** Previous work in rice **Jimiung's paper is only rice correct?** has documented the frequent presence of nonautonomous LTR retrotransposons in or near **correct to say this?** the centromere (Jiang et al 2002). Given that the centromere on chromosome 4 shows signs of strong selection during domestication (Hufford et al. 2012) **doublecheck i'm not misremembering this please**, it is plausible that the increase in abundance of Cent4 seen in domesticated maize is due to the effects of linked selection on a locus important for domestication rather than novel insertions or duplications.

In conclusion, our study pairs bioinformatics and cytology to validate observations of decreasing centromere repeat content through maize domestication. We show that patterns of CentC similarity are consistent with tandem duplications and maizes allopolyploid history has [left no significant differentiation between CentCs from the different maize subgenomes]. We also identify the chromosome 4 specific repeat, cent4, as the TE RLX_sela and show that it has increased through domestication, unlike a majority of the other TEs (?).

Acknowledgements We wish to thank Pacific Biosciences for sequencing resources. We thank Lauren Sagara, Gernot Presting, NSF summer exchange program interns Cesar, Aurelio, and Siddharth Bhadra-Lobo for helpful discussion.

needs last names

USNSF grant IOS-0922703

Supplemental Material

Table 1 PacBio Read Counts and Tandem CentC

Maize Line	Reads over 600bp	Reads with ≥ 4 CentC	% Reads Showing Tandem CentC
B73	237995	72	0.030252736
<i>luxurians</i>	156964	79	0.050330012
<i>mexicana</i>	141939	150	0.1056792
<i>parviglumis</i>	227050	89	0.039198414