The Genetics and Epigenetics of Maize Centromeres

A Dissertation Submitted to the Graduate Division at University of Hawaii at Mānoa

In Partial Fulfillment of the Requirements for The Doctor of Philosophy Degree

in the Field of Molecular Biosciences and Bioengineering

July 2022

By

Daniel Joseph Laspisa

DISSERTATION COMMITTEE

Gernot Presting, Chairperson

Robert Paull, University Representative

Guylaine Poisson

Michael Kantar

Clifford Morden

Michael Muszynski

This work is dedicated to my grandfather David Zaval Sr., his love for nature was an inspiration.

Acknowledgements

First, I would like to express my sincere gratitude to my advisor Dr. Gernot Presting for providing the opportunity to work in his lab. His patience and diligence during my time in his lab was key to my growth as a researcher and his thoughtful guidance and creative thinking were invaluable. I truly admire and respect his passion and resolve to uncover the truths of the natural world. I would like to thank Kevin Schneider for his patience, expertise and tutelage which were a boon early in my PhD. I am also grateful to my colleague Ryan Shontell, our morning coffee trips were quite productive, it was a great help to bounce ideas off you and have a good laugh.

I would like to thank my committee for their patience and time in reviewing this dissertation and helping me progress as a researcher. I truly appreciated the many discussions with Dr. Kantar whose thoughtful recommendations on both my research and career goals helped keep me on a good path. I would also like to extend my thanks to the University of Hawaii at Manoa, Information Technology Services – Cyberinfrastructure, High Performance Computing center, funded in part by the National Science Foundation MRI award #1920304 for technical support and for their tireless assistance to myself, and the greater UH research community. I also extend my thanks to the National Science Foundation as well as the United States Department of Agriculture for grants “Functional Genomics of Maize Centromeres” and “Characterizing the interplay between Maize Retrotransposons and the Epigenome”, and the University of Hawaii, without whom this research would not be possible.

I would like to thank my family for their love and support. I miss you all terribly, and I think I am ready to come home. Finally, I would like to thank my partner, Amanda Olson for her patience, love, and steadfast support through our time here. We have seen many ups and downs on our adventure, but they have only brought us closer together. This chapter of our lives is about to close, but I’m truly excited to write the next one with you by my side.

ABSTRACT

Centromeres are chromosomal structures defined epigenetically by the histone cenH3. These structures are required for the faithful segregation of chromosomes during cell division, and survival of all eukaryotic organisms. The sequences of the centromere are highly repetitive, primarily containing tandem repeats and transposable elements. Remarkably, the repeats and proteins involved in centromere function evolve rapidly though the function of centromeres is highly conserved, a phenomenon referred to as the "centromere paradox". This has been interpreted to be due to evolutionary pressure on both molecules based on their interaction (centromere drive). Recent investigations into domesticated maize have shown that artificial selection and genetic bottlenecks during domestication heavily influenced the evolution of domesticated centromeres, rather than the interplay of centromeric repeats and proteins as hypothesized by meiotic drive. Disruption of the ancestral centromere has the potential to cause the formation of a neocentromere in the pericentromere or chromosome arm. In *Zea mays,* loss of the centromeric tandem repeat centC has the potential to destabilize centromeres. When the centC is lost cenH3 can either spread into the chromatin flanking centC, or jump to a new locus entirely. Our investigations have determined that the jump from CEN5M to CEN5L or CEN5R was the result of transcribed genes flanking CEN5M which restricted the spread of cenH3 into flanking regions following loss of centC. The pericentromeres provide relatively ideal locations for neocentromere formation due to low transcription. The newly formed neocentromeres were stabilized through the loss of a H2A.Z loci that were shown to destabilize cenH3, via deletions and suppression by cenH3. Regions of limited recombination surrounding centromeres were defined by structural variation and selection in CEN2 and it was observed that a tight linkage between a domestication locus to a centromeric haplotype region can reduce the observed centromeric diversity likely due to linkage drag. Following the formation of neocentromeres they were rapidly colonized by CR2 retrotransposons which expanded in a burst event during domestication. There was

extensive recombination among these elements, but a preferred recombination was not responsible for their expansion. It is possible that expansion of the CR2 subfamily was the result of horizontal transfer of a foreign CR2 element, or the result of epigenetic changes in the ancestral centromere following neocentromere formation. Regardless of the trigger for expansion, the CR2 insertions served to further stabilize the neocentromeres by increasing the repeat content and pushing sensitive DNA such as genes and regulatory regions out of the neocentromeres. Overall, these investigations suggest the story of domesticated maize centromere evolution was one of stability. The process of domestication destabilized the centromeres, and the centromeres have since been evolving to reacquire stability in a new genomic context.

TABLE OF CONTENTS

Acknowledgements iii
ABSTRACT iv
LIST OF FIGURES ix
LIST OF TABLES x
LIST OF ABBREVIATIONS x
CHAPTER 1. Introduction and Literature Review 1
1.1 An introduction to *Zea mays* and domestication 1
1.2 Centromere composition, structure, and epigenetics 2
1.3 Recombination and Structural Variation in Centromeres 5
1.4 Neocentromeres 6
1.5 Centromeric Retrotransposons 7
1.6 Recombination in CR elements 8
1.7 Transposable element evolution 9
1.8 Evolution of Centromeres 10
1.9 The value of studying centromeres 12
1.10 Progression and development of *Zea mays* physical maps 12
CHAPTER 2. Assembly of the *Zea mays* B73 RefGen_v4CEN Physical map 14
2.1 Abstract 14
2.2 Introduction 14
2.3 Methods 16
2.3.1 Extraction and annotation of RefGen_v4 centromere and pericentromere regions 16
2.3.2 Anchoring all assembled unitigs that contain centromere repeat sequence 16
2.3.3 Gap Closure 17
2.3.4 Assembly 17
2.4 Results 18
CHAPTER 3. H2A.Z and transcription influence neocentromere positioning in CEN5 of *Zea mays*. 22
3.1 Abstract 22
3.2 Introduction 22
3.3 Methods 25
3.3.1 Preparing and sequencing H2A.Z libraries 25
3.3.2 CenH3, RNAseq and H2AZ mapping 26
3.3.3 Peak calling and identification of shared peaks 26

3.3.4 Mapping to CR2s 27
3.3.5 Genome-Wide 101mer mapping 27
3.4 Results 27
3.4.1 Transcription prevents centromere expansion 27
3.4.2 Low transcription in potential neocentromere positions CEN5L and CEN5R 28
3.4.3 Active neocentromeres contain smaller and fewer H2A.Z peaks 31
3.4.4 Reduced H2A.Z in active neocentromeres resulted from reduced H2A.Z signal and deletion of the underlying sequence 33
3.4.5 H2A.Z is not enriched in CR2 elements 36
3.4.6 H2A.Z epigenetic signal was resilient to cenH3 invasion 38
3.5 Discussion 40
3.5.1 Model of neocentromere positioning in CEN5 40
CHAPTER 4. Recombination near active centromeres is limited by selection and structural variation 42
4.1 Abstract 42
4.2 Introduction 42
4.3 Methods 45
4.3.1 CenH3, RNAseq and H2AZ mapping 45
4.3.2 Genome Wide 101mer mapping 45
4.3.3 Mauve alignments using whole genomes 45
4.3.4 Substitution rate calculation 46
4.3.5 Generating phylogenetic trees 47
4.3.6 HM3 SNP dating 47
4.4 Results 48
4.4.1 Distribution of meiotic recombination in maize centromeres 48
4.4.2 H2A.Z was correlated with centromere proximal recombination in B73 50
4.4.3 CEN2 had three post-domestication haplotypes among domesticated maize 50
4.4.4 There was extensive structural variation in CEN2 among the NAM lines 54
4.4.5 There was a tight linkage between the CEN2 haplotype region and the CEN2 domestication locus 56
4.4.6 The CEN2 domestication locus was acquired by HAP1 and HAP3 from HAP2 56
4.4.7 The CEN2 domestication locus and structural variation defined the CEN2 haplotype region 60
4.4.8 Recombination in CEN5 and selection for Region-D 60
4.4.9 Candidate domestication genes in CENs 1, 4, and Region-D 63
4.5 Discussion 64

4.5.1 Model of CEN2 evolution 64
4.5.2 Model of centromere haplotype formation in maize 64
4.5.3 Potential consequences centromere haplotype formation 65
4.5.4 Future directions 66
CHAPTER 5. The expansion and evolution of the CR2 subfamily in domesticated maize 66
5.1 Abstract 66
5.2 Introduction 67
5.3 Methods 69
5.3.1 CR2 extraction 69
5.3.2 CR2 phylogenetic analysis, haplotyping 69
5.3.3 CR2 recombination and consensus sequence generation 69
5.3.4 CR1 and CR2 fractions for maize lineages 70
5.3.5 SNP identification via short read mapping and SNP enrichment 70
5.3.6 UMR analysis 70
5.3.7 Identification of the oldest element in the CR2 Clades 71
5.3.8 Verification of shared elements in the unique list of CR2s 71
5.4 Results 72
5.4.1 Inventory of the CR family members in B73 72
5.4.2 The CR2 subfamily expanded in a post-domestication burst 74
5.4.3 A preferred recombination was not responsible for CR2 expansion 77
5.4.4 SNPs in recently active CR2 clades in domesticated maize were enriched in *Zea luxurians* 81
5.5.5 Unmethylated regions in CR1 and CR2 are correlated with the active centromere 84
5.5 Discussion 87
5.5.1 CentC loss during domestication led to the decrease of CR1 in modern inbreds 87
5.5.2 Expansion of CR2 via horizontal transfer of a foreign CR2 88
5.5.3 Expansion CR2 as a consequence of neocentromere formation 89
5.5.4 Model of CR2 Expansion 90
CHAPTER 6. Summary and Conclusions 91
6.1 Summary of major findings 91
6.2 Interplay of epigenetics and transcription in the centromere 93
6.3 The impact of genes and structural variation on centromere evolution 94
6.4 Functional implications of CR2 elements in neocentromeres 94
6.5 Model of centromere evolution in domesticated maize 96

6.6 Conclusion 97
6.7 Future directions 97
REFERENCES 100

LIST OF FIGURES

Figure 3.1. CEN5L and CEN5R neocentromere loci have low transcription regardless of lineage. 30
Figure 3.2. CEN5L/CEN5R ratios of H2A.Z content among the 26 NAM lines. 32
Figure 3.3. H2A.Z distribution among the NAM lines. 35
Figure 3.4. H2A.Z and cenH3 coverage over a CR2 consensus. 38
Figure 3.5. Isolated H2A.Z loci lacking RNA-seq can form due to gene damage. 40
Figure 4.1. Overview of B73 v5 centromeres. 49
Figure 4.2. CEN2 HapMap3 SNP tree for REGON13_SYN. 53
Figure 4.3. An overview of CEN2 divergence among the NAM lines with respect to the B73 v5 physical map. 54
Figure 4.4. An overview of CEN2 recombination among the NAM lines and a comparison of the CEN2 domestication locus and the two flanking homeobox genes in *Zea mays* and *Sorghum bicolor* 59
Figure 4.5. Recombination across the multi-haplotype CEN5 and region D among domesticated maize. 62
Figure 5.1. Insertion time distribution for CR elements of the B73 genome. 74
Figure 5.2. An increase in CR2 content in the genome of domesticated maize correlates with a reduction in CR1. 76
Figure 5.3. Log10 transformation of cumulative insertions in B73 over the last 600ka. 77
Figure 5.4. Date distributions and recombination among the clades of the CR2 subfamily. 80
Figure 5.5. Enrichment of 91 SNPs to the “G” consensus CR2 in domesticated maize teosinte and Zea lines. 83
Figure 5.6. UMR, H2A.Z and cenH3 coverage for CR2-G (a.) and CR1-R5 (b.). 86

Figure 5.7. Distribution of UMRs across the CEN5 brown lineage (CML277 and CML322). 87

LIST OF TABLES

Table 2.1. Statistics for the subset of RefGen_v4 data used to assemble RefGen_v4CEN. 19

Table 2.2. Assembly statistics of the centromeres and flanking regions. 20

Table 2.3. Assembly statistics of the functional centromeres of B73. 21

Table 4.1. Estimated divergence dates for CEN2 using Mauve and HapMap3 SNP data. 53

LIST OF ABBREVIATIONS

FISH – Fluorescence *in situ* hybridization

CR – Centromeric Retrotransposon

CEN – Centromere

LTR – Long terminal repeat

TSD – Target site duplication

PPT – Polyprotein

PRT – Protease

RT – Reverse transcriptase

GAG – Group specific antigen

INT – Integrase

SNP – Single nucleotide polymorphism

NAM – Nested association mapping (used to reference the genetically diverse NAM founder lines)

TIL – Teosinte inbred line

BKN – A group of maize landraces in HapMap3

ChIP – Chromatin immunoprecipitation

DSB – Double strand break

CHR - Chromosome

HAP - Haplotype

Os – *Oryza sativa* (rice)

Si – *Setaria italica* (foxtail millet)

Bd – *Brachypodium distachyon*

Sb – *Sorghum bicolor*

Zm – *Zea mays*

At – *Arabidopsis thaliana*

CHAPTER 1. Introduction and Literature Review

1.1 An introduction to *Zea mays* and domestication

Zea mays (maize, corn) is a cereal grain crop in the *Poaceae* family along with other cereal grains *Oryza sativa* (rice), *Triticum aestivum* (wheat) and *Hordeum vulgare* (barley) among others that together represent ~40% of all calories consumed by humans [1]. Domestication of maize occurred ~10 thousand years ago (ka) from the wild grass *Zea mays* subsp. *parviglumis* in the Balsas River Valley region of what is now Mexico [2-4]. Due to its food value, maize percolated through Mesoamerican religion and culture. Indeed, in the Popol Vuh, a 16th century mythological and historical account of the K'iche' people, mankind is born from maize, and maize is deified in various contexts. As a cornerstone of diets globally, a cultural icon, and a model organism, *Zea mays* and its domestication has been subject to intense study. Two analyses on maize domestication have estimated that a population of between ~500 to 5,600 individuals went through a bottleneck lasting between 300-2,800 years resulting in a 40% loss in genetic diversity [5, 6]. Similar losses of diversity have been documented in a variety of domesticated plant systems [7-9]. Intense selective pressure during the domestication of annual species often results in a single haplotype over certain genes resulting in a sweep within the domesticated organism. There have been considerable efforts to uncover and study these domestication loci as they are often linked to genes of agronomic importance [10-14]. Intense selection also has the potential to decrease diversity in nearby regions due to genetic linkage [14]. Consequently, this selection can decrease the diversity of large loci if recombination is limited [15]. It can also perpetuate minorly deleterious alleles due to linkage drag which may contribute to inbreeding depression (a decrease in fitness of offspring from related individuals). The identification and impact of these deleterious alleles is of considerable interest to plant breeders.

The genetic and regulatory changes acquired during domestication led to major phenotypic changes in the domesticate. These phenotypes are well summarized in in Stitzer *et al.* 2018 but include increased apical dominance, reduced shattering, and increased yield [16]. However, the genetic basis for the formation of the domesticated maize phenotype are still not completely understood. Initial analysis suggested the domestication phenotype was the result of relatively few large effect loci [10]. However more recent analyses suggest that it is more likely the result of hundreds of genes that all contribute to the domesticated phenotype [14]. Certain phenotypes found in maize such as separation of male and female flowers (simple controlled crosses), large number of seeds per plant (easy identification of segregation ratios for kernel phenotypes) and large chromosomes (easily visualized) and relatively short generation time, led to its adoption as a model organism [17]. As such, a great variety of genetic tools such as SNP/haplotype maps, and physical maps (among countless other resources) have been developed for *Zea mays*. The volume of high-quality data for *Zea mays* is a boon for plant geneticists and permits the study of more challenging regions of the genome including the highly repetitive regions such as centromeres.

1.2 Centromere composition, structure, and epigenetics

Centromeres are chromosomal structures that are required for the faithful segregation of chromosomes during cell division [18]. At the chromosomal level, centromere position is variable and may be metacentric, acrocentric or telocentric, aligning to either the center, offset, or terminal regions of the chromosome respectively [19]. Chromosomes may also have one of two centromere variants, being either holocentric or monocentric. Typically, plant chromosomes (including maize) are monocentric being characterized by a regional centromere and defined constriction point [20, 21]. Conversely, holocentric chromosomes are characterized by having a centromere almost the entire size of the chromosome, and contain no observable constriction [21]. However, cases do exist where monocentric

chromosomes can have multiple centromeric domains [20]. This level of plasticity is remarkable given centromere function is so critical to the survival of eukaryotic organisms.

At the regional level, centromeres are heterochromatic and remain condensed by nucleosomes at interphase [22]. While the core of many centromeres contains tandem repeat arrays (for additional background on tandem repeats see *SI Appendix Background SB1*), the active centromeres themselves are not defined by the DNA, but rather epigenetically by the centromeric histone variant cenH3 (CENP-A in humans and CID in *Drosophila*) [23]. Active centromeric regions are identifiable through the measurement of the distribution of cenH3 via chromatin immunoprecipitation sequencing (ChIP-seq) or visualized via fluorescence *in situ* hybridization (FISH) [23-25]. However, the existence of cenH3 in a region does not necessarily mean that it will become an active centromere. Indeed, studies of CENP-A have shown that ectopic CENP-A islands are common [26], and lack centromere activity.

Mammalian centromeric histones manifest as an octomer containing two each of histones H2A, H2B, H4 and CENP-A that takes the place of the canonical H3 histone [27]. However, some studies have observed tetrameric hemisomes [28]. A consensus structure has not been released for cenH3 in *Zea mays* and centromeric histones vary greatly, particularly in the N-terminal region and in long loop 1 regions [29-31]. The nucleosomes package around 146 bp of DNA wrapped in a left-handed orientation around the octomer [29]. Interestingly, it was found that DNA wraps more loosely around centromeric nucleosomes, though the reason for this is unknown [32]. The cenH3 histones act as the foundation upon which the multiprotein kinetochore complex is constructed, and is physically moves chromosomes toward opposite poles during cell division [33]. Studies have shown that human CENP-A nucleosomes are dispersed in blocks along the centromeric chromatin flanked by blocks of the canonical H3 containing nucleosomes [34]. The histones self-organize and position the centromeric histones to allow for recruitment of the outer kinetochore proteins [35]. This structure forms the chromatin basal layer of the active centromere.

Several proteins have been shown to attach directly to centromeric chromatin that form the foundation of the kinetochore complex. CENP-T, CENP-W, CENP-S and CENP-X have been shown to form a heterotetramer complex that binds directly to, and supercoils centromeric chromatin [36]. The kinetochore itself is the protein mechanism that binds to centromeric chromatin and interacts with spindle fibers through its associated motor proteins. Experiments have uncovered many kinetochore proteins a list can be found in Table 1 of Yu *et. al.* 2000 [37]. Studies in *Lilium Longiforum* have revealed the macro structure of the plant centromere to be roughly spherical and embedded in the centromeric chromatin [38]. Animal kinetochores differ in that they can be subdivided into inner and outer kinetochore plate macrostructural domains [39]. Plant kinetochores generally lack the plate structures seen in animal kinetochores, but are far less characterized. Regardless, many conserved kinetochore assembly protein homologs have been identified in maize including a CENP-A, CENP-B and CENP-C homologs, with CENP-C shown to play an active role in early kinetochore assembly [40]. To maintain kinetochore activity, the cenH3 domain must be large enough to assemble the kinetochore, though it is currently unknown what the minimum cenH3 requirement is for a stable kinetochore.

Flanking the active centromere/kinetochore region are regions of pericentromeric heterochromatin. The pericentromeric heterochromatin is populated by histones with post translational modifications including H4K20me1, H3K9me1, H3K9me2, and H3K27me1, all of which repress transcription [41-43]. In *Arabidopsis thaliana* the pericentromeric regions are repetitive containing a high density of retrotransposons as well as various local and global tandem repeats and the DNA is highly methylated [44]. However, despite the characterization of pericentromeric heterochromatin little is known about whether/how it contributes to centromere function in plants though it is suspected to be involved in orienting sister chromatids via the protein cohesin [45, 46].

Indeed, epigenetics play a central role in the formation of centromeres, but while the cenH3 histone variant epigenetically defines active centromeres, H2A.Z (an H2A histone variant) is often associated

with genes/transcription. In genes where H2A.Z is promoting transcription, it accumulates as a nucleosome dyad proximal to transcription start sites (TSS) and flanking a nucleosome deficient region (NDR). H2A.Z is also the most conserved of the histone variants [47, 48]. The relationship between H2A.Z and transcription is variable with prior studies observing H2A.Z both repressing and activating transcription [49, 50]. Notably, the H2A.Z histone variant is often absent from the most highly expressed genes as well as those with no expression [51]. It also implicated to have multiple roles including nucleosome turnover, DNA repair, and DNA methylation (for more information of DNA methylation in centromeres see *SI Appendix Background SB2*) [51-53]. Analysis of H2A.Z and DNA methylation has repeatedly shown a negative correlation across various organisms [54-56]. The antagonistic relationship between H2A.Z and DNA methylation indicate H2A.Z can protect genes from DNA methylation and vice versa.

Beyond a transcriptional role, H2A.Z has been implicated in promoting the initiation of meiotic recombination in fission yeast and Arabidopsis suggesting H2A.Z also plays a critical role in facilitating recombination [57, 58]. Using fission yeast H2A.Z mutants, it was found that H2A.Z increases the efficiency of binding for meiotic DSB proteins and balances chromosome compaction. However, H2A.Z was not found to be enriched at recombination hotspots, suggesting H2A.Z is not directly required for meiotic recombination, but can facilitate it. However, the behavior or H2A.Z in the centromeres of maize is unknown.

1.3 Recombination and Structural Variation in Centromeres

Centromeres are generally characterized by little meiotic recombination, however there is significant structural diversity observed among otherwise closely related haplotypes [24]. While it is true that there is little traditional recombination occurring, multiple gene conversion events have been observed in maize centromeres, leading to the hypothesis that centromeres convert rather than cross over resulting

in the observed diversity [59]. A detailed study of *Oryza sativa japonica* CEN8 has shown evidence of low frequency unequal recombination, segmental duplication, as well as amplification and reshuffling of the centO tandem repeat [60]. Evidence for these mechanisms as well as evidence of double strand break (DSB) have been observed in maize CEN10 [61]. These studies show that while centromeres are resistant to traditional chromosomal crossover, they are subject to a multitude of mechanisms capable of introducing variation. Little is known about the location of centromere proximal recombination breakpoints among teosinte, landraces, and domesticated maize that is critical to understanding maize centromere domestication and evolution.

It is important to note the centromeres serve as the attachment point of spindle fibers during cell division and are subject to mechanical stress. Evidence exist for frequent large scale genomic rearrangements in centromeric regions, particularly due to DSB and its subsequent repair [61]. These events can lead to inversions, duplications, fusions, translocations, and deletions of centromere proximal genomic sequence have been extensively documented [24, 61-64]. Additionally, DSB and repair have resulted in the degradation of collinearity in centromere proximal regions, as observed between sorghum and rice [65]. These events can also shrink centromeres forcing cenH3 to expand into neighboring regions to maintain the minimum cenH3 required for a stable kinetochore [24, 61]. When a centromere locus is destabilized, the centromere appears to jump to a new location and results in the creation of a *de-novo* neocentromere.

1.4 Neocentromeres

The term neocentromere, can be used to describe two phenomena, either the motility of heterochromatin domains on the chromosome arms during cell division, or *de-novo* centromere formation physically distant from an ancestral centromere locus [24, 66, 67]. For simplicity, the term neocentromere will refer to *de-novo* centromere formation in this dissertation. Neocentromeres are

often smaller than the native centromeres and generally form in pericentromeric regions [68]. In plants, neocentromeres have been studied in detail in barley and maize [69, 70]. *Zea mays* neocentromeres are common in the centromeric lineages of multiple chromosomes [24]. In fact four different active centromeric loci exist in CEN5 alone (CEN5R, CEN5L, CEN5L', CEN5M) [24]. However, little is known about what directs neocentromere positioning following disruption of the ancestral centromere. Movement of the centromere in *Zea mays* is correlated with the loss of tandem repeat at the ancestral centromere, reported to be the initial disruption causing formation of the neocentromeres [24]. Interestingly, neocentromeres tend to acquire repetitive DNA elements over time. In *Zea mays*, this is due to invasion of the neocentromere by centromere targeting retrotransposons [24]. The movement of centromere loci also has the potential to place genes in, or proximal to active centromeres, especially if it moves to a formerly euchromatic region. The genes that end up in centromeric heterochromatin have been observed to have altered transcription levels [71]. Previous investigations in *Oryza sativa* spp. *japonica* have shown that actively transcribed genes can persist in the active centromere [72]. While counterintuitive, heterochromatic transcription can occur by several mechanisms reviewed by Yasuhara et al. 2006 [73]. The many neocentromere loci in *Zea mays* suggest that there are many centromere proximal genes, however the transcriptional activity and epigenetic interactions with cenH3 are not known.

1.5 Centromeric Retrotransposons

Following the formation of the neocentromeres in maize, they were colonized by centromere targeting retrotransposons (CRs) first identified by Jiang *et al.* 1998 in Sorghum centromeric BAC clones containing a high concentration of retrotransposon related sequence [74]. FISH analysis of the retrotransposon related centromere sequence revealed that Ty3-Gypsy sequence is centromere specific (for a primer on TEs see *SI Appendix Background SB3*). Furthermore, they showed that CRs are present in multiple

grasses (Sorghum, Wheat, Maize, and Rye). Centromeric retrotransposons have been classified as members of the *Chromoviridae* and characterized by the presence of a CR motif on the integrase domain [75-78]. This domain is proposed to assist in centromeric targeting, though the exact mechanism by which CRs target centromeric chromatin is unknown. Centromeric retrotransposons are subdivided into six subfamilies (CR1 – CR6) [79]. The most active members of the CRs in maize following domestication are CR2 and CR1. Analysis of CEN5 in *Zea mays* indicates most of the post domestication CR insertions are from the CR2 subfamily, which inserted into the active centromere locations in each lineage [24]. The CR1, CR2, and CR3 subtypes are capable of preferentially integrating into centromeric regions by an unknown mechanism (for details of TE transposition see SI *Appendix Background SB3.2*). The integration of these elements targeted to a relatively small genomic region (~2Mb) often results in the formation of nested insertions [80]. Interestingly, TEs need not be autonomous (i.e. containing all five complete uninterrupted functional domains) to transpose. Both Class 1 and Class 2 elements (for more details on TE classes see *SI Appendix Background SB3*) can use the transposition machinery of closely related elements to transpose [81]. An example of this can be found among the CR elements in the non-autonomous CR3 derivative CentA. CR1 and CR2 elements dominate the centromeric regions of maize alongside the CentC tandem repeat. As CR elements insert into genomic DNA are subject to various mechanisms that can result in the formation of recombinant subtypes.

1.6 Recombination in CR elements

Previous studies in CR1 have shown evidence of multiple recombination events giving rise to several CR1 subtypes. The parental CR1 subtypes (A and B) produce five major recombinant subtypes (R1-R5) though only R4 and R5 are active post-domestication [82]. It is postulated that the recombinants are formed via intrastrand homologous recombination of nested CR1 elements resulting in full-length hybrid elements that could then transpose [82]. This mechanism is like that which forms “solo” CR1 long terminal repeats

(LTRs); CR1 elements which have recombined over the LTR effectively deleting the untranslated region (UTR) and polyprotein (PP) regions but maintaining a single LTR with target site duplications (TSDs). The mechanism of transpositional recombination is not ruled out, however. In this mechanism, two different CR1 elements recombine during cDNA synthesis in a virus like particle (VLP) forming recombinant elements as observed in retroviruses [83]. In addition to forming new recombinant elements, CR1 and CR4 have been shown to form tandem repeats (CTR1 and CTR4) in CEN9, and CEN6 as the result of interelement recombination [84]. While the lineages of CR1 have been studied in detail, little is known about the evolution of the CR2 lineage, which has expanded considerably in maize centromeres since domestication.

1.7 Transposable element evolution

The expansion of the CR2 lineage is of interest because of the burst of activity that has occurred since domestication. Transposon bursts are an evolutionary phenomenon that have the capacity to radically restructure genomic DNA. Indeed, transposon amplification is one of the primary mechanisms of genome expansion in plants [85]. The bursts generally coincide with the generation of new phylogenetic groups, and are identified as high frequency TE insertions over a relatively short period of time [86, 87]. Most LTR retrotransposons are relatively young, though insertions up to ~15 million years have been reported [88]. The high ratio of young retrotransposons suggests either older elements are being purged or the older elements have become too nested and damaged to be reliably identified. It has been proposed that retrotransposon bursts are accompanied by DNA loss following the burst event to compensate for genomic expansion [89, 90]. Studies have shown that genomic DNA loss through illegitimate recombination and unequal homologous recombination are common, with illegitimate recombination being the predominant mechanism [88, 91]. These mechanisms generate solo LTRs, truncated elements, and deletions that serve to balance genomic DNA growth [90]. Rapid transposon

insertion is characteristic of neocentromere formation in *Zea mays* and has contributed to speciation in the *Oryza* genus [24, 86]. The mechanisms described above highlight the active role TEs play in host evolution.

Indeed, TE insertions particularly in the 5' UTR or introns of genes have the capability of altering the host genomes expression. These expressional differences are well characterized and led to the very discovery of TEs by Dr. Barbara McClintock [92, 93]. Due to the general ubiquity of retrotransposons across the plant kingdom, it is possible that retrotransposons can mediate transfer of genetic material between distantly related species as observed with the BovB LINE element in vertebrates and a *copia* LTR element between *Drosophila* species [94, 95]. TE insertions in the host genome can sway an organism's evolution through the alteration of gene function and expression, facilitation of chromosomal rearrangements, gene capture, and expansion of the host genome through amplification as described above. Currently it is unknown if retrotransposons play a "functional" role in the genome or more specifically the centromere, however they are generally considered to be commensal at best, and at worst deleterious due to the possibility of mutation by gene disruption.

1.8 Evolution of Centromeres

Functional centromeres are required for the segregation of chromosomes in eukaryotes. As such, their function is essential to an organism. The molecular machinery (kinetochore and its associated proteins) responsible for this is highly conserved across eukaryotic organisms [96]. However, the DNA and histone constituents found in centromeric chromatin evolve rapidly [97]. The disconnect between the observed rapid evolution of centromeric DNA and histones, and the highly conserved function has been dubbed the centromere paradox. The longstanding hypothesis to explain this paradox is meiotic drive, first observed in *Drosophila obscura* [98]. This hypothesis suggests that the rapid rate of evolution is the result of selfish elements that can increase a chromosomes chance of transmission to the functional

gamete. This hypothesis states that genomic loci, in this case centromeres, can be segregated preferentially during female meiosis. An example of this has been studied in mice that demonstrated centromere strength, indicated by kinetochore protein levels and altered interactions with microtubules, can predict the direction of meiotic drive [99]. A second example of this can be found in the maize knob repeats in the presence of abnormal chromosome 10 (Ab10) [100]. The knobs are heterochromatic regions, which under the right conditions, are capable of binding microtubules [101]. In a heterozygote, the knob containing chromatids are preferentially pulled to the outermost megaspores that are more likely to produce gametes [102]. As such, these knobs can exploit female meiosis to perpetuate themselves. In broad terms, centromere drive postulates that an expansion of a preferred tandem repeat or histone allows for increased transmission in female meiosis and leading to positive selection for the chromosome.

A seminal study into the evolution of the centromeres in *Zea mays* presents a different, but not mutually exclusive hypothesis. It is reported that decreases in genetic diversity caused by selection during domestication influenced the evolutionary trajectory of domesticated maize centromeres [24]. It was found in CEN10, that temperate maize lines that passed through a genetic bottleneck had decreased the levels of the centromeric tandem repeat CentC when compared to tropical lines. The CentC loss occurred within the last 2.5-5.8 ka suggesting the loss is likely tied to domestication. The prevalence of decreased CentC among temperate lines may also indicate a founder effect resulting from maize moving from Mexico into temperate regions of the United States [24]. Analysis of SNP data (HapMap2) across the centromere regions also revealed low genetic diversity over many centromeres, which suggests selection for centromeric haplotypes [24]. It was shown that large centromere proximal inversions can link formerly euchromatic genes to centromeres, and the resulting selection on the genes reduced centromeric haplotype diversity [24]. This mechanism presumably resulted in the formation of single haplotype centromeres CEN1, CEN4, and CEN6 and the formation of two haplotype centromeres CEN2,

CEN3, CEN8, and CEN9. These findings paint a different picture of centromere evolution in domesticated maize. In the proposed mechanism, artificial selection and genetic bottlenecks during domestication heavily influenced the evolution of domesticated centromeres, rather than the interplay of centromeric repeats and proteins as hypothesized by meiotic drive.

1.9 The value of studying centromeres

Centromeres are critical to the perpetuation of eukaryotic life thus makes them worthy of investigation. Though from the perspective of basic science, the study of centromeres gives insight into the mechanics of evolution and chromosome transmission from generation to generation. From a more applied standpoint, the study of maize centromeres has the potential to identify new breeding targets to improve maize productivity by the introduction of domesticated alleles, or removal of deleterious mutations fixed in the centromeres during domestication. Recombination maps and investigation of centromeric structural variation could also potentially be of value for geneticists/breeders working with genes proximal to active centromeres. Outside of agriculture, the study of centromere stability and formation of neocentromeres has implications for human disease, as neocentromeres have been implicated congenital abnormalities, developmental delay, intellectual disability, and at least two cancers (lipomatous tumors and acute myeloid leukemia) [103]. Because of their repeat-rich nature, centromeres represent the last frontier of genome sequencing and have long been considered a “black box”, improvements in sequencing technology specifically long-read technologies, such as Oxford-nanopore and PacBio, have lifted the veil on these elusive genomic regions.

1.10 Progression and development of *Zea mays* physical maps

As a model system and valuable crop, *Zea mays* has benefited from substantial funding, allowing for early adoption of emerging technologies. Arguably one of the most valuable datasets to be released for

an organism is a high-quality physical map. While published genomes are a dime a dozen these days, this was not always the case, and the quality of the assemblies is still variable depending on the organism, methods, and sequencing technology. The first physical map of *Zea mays* was published in 2009 using the inbred line B73 (B73 RefGen_v1, [104]). This physical map was a sequenced using the Sanger methodology on bacterial artificial chromosomes (BACs) that allowed for high nucleotide accuracy at low read depth. In 2012 the B73 genome was updated to B73 RefGen_v2 that closed many gaps and further refined the genome and represented a major advance in the B73 physical map. A third update was released in 2013 B73 RefGen_v3 to further correct the genome, however, the improvement from v2 was marginal. A fourth and major advancement came in 2017 with the release of B73 RefGen_v4 generated using PacBio SMRT sequencing coupled with BioNano optical mapping [105]. The increased read length allowed for entire full-length retrotransposons and other repeats to be fully sequenced which allowed for detailed study of transposable element families as well as the repeat rich centromeres and telomeres. While the centromeres of RefGen_v4 were notably far better than any previous assembly they were still fragmented. To rectify this the B73 genome was manually edited in the centromeric regions resulting in the RefGen_v4CEN genome (see Chapter 2). The fifth and most recent release was a complete *de novo* resequencing of the B73 genome and 25 additional maize lines that were used as the founders of the maize NAM (Nested Association Mapping) population which were chosen to reflect most of the diversity in maize [106]. The lines were all sequenced using the most recent PacBio SMRT sequencing chemistry coupled with BioNano optical mapping to produce truly superb physical maps. This major advancement allows for direct comparisons at the nucleotide level, however methods for comparing whole genome physical maps at the nucleotide level are still developing. The advancement of physical maps in *Zea mays* has progressed rapidly and each new release requires the validation of previous findings. However, the additional data is invaluable to elucidating details regarding the mechanics of phenomena observed in the genome. Furthermore, the

high-quality assemblies of the repetitive centromeres among a diverse group of inbred lines permits the comparative study of centromeric evolution.

CHAPTER 2. Assembly of the *Zea mays* B73 RefGen_v4CEN Physical map

2.1 Abstract

Genome assemblies based on long read sequencing technology have revolutionized the quality of the repeat-rich centromere regions. However, because maize centromeres are highly enriched for the tandem repeat CentC and a group of evolutionarily related centromeric retrotransposons (CR), automated genome assembly leaves gaps, even in the excellent B73 RefGen_v4 reference genome constructed from long-read PacBio data. Manual editing of >140 Mb spanning the ten centromeres of maize inbred B73 resulted in the closure of an additional 127 sequence gaps and the addition of >8.4 Mb of previously unanchored sequence containing 24 genes, 2 Mb of CR repeat and 887 kb of CentC. The functional centromeres of five maize chromosomes were closed completely, including a 7 Mb region spanning the extremely CR2-rich CEN2. This improved assembly, **B73 RefGen_v4CEN**, was completed in February 2019 and is available at https://doi.org/10.25739/7y1p-5169, both as pseudomolecules and as centromere assemblies alone. This assembly has since been superseded by the B73 v5 physical map and the additional 25 NAM founder physical maps.

2.2 Introduction

Until the recent release of the B73_v5 and NAM assemblies [106] the B73 RefGen_v4 assembly [105] has been the highest quality *Zea mays* assembly available, with notably better centromere assemblies than previous B73 reference assemblies. However, RefGen_v4 was still fragmented and incomplete in the centromeres. We wanted to ensure we had the best centromere physical maps we could obtain so

we generated an intermediate physical map B73_RefGen_v4CEN using unplaced contigs and reads from the RefGen_v4 assembly. Manual editing was used to close all centromere proximal gaps except those flanked by >10 kb of tandem repeat (primarily the centromere specific CentC). For each chromosome, we edited the active B73 centromere, including the ancestral centromere location (Often a CentC tandem repeat locus), all regions on which a neocentromere formed in one or more NAM parents, and the centromeric haplotype regions consisting of a single or two haplotypes [24].

Previous analysis has reported that centC loss caused by inbreeding during domestication resulted in a destabilization of the ancestral centromeres and resulted in neocentromere formation. The neocentromeres were then populated by centromeric retrotransposon insertions that target active centromeres [24]. Although the focus of this project was assembly of the ten functional centromeres of inbred B73 (approximately 20 Mb) as defined by the histone H3 variant cenH3, we extended our efforts to include the ancestral centromere locations, as well as the formerly pericentromeric regions that gave rise to various neocentromeres in one or more of 27 maize inbred lines following domestication. These regions had been previously identified via cenH3 ChIP-seq in inbred lines B73, Mo17, W64A and most of the NAM lines [24]. The centromere regions of domesticated maize had been found to have very little genetic diversity in that seven of the ten maize centromeres are represented by only one or two pre-domestication haplotypes. The edited region of these chromosomes were extended to the edges of the one (CEN1, CEN4 and CEN6) or two haplotype (CEN2, CEN3, CEN8 and CEN9) centromere regions described previously [24], resulting in a total of >140 Mb of the maize genome being subjected to manual editing. Here we describe the methods used for the creation of B73 RefGen_v4CEN as well as basic statistics regarding the updated assembly including the number of gaps closed and total sequence added.

2.3 Methods

2.3.1 Extraction and annotation of RefGen_v4 centromere and pericentromere regions

MUMmer [107] was used to identify the centromere and major haplotype regions [24] in RefGen_v4 as defined in RefGen_v2 [104]. Following extraction of these sequences from RefGen_v4 the regions were analyzed using Junction Viewer (JV) [108] to visually identify gaps, tandem CentC repeats [109], CR elements belonging to subfamilies CR1 through CR6 [110, 111], genes, direct and inverted repeats, organellar genome fragments, and other retrotransposon junctions. RefGen_v4 pseudomolecules were obtained from NCBI (PRJNA10769), unitigs (provided by the Ware laboratory) and reads (SRP067440) were provided by the Maize B73, AGPv4 Consortium. A subset of the RefGen_v4 sequence data was used to assemble RefGen_v4CEN (**Table 2.1**) including 3,303 high quality unitigs, 145,880 low quality unitigs and 4,047,379 reads.

2.3.2 Anchoring all assembled unitigs that contain centromere repeat sequence

All high quality unitigs were examined for CentC, CR1 and CR2. Thirty-eight of the 57 CentC-containing high quality unitigs had been placed in the original RefGen_v4 assembly. Thirteen additional CentC-containing HQ unitigs were anchored during the editing process. Three of these were chimeric (utg3669, utg15323 and utg8675) and required trimming of the misplaced sequence (based on RefGen_v2) prior to incorporation into RefGen_v4CEN. Four CentC-containing high quality unitigs, which are comprised of 40% to >80% CentC, remain unplaced because they did not contain unique junctions that could be mapped to a centromere. Two high quality unitigs containing CR2 remain unanchored: chimeric utg115865 (695kb) contains a single CR2 that had been improperly joined at one end and utg129473 (345kb) contains one CR2 near a sequence similar to rDNA.

2.3.3 Gap Closure

Using the centromere JV images as a guide, unique junctions flanking the gaps in the RefGen_v4 assembly were used to identify overlap, first with singleton or repeat unitigs, then raw long reads or, if needed, using sequence from RefGen_v2. The highest quality sequence containing the junction was imaged with JV and compared visually to the anchored sequence to confirm overlap. A gap was considered closed if a unitig or read covered two unique junctions on either side of the gap. Chimeric unitigs, for which different regions mapped to more than one chromosome or two distinct regions on the same chromosome of RefGen_v2, were split at the chimeric point prior to assembly.

Gaps within the tandem repeats CentC and Cent4 [112] were filled by searching for all unique junctions from the corresponding centromeric regions of B73 RefGen_v2. If a CentC or Cent4 containing fragment was flanked by approximately 10kb of CentC, Cent4, or organellar sequence, a gap of 100 Ns was inserted between the fragments.

2.3.4 Assembly

After all unitigs that could be used to close the gaps were identified, they were imported into Geneious 6.0 [113]. The B73 RefGen_v4 assembly region was extracted and broken into contigs at gaps. Gapped regions were assembled using the “Map to Reference” function, aligning one unitig at a time to walk across each gap. The sensitivity parameters varied from unitig to unitig, but “medium sensitivity” was used to assemble most unitigs.

Sequences were manually edited to introduce the least number of ambiguities. Low-quality sequence that overlapped with high quality sequence was trimmed to retain the latter. For chromosomes 2, 4, 5 and 10, overlapping low quality unitigs were trimmed to remove overlap. For chromosomes 1, 3, 6, 7, 8 and 9 the unitig overlap was manually edited for continuity at the nucleotide level. Raw PacBio reads were trimmed if overlapping a unitig or another read to introduce as little error from a raw read as

possible. A final JV annotation of the region was constructed to do a manual verification of the updated assemblies.

2.4 Results

The RefGen_v4CEN assembly contains an additional ~8.4 Mb of sequence data resulting in the closure of 127 gaps (**Table 2.2**). Approximately 2.9 Mb of centromeric repeat, including ~1.18Mb of CR2, were added to the regions. In the functional B73 centromeres, 16 gaps were closed and ~2Mb of N's were replaced with sequence that included more than 1Mb of CR2 (**Table 2.3**). Five of the ten B73 active centromeres (CENs 2, 4, 5, 8, and 9) are complete and contain no gaps. Approximately 5.3 Mb of sequence is estimated to be missing, mostly from CEN1, CEN6 and CEN7 which contain a high concentration of centC tandem repeat [114]. CentC still presents a challenge to assembly even with modern long read platforms as tandem repeats may not contain enough unique markers or variation to bridge contigs. Advancements in sequencing platforms allowing for even longer reads with high read depth will eventually allow the complete closure of the centromeric tandem repeats.

Chimeric sequences also present a challenge to centromere assembly. The high density of repeat elements (both tandem repeats and retrotransposons) increases the likelihood of mis-assembly of similar but unique repeat elements (i.e. two unique insertions of a CR2 retrotransposon). These chimeras often occur in regions of low sequence coverage such as the termini of unitigs. In these locations, sequence reads originating from different retrotransposons or tandem repeat sequences (often from different chromosomes) can be falsely joined. Because the chimeric unitig cannot be accurately joined to the next unitig in the assembly, they often appear adjacent to gaps. These chimeras are identified through analysis of retrotransposon junctions or by cross reference with RefGen_v2 data. Chimeras can exceed 10kb in length and can be challenging to identify, as such they are a major obstacle in assembling high quality centromeric sequence.

Sequence Type	Number of Sequences	Max # of reads per contig	Min # of reads per contig	Median length	Total bases
High quality unitigs	3,303	11,139	3	432,132	2,099,336,497
Repeat unitigs	14,338	1,587	2	21,816	329,013,878
Singleton unitigs	131,542	1	1	9,461	1,467,086,700
Reads	4,047,379	N/A	N/A	8,630	40,351,155,710

Table 2.1. Statistics for the subset of RefGen_v4 data used to assemble RefGen_v4CEN. All values represent nucleotides unless noted otherwise.

CEN	Manually Edited Region (Mb)	# Gaps Closed	Change in Length (v4CEN – v4)	Replaced "N"	CR1 Added	CR2 Added	CentC Added
CEN01	13.949	15	171,109	458,285	73,294	37,834	112,428
CEN02	7.001	9	9,566	683,782	52,790	221,512	14,503
CEN03	9.998	9	710,338	443,926	150,893	218,558	179,428
CEN04	20.3	29	626,614	423,723	145,883	237,414	0
CEN05	25.839	10	-22,609	773,775	30,175	81,505	48,302
CEN06	14.29	4	511,716	6,539	180,632	93,497	122,457
CEN07	14.238	14	-72,634	355,942	39,372	52,513	94,547
CEN08	10.353	11	279,166	401,666	50,736	164,266	99,367
CEN09	11.899	11	-50,855	113,232	50,628	6,125	100,027
CEN10	12.929	15	2,274,592	393,291	120,038	69,483	116,841
Total	**140.796**	**127**	**4,437,003**	**4,054,161**	**894,441**	**1,182,707**	**887,900**

Table 2.2. Assembly statistics of the centromeres and flanking regions. In total, 127 gaps were closed and ~8.4Mb of sequence (almost half of which replaced Ns) was added to RefGen_v4CEN. All values represent nucleotides unless noted otherwise.

CEN	SIZE (Mb)	Missing Sequence (Mb)	# of Gaps	Replaced "N"	CR1 Added	CR2 Added	CentC Added
CEN01	0.27	1.52	3	-1,885	19,332	36,995	67,206
CEN02	2.11	complete	0	683,682	52,100	215,591	14,537
CEN03	1.56	0.17	4	372,341	111,812	186,115	104,141
CEN04	1.75	complete	0	313,321	115,093	230,098	0
CEN05	1.96	complete	0	277,105	-1,312	54,564	0
CEN06	0.52	0.71	4	-1,203	104,417	48,975	90,980
CEN07	0.19	2.3	3	-1,251	27,789	37,098	53,688
CEN08	2.02	complete	0	357,416	14,026	145,603	0
CEN09	1.72	complete	0	1,004	0	-8,741	0
CEN10	1.81	0.6	2	70,840	102,337	63,708	57,257
Total	**13.91**	**5.3**	**16**	**2,071,370**	**545,594**	**1,010,006**	**387,809**

Table 2.3. Assembly statistics of the functional centromeres of B73. N's were added to CENs 1, 6 and 7 in tandem repeat gaps of unknown length resulting in an increase in total ambiguities in the major haplotype regions of CEN1, 6 and 7. Loss of CR2 sequence in CEN9 was due to removal of chimeric CR2 sequence. All values represent nucleotides unless noted otherwise.

CHAPTER 3. H2A.Z and transcription influence neocentromere positioning in CEN5 of *Zea mays*.

3.1 Abstract

Centromeres are required for the faithful segregation of chromosomes during cell division and are defined epigenetically by the histone variant cenH3. Centromeres typically contain long stretches of tandem repeat (centC in *Zea mays*). CentC loss in maize, thought to have been caused by inbreeding during domestication, reduced the size of the active centromeres. Transcription of housekeeping genes flanking the ancestral CEN5M prevented the spread of cenH3 following centC deletion and resulted in the centromere relocating and forming *de novo* neocentromeres either upstream (CEN5L) or downstream (CEN5R) of the ancestral CEN5M. Both neocentromere loci contain transcribed regions and H2A.Z loci, that conflicts with cenH3. However, H2A.Z levels are reduced in active neocentromere regions relative to the inactive neocentromere regions, furthermore H2A.Z can be suppressed by cenH3, stabilizing the neocentromere. Two deletions were uncovered in CEN5L that resulted in the loss of H2A.Z loci that may have also had a stabilizing effect on the neocentromere. Additionally, the insertion of centromere targeting CR2 retrotransposons, are generally not enriched for H2A.Z, can further stabilize the neocentromeres by pushing H2A.Z/genes out of the neocentromere and increasing repeat content. CEN5L and CEN5R are not ideal centromere loci but through suppression/loss of H2A.Z/transcription, and the accumulation of repeats, the loci are stabilized.

3.2 Introduction

Centromeres are epigenetically defined by the centromeric histone variant cenH3 (CENP-A in humans and CID in *Drosophila*) and are required for the faithful segregation of chromosomes during cell division [23]. The centromeres of most organisms are characterized by long stretches of centromere specific tandem repeats (centC in *Zea mays*). Though maize is atypical, in that many of its centromeres have

moved from the ancestral tandem repeat loci and formed neocentromeres in the pericentromeric regions. One such centromere is that of chromosome 5 of *Zea mays* (CEN5). It was reported that loss of centC due to inbreeding during domestication resulted in a destabilization of the ancestral centromere locus (CEN5M) and subsequent neocentromere formation at one of two loci, CEN5L, or CEN5R [24]. The centromeric positions and evolution of CEN5 have been investigated in detail and thus, CEN5 provides a suitable exemplar to investigate the factors influencing neocentromere positioning and stability in maize.

Neocentromere formation in *Zea mays* manifests in two ways. In the first case, the centC is encompassed within, or overlapping, the active centromere. In these situations, the loss of tandem repeat led to expansion of cenH3 into the formerly pericentromeric regions directly flanking the centC. In second case, the neocentromere is distant from the centC loci and the centromere appears to have jumped from its ancestral position. CEN5 is a case where both CEN5L and CEN5R are distant from the centC suggesting a jump. However, little is understood as to why the neocentromeres jumped from the ancestral position, rather than just expanding into the regions flanking the centC.

Previous analysis of CEN5 has reported eight primary lineages among the genetically diverse NAM founder lines (*SI Appendix Figure S1,* Schneider *et al.* 2016 see Figure 2 [24]). All but CML333, Il14H and P39 fall into five colored lineages, Blue, Orange, Brown, Green, and Red which diverged ~18 thousand years ago (ka). Analysis of centromere positions among these colored lineages showed that neocentromere positioning is not genetically linked, as the closely related Blue and Orange lineages (diverged 9ka) have centromeres at the CEN5L and CEN5R locus respectively and Red and Green lineages (diverged 7ka) have centromeres at the CEN5R and CEN5L locus respectively. This suggests that factors such as epigenetics or transcription are responsible for the positioning of neocentromeres.

The factors affecting neocentromere positioning have been investigated previously in vertebrates and yeast by inducing neocentromere formation through deletion of the active centromere [115, 116].

Studies conducted in chicken DT40 cells indicated that neocentromeres generally form in regions of low transcription and that CENP-A (cenH3) occurs at low levels outside of the kinetochore region which was proposed to seed neocentromeres. It was also shown that CENP-A density is greatly reduced over transcribed regions in active neocentromeres. Furthermore, a high proportion of the neocentromeres observed in the study occurred proximal to the ancestral centromere, consistent with studies in *Drosophila*, and suggests heterochromatin may play a role in facilitating neocentromere formation [116-118]. Not all neocentromeres observed were proximal to pericentromeric heterochromatin but follow up studies in chicken DT40 cells reported neocentromeres that lack pericentromeric heterochromatin, associate with non-pericentromeric heterochromatic regions [119].

Other epigenetic factors such as the histone variant H2A.Z (replaces the canonical H2A histone) may also play a role in neocentromere positioning. Studies in the yeast *Schizosaccharomyces pombe* have shown that neocentromeres preferentially form at subtelomeric regions which contain lower levels of Pht1 (H2A.Z) variant histones. They report H2A.Z is disruptive to neocentromere formation and the loss of H2A.Z content in chromatin improves neocentromere function suggesting reduction in H2A.Z represents an intermediate step in the establishment of stable neocentromeres. H2A.Z is the most conserved of the histone variants and it generally accumulates as two defined nucleosome positions flanking a nucleosome deficient region (NDR) containing the transcription start site (TSS) [47, 48]. The relationship between H2A.Z and transcription is complex, and it has been reported that H2A.Z positioning in +1 nucleosomes (promoters) can activate transcriptional activity in genes and H2A.Z distribution across gene bodies is repressive of transcription [49, 50, 120]. Furthermore, H2A.Z is generally absent from the most highly expressed genes and those with no expression [51]. It has also been implicated to have either direct or indirect roles in nucleosome turnover, DNA repair, and DNA methylation [51-53]. Here we examined the relationships between cenH3, H2A.Z, and transcription in the well documented *Zea mays* CEN5 using the recently sequenced, and genetically diverse NAM founder lines [106]. Our analysis

suggested H2A.Z and transcription disrupt cenH3 resulted in the neocentromere being positioned in the locus least disruptive to centromere function.

3.3 Methods

3.3.1 Preparing and sequencing H2A.Z libraries

A peptide corresponding to the N-terminal amino acids of ZmH2A.Z (AGKGGKGLLAAKTTAAKSTDKDKDRKKAPVSC) was synthesized by Zidian Xie, and rabbit polyclonal antibodies (anti-ZmH2A.Z antibody) against this peptide were generated and purified by GenScript. Immature ear (B73) or leaf (Mo17 and NAM lines) tissue inbred lines were used to perform H2A.Z chromatin immunoprecipitation (ChIP) by Zidian Xie according to previous published protocols (Schneider et al, 2016, Xie and Presting, 2016). Briefly, a fine powder was obtained from maize ear (B73) or leaf tissues by grinding in liquid nitrogen using mortar and pestle. After cross-linking with 1% formaldehyde on ice for 20 minutes, the samples were filtered using two layers of Miracloth and washed with M1 buffer (11.9 % hexylene glycol, 10 mM KPO4, pH 7.0, 100 mM NaCl, 5 mM beta-mercaptoethanol, 0.1 mM PMSF, 1× Plant protease inhibitor cocktail) and M2 buffer (8.85 % hexylene glycol, 10 mM KPO4, pH 7.0, 10 mM MgCl2, 0.5 % Triton X-100, 5 mM beta-mercaptoethanol, 100 mM NaCl), crude chromatin was digested by micrococcal nuclease in MNB buffer (50 mM Tris–HCl, pH 8.0, 2.5 mM CaCl2, 4 mM MgCl2, 0.3 M Sucrose) at 37°C to generate mononucleosomes. After pre-clearing with rabbit anti-IgG antibody and protein A dynabeads, the digested nucleosomes were incubated with anti-ZmH2A.Z antibody at 4°C overnight. The antibody-nucleosomes were purified with protein A dynabeads, followed by several rounds of washing. Chromatin immuno-precipitated DNA (ChIP DNA) was purified after reverse cross-linking overnight at 65°C plus RNase A and Proteinase K treatments. Library preparation and sequencing for the B73 H2A.Z ChIP sample were as described in Schneider *et al*, 2016. Illumina libraries of the other 26 maize lines were prepared with the NEBNext Ultra II DNA Library

Prep Kit and sequenced with the NovaSeq Reagent Kit v1.5_150x150 bp Sequencing. A table with read counts per dataset can be found in *SI Appendix Table ST5*.

3.3.2 CenH3, RNAseq and H2AZ mapping

Paired cenH3 reads for B73 cenH3 ChIP-seq, RNAseq and H2A.Z (PRJNA811190) were mapped uniquely to the B73_v5 reference genome using bowtie (cenH3 and H2A.Z ChIP-seq) or TopHat2 (RNAseq) allowing no mismatches for ChIP-seq and one mismatch for H2A.Z and RNAseq. An insert size of 2000 and 300 was used for cenH3 and H2A.Z datasets respectively. A combination of samtools mpileup and bcftools call were used to generate a VCF file from which raw coverage data was obtained (samtools mpileup -v -u -t DPR -BQ0 -d10000000 -A -f GENOME.fasta BAM_ALIGNMENT.bam | bcftools call -Ov -m -A -V indels --threads 20 > VARIANT_CALL.vcf).

3.3.3 Peak calling and identification of shared peaks

Active centromere regions were manually identified from cenH3 coverage data where cenH3 rises above ~2X background. H2A.Z peaks were called using MACS2 (macs2 callpeak –broad –broad-cutoff 0.1 and the output was filtered requiring a fold change >=3). Peaks were extracted from the dataset that correspond to CEN5L and CEN5R. Corresponding regions in the NAM lines were identified using shared gene markers (*SI Appendix Tables ST1 and ST2*). To identify shared peaks the sequences called as peaks in the A-D (CEN5L) and E-H (CEN5R) regions were extracted from the respective genomes. The set of sequences called as peaks in A-D and E-H were aligned to themselves using blastn. Shared peaks were identified as blast hits overlapping by >=100nt in the same orientation. The list of shared peaks was then manually curated for continuity.

3.3.4 Mapping to CR2s

B73 H2A.Z data was mapped to the B73_v5 physical map allowing multimapping to account for the two CR2 LTRs (Bowtie, X 300, m 2). Peaks were called with MACS2 on the resulting BAM file (default parameters). Using the MACS2 peak output file the peak coordinates were compared to the coordinates of CR2 elements identified in B73_v5 via blastn looking for any coordinate overlap. Identified were 42 peaks that overlapped with CR2 elements. H2A.Z and cenH3 were mapped to the CR2 consensus with bowtie (X 300, m 2) coverage data was extracted with bcftools as described in the previous section, CenH3, RNAseq and H2AZ mapping.

3.3.5 Genome-Wide 101mer mapping

The 101mers moving at 1nt per step were generated using a custom Perl script for each of the 10 chromosomes in the v4CEN genome. Any duplicated 101mer, and any 101mer with >2 ambiguous bases were removed. The filtered 101mer list for each chromosome was mapped uniquely against the v4CEN genome allowing no mismatches using bowtie. The resulting bam files were then merged using samtools merge. Coverage values were obtained for the 101mers using the same methods used for obtaining cenH3 coverage values.

3.4 Results

3.4.1 Transcription prevents centromere expansion

The CEN5L and CEN5R loci are both separated from the ancestral CEN5M by more than 1Mb, giving the appearance that the active centromere jumped from CEN5M to either CEN5L or CEN5R. RNAseq data from four tissues (root, shoot, ear and tassel) for the 26 NAM lines (NCBI: SRP011480) mapped to their respective physical maps revealed evidence of transcription in genes immediately upstream and downstream of CEN5M in all 26 NAM lines. CEN5M is flanked upstream by the housekeeping gene

AUGMIN subunit 6 (npi571) and is expressed nearly universally (TAIR). Downstream, CEN5M is flanked by six gene models expressed at variable levels among the 26 NAM lines (**Figure 1***, SI Appendix Table ST1*). The fact that the CEN5M did not spread over these genes suggests cenH3 was not present in transcribed regions. Previous analysis of the expressed *MAMDC2* gene found in an artificially generated neocentromere of chicken DT40 cells displayed a conflicting relationship between cenH3 and transcription, as well as an overall decrease in transcription level [116]. This is consistent with our observations of transcribed genes in the active centromeres of chromosomes 3 and 4 in *Zea mays* B73. The two transcribed genes in the active centromeres display low cenH3 and H2A.Z across the gene body suggesting neither histone variant was present (*SI Appendix Figure S2*). Furthermore, a previous review has suggested that transcribed genes may prevent cenH3 expansion [121]. Given the evidence that cenH3 and transcription were conflicting forces, it was reasonable to conclude that the transcribed genes flanking CEN5M prevented CEN5M from expanding after loss of the centromeric tandem repeat. Being unable to maintain the minimum cenH3 for kinetochore function, CEN5 was forced to move to either the CEN5L or CEN5R locus.

3.4.2 Low transcription in potential neocentromere positions CEN5L and CEN5R

Transcription was found to be low in both CEN5R and CEN5L regions regardless of whether that line had an active CEN5L or CEN5R (**Figure 3.1**). Since all 26 NAM lines share the transcriptional voids in CEN5L and CEN5R we can conclude that transcription alone was not a sufficient indicator of where a neocentromere would jump. Previous analysis of transcription levels over a set of eight genes in neocentromeres generated in chicken DT40 cells indicated neocentromeres preferentially form in regions of decreased transcription. They go on to suggest that because neocentromere formation is disruptive to transcription and vice versa, neocentromere formation in gene rich loci could potentially be detrimental to the organism [116]. Both CEN5L and CEN5R formed in the pericentromeric regions of

the former CEN5M. Pericentromeric DNA contains a high repeat density (including retrotransposons), is often heavily methylated, and the nucleosomes carry post translationally modified variants (H4K20me1, H3K9me1, H3K9me2, and H3K27me1) repressive to transcription [41-43, 122]. It stands to reason that the low transcription density of pericentromeres makes them suitable sites for stable neocentromere formation. To verify that the regions of low transcription in CEN5L and CEN5R were not a result of the limited number of tissues used in the transcription analysis, we mapped an additional 73 RNAseq datasets to B73_v5 (*SI Appendix Figure S3*). It was found that transcription remained low in the B73_v5 CEN5L and R regions regardless of tissue/treatment. Notably, there were three total transcribed peaks that occurred within the CEN5L and CEN5R regions, though only one corresponded to a gene model (Zm00001eb235330, gene marker "F" in CEN5R region of B73). Previous analysis of genes in *Oryza sativa* had shown four gene models were expressed in the ~750kb kinetochore region of CEN8 suggesting that transcription can persist in active centromeres [72]. Taken together, these data suggested transcription generally conflicted with cenH3, but some active transcription of genes in neocentromeres was permitted.

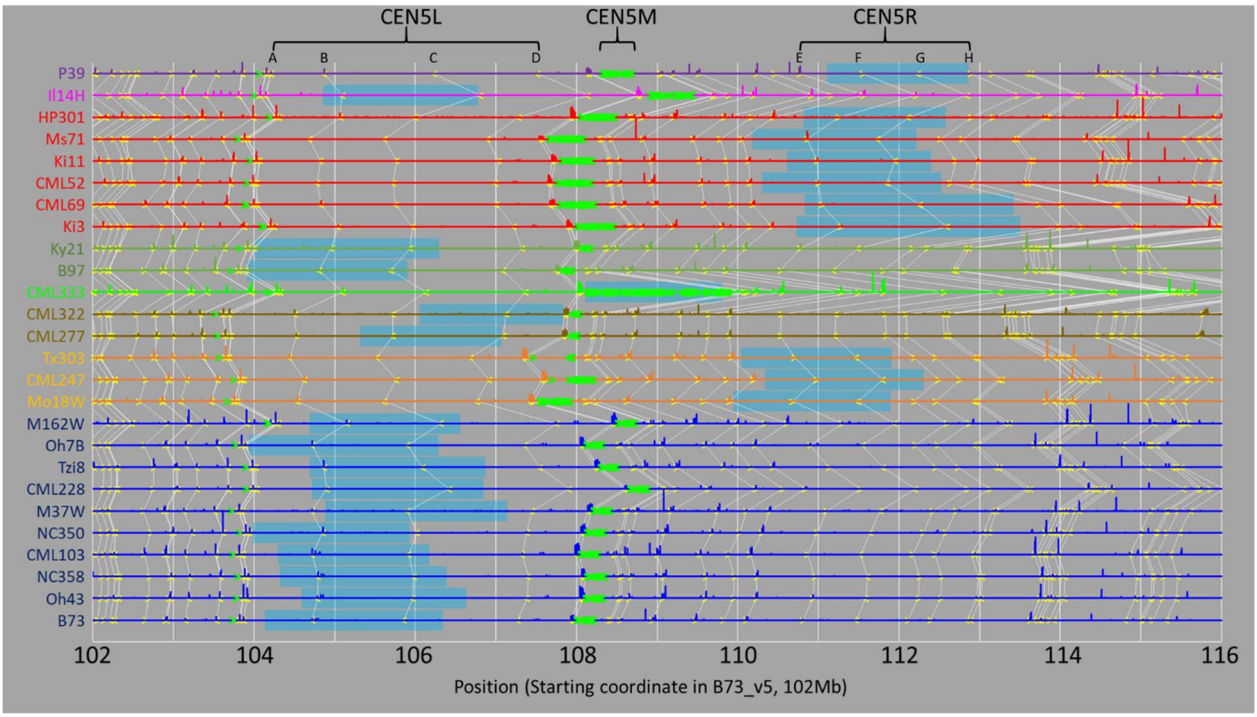


Figure 3.1. CEN5L and CEN5R neocentromere loci had low transcription regardless of lineage. Combined RNAseq coverage data for root, shoot, ear, and tassel tissues overlaid onto schematic data for CEN5L-R including the active centromere loci (light blue shade), centC (green arrows) and gene markers shared by the 26 NAM lines (yellow arrows connected by white lines). The lines were sorted and colored based on the lineages published in Schneider et al 2016 [24]. Corresponding CEN5L and CEN5R regions among NAM lines were defined by shared gene markers A-D (CEN5L) and E-H (CEN5R) (*SI Appendix Table ST2*). The ancestral centromere locus CEN5M was flanked by transcribed genes while the pericentromeric CEN5L and CEN5R regions displayed low transcription density. The neocentromeres of CEN5 were found to have jumped from the ancestral CEN5M to their current location rather than spread out of the centC as seen in other centromeres (CEN2) the flanking transcription prevented the centromeres from sliding and forced them to jump into CEN5L or CEN5R.

3.4.3 Active neocentromeres contain smaller and fewer H2A.Z peaks

H2A.Z is a highly conserved histone variant implicated either directly or indirectly, in telomeric silencing, genome stability, cell cycle progression, DNA repair, recombination and transcriptional regulation (reviewed in, Zlatanova *et al.* 2008) [123]. We sought to determine the relationship between the histone variants cenH3 and H2A.Z in the active centromeres of the NAM lines in CEN5. The ratio for total H2A.Z peak associated nucleotides and the per-nucleotide H2A.Z coverage were summed in CEN5L and CEN5R for corresponding regions in the NAM lines (**Figure 3.2**). The ratios resolved into three major clusters. CML333 (CEN5M) and P39 (a recently formed CEN5R with only two post-domestication CR2s) had very similar ratios (~1.5). All CEN5R lines except HP301 (a Red line contaminated with a Green Ky21 related line *SI Appendix Table ST6*) had coverage sum ratios >1.5 indicating CEN5R lines had more H2A.Z in CEN5L than CEN5R. The opposite was true for the CEN5L lines including Il14H, and the Blue, Green, and Brown lineages all of which had coverage sum ratios <1.5 indicating they had more H2A.Z in CEN5R than CEN5L. From this we concluded that H2A.Z content was reduced in active centromere regions. This was consistent with previous analysis of H2A.Z and cenH3 in *Schizosaccharomyces pombe* neocentromeres that showed a reduction of H2A.Z content in chromatin improved neocentromere function. They further reported that if scarcity of H2A.Z at neocentromere loci is not established, the neocentromere will fail to propagate efficiently suggesting reduction in H2A.Z represents an intermediate step in the establishment of stable neocentromeres [115]. Taken together, we concluded that there was an overall decrease in H2A.Z in active neocentromeres regardless of lineage that may serve to stabilize the neocentromeres.

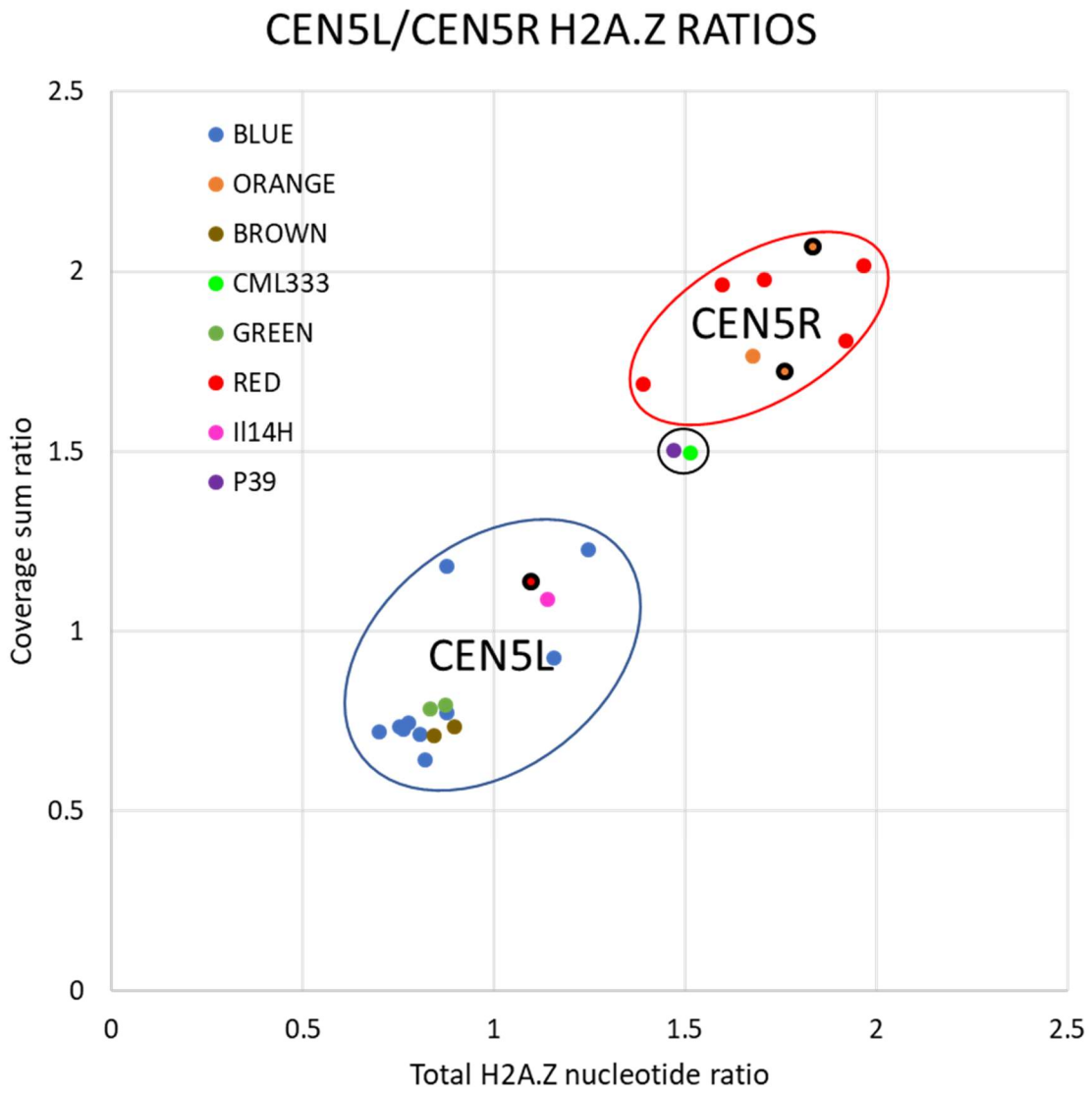


Figure 3.2. CEN5L/CEN5R ratios of H2A.Z content among the 26 NAM lines. The x axis corresponds to the ratio of total nucleotides associated with H2A.Z peaks called with MACS2. The y-axis represents the ratio of the per nucleotide coverage sum for CEN5L/CEN5R (*SI Appendix Table ST3*). lines were colored based on their lineage. Blue, Green, Brown and Il14H are CEN5L lines (blue circle). Red, Orange were CEN5R lines (red circle). P39 contains only two post domestication CR2 insertions suggesting it was a very recent CEN5R neocentromere formation placing it close to the CEN5M CML333. We observed low ratios in CEN5L lines and higher ratios in CEN5R lines indicating that H2A.Z loci were both less frequent, and at a lower coverage in regions occupied by neocentromeres. The HP301 sample was contaminated with a CEN5L-containing Ky21 related green line (*SI Appendix Table ST6,* Schneider et al. 2016 supplemental figure S1) and represented a heterozygote from of the RED/GREEN lineages that resulted in a skewed ratio. The orange lines Tx303 and CML247 (Black bordered orange points) were found to have variable signal to noise ratios in CEN5L and CEN5R resulting in more false positive peak calls by MACS2. To correct this, the fold change threshold was increased from 3 to 3.5 in CML247 and Tx303 (*SI Appendix Figure S4*).

3.4.4 Reduced H2A.Z in active neocentromeres resulted from reduced H2A.Z signal and deletion of the underlying sequence

To further analyze the relationship between H2A.Z and cenH3 in CEN5 we identified H2A.Z peaks shared between NAM lines (**Figure 3.3a**). Five peaks were identified that were phylogenetically significant (i.e., shared by CEN5L lines only or CEN5R lines only, **Figure 3.3** peaks 4-8). Peak 4 was present only in the CEN5R lines and peaks 5-8 were present in CEN5L lines only. We observed no H2A.Z peaks shared by only CEN5L lines in the CEN5L region, and no peaks was shared by CEN5R lines only, in the CEN5R region. Remarkably these five H2A.Z peaks were shared between distantly related lineages (Blue/Green, ~18ka and Orange/Red ~18ka) but the neocentromeres formed roughly 10ka. Furthermore, the underlying sequence for these peaks was present in the physical map of most NAM lines that were missing the peak. This being the case, we concluded that H2A.Z vacated these regions in inbreds that were missing the peak (**Figure 3.3b**). This suggested that cenH3 can both suppress H2A.Z in active centromeres, and that H2A.Z loss occurred after neocentromere formation. It was notable that some peaks display both cenH3 and H2A.Z above background coverage levels, suggesting there is some variability in whether H2A.Z or cenH3 occupies these loci.

In addition to the five phylogenetically significant peaks, three peaks were observed in CEN5L that were missing from a single-colored group, one peak was missing in Green and two in Blue. It was found that the loss of these peaks was caused by independent deletions in the Blue (~113kb) and Green (~308kb) CEN5L. These deletions took place in the regions now occupied by the active centromeres <10ka, though it was unknown whether the deletions occurred before or after neocentromere formation. The addition of the deleted peaks back to a CEN5L Blue line was not sufficient to replicate the H2A.Z ratios observed in CEN5R lines suggesting the trends observed in **Figure 3.2** were not caused by a single peak. We examined the NAM lines for additional deletions and uncovered a third, shared in the active centromere region of the brown lines (~85kb) though this deletion did not result in the loss of an H2A.Z locus (*SI*

Appendix Table ST4). Previous studies have shown that chromatin containing H2A.Z has a poor association with the cenH3 loading chaperone Scm3 (HJURP in vertebrates) which had a destabilizing effect on neocentromeres of *Schizosaccharomyces pombe* [115]. The fact that cenH3 can suppress H2A.Z nucleosomes, and the deletion of H2A.Z peaks in CEN5L suggest that if a gene (and its accompanying H2A.Z locus) within a centromere is unnecessary to the point it can be silenced or deleted, then deletion of H2A.Z loci within a neocentromere may have a stabilizing effect by improving the efficiency of cenH3 loading in neocentromeres. Taken together the observation that H2A.Z was reduced in active centromeres, was the result of silencing or deletion of multiple H2A.Z peaks following neocentromere formation, that had a stabilizing effect on the neocentromere loci.

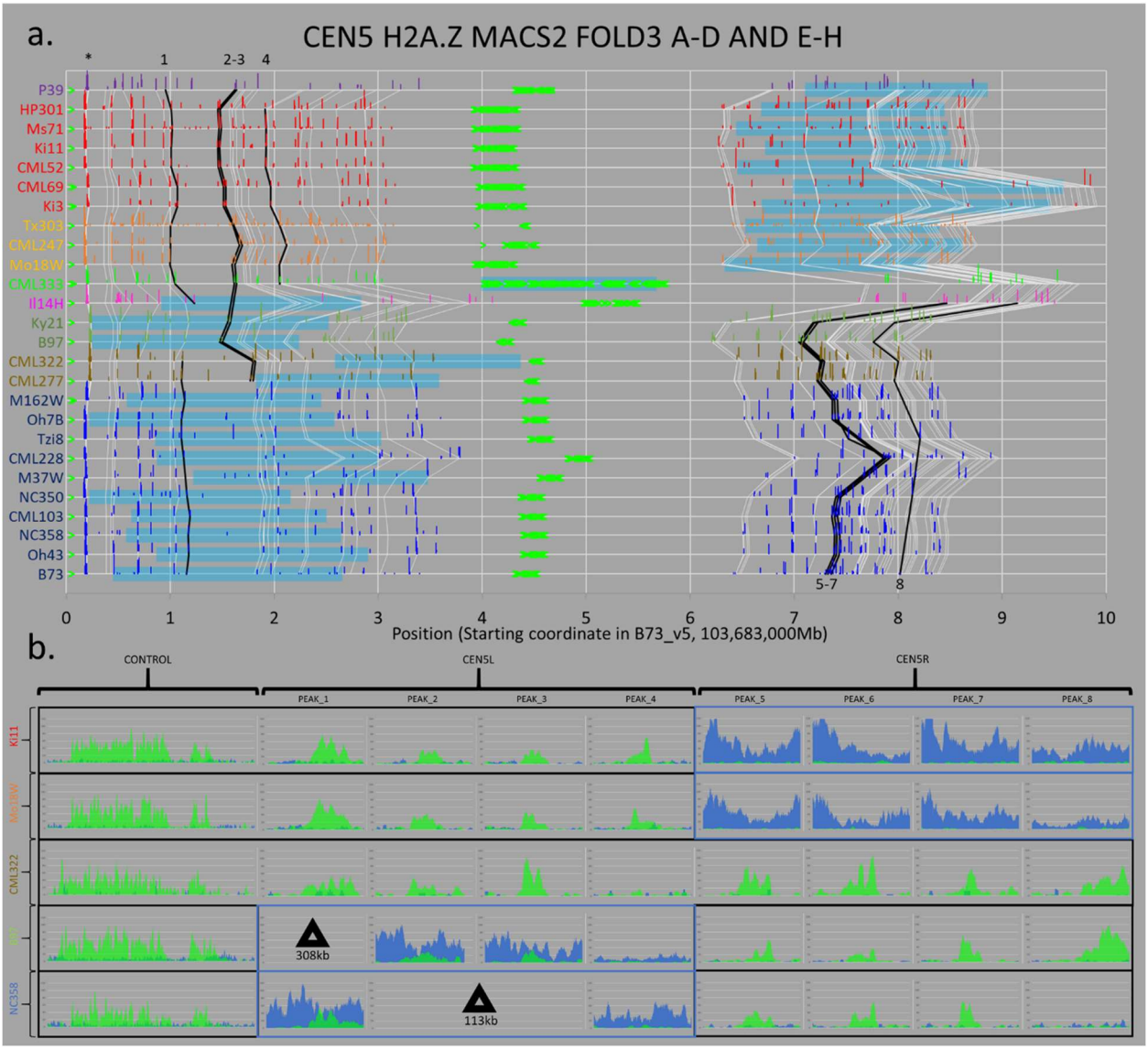


Figure 3.3. H2A.Z distribution among the NAM lines. a. H2A.Z coverage data for A-D (CEN5L) and E-H (CEN5R) (SI Appendix Table ST2) region peaks called by MACS2 overlaid onto a schematic of CEN5L-R included the active centromere loci (light blue shade), CentC (green arrows). The lines were sorted by centromere position (CEN5L, CEN5M and CEN5R) and colored based on the lineages published in Schneider et al 2016. Sequence shared by all 26 NAM lines and called as an H2A.Z peak in at least one line were connected by white lines. The asterisk represents an ~20kb region shared by all 26 NAM lines rich in H2A.Z upstream of the A-D region and was used as a control region to normalize H2A.Z within each line in panel **b.** Per nucleotide coverage data for H2A.Z (green) and cenH3 (blue) normalized to the CEN5L control peak. Black lines numbered 1-8 represented peaks that were phylogenetically significant. Peak 1 and peaks 2-3 represent H2A.Z peaks lost via independent deletions in the green and blue lines respectively. Peak 4 was called only in the CEN5R lines. Peaks 5-8 were called in CEN5L lines only. We did not observe any peak shared by CEN5L lines in CEN5L, the same holds for CEN5R. The fact that the phylogenetically distant blue/green lineages share H2A.Z peaks in CEN5R and the underlying sequence of the peaks was conserved in most CEN5R lines and suggested that H2A.Z was lost from the positions in CEN5R lines. Panel **b.** shows per-nucleotide coverage data over the control region and the eight

phylogenetically significant peaks for a representative of each of the colored lineages. Peaks 1-3 were deleted, however the underlying sequence for peaks 4-8 were present in most NAM lines and suggested cenH3 suppressed the H2A.Z peaks at these loci. A combination of peak deletion and suppression of H2A.Z by cenH3 that resulted in the observed H2A.Z ratio differential between CEN5L and CEN5R. Per-nucleotide coverage data for these peaks and additional peaks across CEN5L and CEN5R among all colored lines.

3.4.5 H2A.Z is not enriched in CR2 elements

Studies of CEN5 in maize have shown that CEN5L and CEN5R have been actively targeted by a group of Ty3-Gypsy retrotransposons (CR2) since domestication (~10ka) which have increased repeat content in the neocentromeres [24]. To evaluate the interaction of H2A.Z and CR2 we initially focused on CR2 elements overlapping at a H2A.Z peak within the active centromere, that inserted post-domestication, and were uninterrupted. Only three elements in B73 fit these criteria, one in CEN1 and two in CEN8 (*SI Appendix Figure S10*). Investigation of H2A.Z and cenH3 over these elements with 2kb flanking sequence indicated that the H2A.Z peaks were between 17X and 31X lower than the H2A.Z control peak located upstream of CEN5 (**Figure 3b**). Furthermore, the H2A.Z peaks appeared to only overlap the LTR junction and did not extend into the LTR where the CR2 promoter was located. We expanded our investigations to include elements that were associated with H2A.Z peaks and inserted post domestication but were not located in the active centromeres. Two additional elements fitted these criteria on chromosomes 1 and 4. The element on chromosome 1 was ~35Mb downstream of the active centromere and had substantial coverage, but this peak overlapped the junction with the 3' LTR. The peak on chromosome 4 was in the 5' LTR but was at a coverage level ~35X lower than the control peak. We also investigated five additional elements with H2A.Z peaks that were present in the active centromere and uninterrupted with pre-domestication insertion dates and found that H2A.Z peaks were in the 5' LTRs of these elements, but again H2A.Z coverage was 20-35X lower than the control peak. In general, we did not see significant enrichment of H2A.Z in the 5'LTR of post domestication uninterrupted CR2s within or outside the active CENs. However, we saw low level H2A.Z coverage over uninterrupted, pre-domestication CR2

5’ LTRs in the active centromere. This suggested that H2A.Z did indeed localize to CR2 LTRs. This was further supported by B73 H2A.Z ChIP mapped to a consensus CR2 that show H2A.Z primarily associates with LTRs in the U3 region (**Figure 3.4**). Peaks that appeared across the polyprotein region may repress CR2 transcription similar to H2A.Z found in gene bodies [120]. Comparison of the maximum H2A.Z coverage across the consensus CR2 to the maximum coverage over the CEN5 control peak suggested that CR2s were not enriched for H2A.Z in B73 (*SI Appendix Table ST6*). It was likely that only a subset of the CR2 elements were associated with H2A.Z but, we were unable to determine specific CR2s enriched for H2A.Z due to the limited variation between recently inserted elements. Most CR2 elements inserted within the last 10 thousand years, and they have not yet acquired enough variation to be uniquely mappable [24]. Notably, transcriptional analysis of CRR2 (the CR2 ortholog in rice) found the element was transcribed in root, leaf and panicle tissues but this transcription was low and derived from relatively few CRR2 elements consistent with our conclusion that only a subset of CR2 elements may be transcriptionally activated by H2A.Z [124]. Regardless, the increase in CR2 repeat content at neocentromeres may in fact serve a functional role by providing a repeat substrate to efficiently repair double strand breaks frequently observed at centromeres [61, 121, 125]. It was conceivable that the acquisition of CR2 insertions may have further stabilize neocentromeres by increasing the repeat content, and simply pushing persistent H2A.Z loci or genes out of the active centromeres over time.

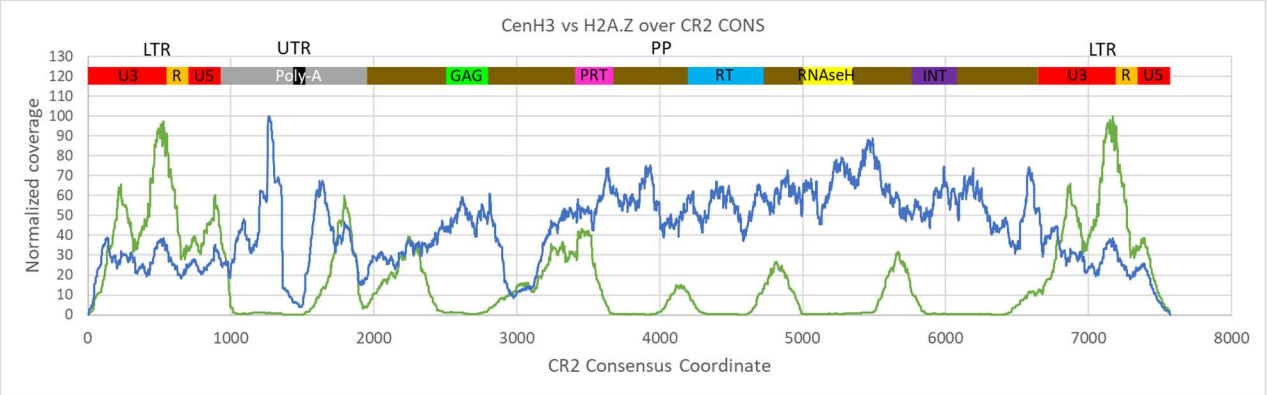


Figure 3.4. H2A.Z and cenH3 coverage over a CR2 consensus. The relative position within the CR2 is annotated with the long terminal repeats (LTRs), untranslated region (UTR), and Polyprotein (PP). The PP included five functional domains: the group specific antigen (GAG), Protease (PRT), Reverse transcriptase (RT), RNAseH and Integrase (INT). The LTRs were further subdivided into the U3 (contains the CR2 Promoter), R, and U5 subregions. H2A.Z and cenH3 coverage (B73) were obtained via bowtie mapping to the CR2 consensus and normalized to 100. H2A.Z peaks (green) were found in the LTRs and across the polyprotein. It was notable that H2A.Z peaks were generally not covered by the functional domains (except for the protease). CenH3 coverage (blue) was relatively consistent throughout.

3.4.6 H2A.Z epigenetic signal was resilient to cenH3 invasion

Notably some H2A.Z loci in the CEN5L and CEN5R regions were not associated with gene models or transcription (*SI Appendix Figure S6*). Comparisons of these isolated H2A.Z loci in maize to syntenic regions of the *Sorghum bicolor* physical map (Sorghum_bicolor_NCBIv3, PRJNA13876) suggested that some of these isolated H2A.Z loci were formerly associated with genes (**Figure 3.5**). The isolated CEN5 peak "o" was found to be separated from a parent gene. CEN5-k and CEN5-o were both orthologous to SORBI_3004G075000. In this case, the 5' H2A.Z peak (CEN5-o) was separated from the parent gene (CEN5-k) by a helitron related gene graveyard (*SI Appendix Text SX1, Figure S7-8*). The gene models corresponding to CEN5-k (v4 gene model Zm00001d015670) and it's ortholog (SORBI_3004G075000) corresponded to different HSPs suggesting one or both gene models were incomplete. Three additional examples of gene derived isolated H2A.Z peaks were uncovered in CENs 4, 9, and 10 **(***SI Appendix Text SX2, Figure S9*). In all cases, the loci were separated from their parent genes via transposable element insertion, or the parent gene was destroyed by deletion, or frameshift mutation. The fact that H2A.Z loci

were retained following the loss of the associated gene, suggested that H2A.Z was resilient to cenH3 invasion. H2A.Z peaks retained in active centromeres of B73 that were not associated with transcription (max raw RNAseq coverage <25) display a negative correlation between cenH3 and H2A.Z suggesting that in general, the two histone variants did not occur in the same nucleosome, though this inverse correlation may be skewed by nucleosome positioning (*SI Appendix Figure S11*).

Studies in *Arabidopsis* indicate the deposition of H2A.Z occurs independently of the cell cycle via the SWI2/SNF2-Related 1 (SWR1) complex while cenH3 was loaded in late G2 phase of the cell cycle [126, 127]. Additionally, studies in humans have shown that CENP-A (cenH3) variant histones are maintained at centromeres and are replenished after each cell cycle to maintain centromere integrity (this process is reviewed in Stellfox *et al.* 2013) [128]. The fact that H2A.Z maintenance was dynamic and affected by various environmental factors (temperature, water stress, pathogen infection, and nutrient stress) while cenH3 deposition is restricted to the G2 phase, would suggest that H2A.Z deposition can supplant cenH3 in active centromeres in some circumstances [120, 129-132]. The dynamic nature of H2A.Z loading and competitive interaction of the H2A.Z/cenH3 suggested that higher H2A.Z density would destabilizecentromeres. This was supported by studies in *S. pombe* which have shown H2A.Z hinders the CENP-A chaperone Scm3 (HJURP in vertebrates) leading to high rates of aneuploidy [115]. The maintenance of H2A.Z peaks in the active centromeres of maize suggested that like transcription, some H2A.Z was permitted in neocentromeres. However, the fact that we observed H2A.Z loci that were resistant to cenH3 invasion and cenH3 suppressed H2A.Z in active centromeres, suggested that the dominant epigenetic signal at a particular locus was variable depending on centromere position and H2A.Z maintenance at that locus. What makes a specific locus susceptible to cenH3 invasion was beyond the scope of these investigations.

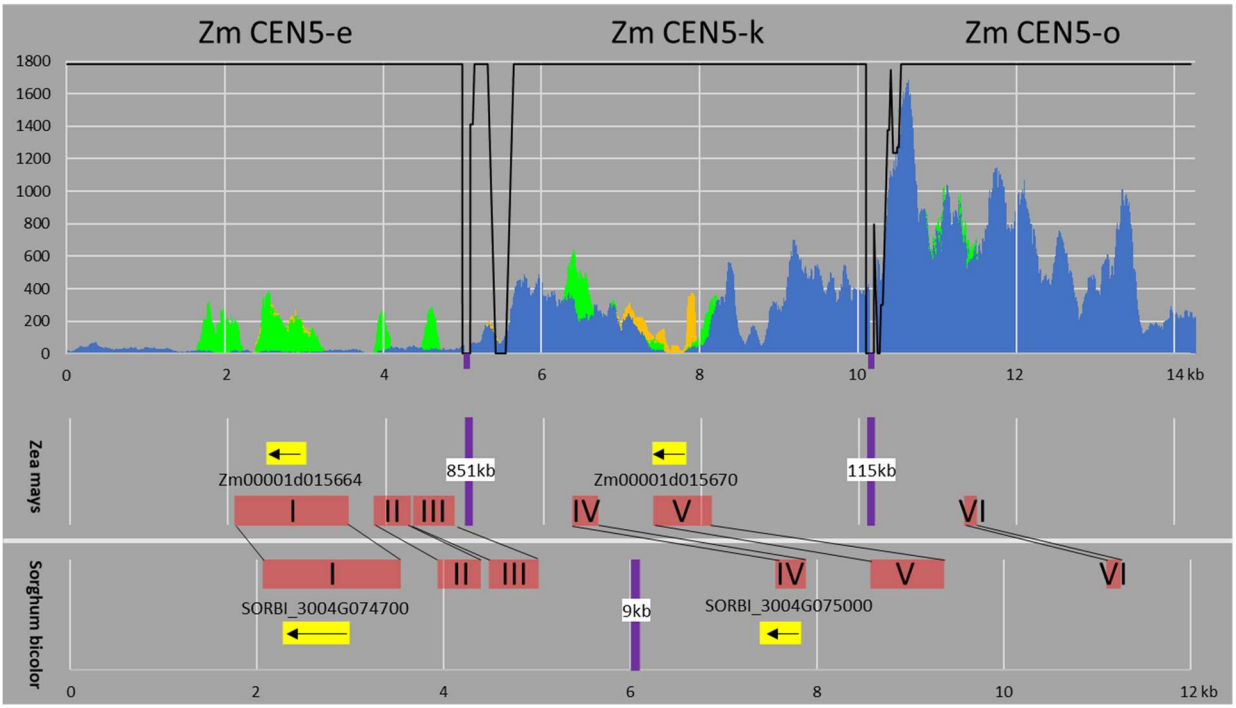


Figure 3.5. Isolated H2A.Z loci lacking RNA-seq can form due to gene damage. An example of isolated H2A.Z peak formation on chromosomes 5. The top panel contains a stacked chart of cenH3 (blue), H2A.Z (green) and RNAseq (orange) raw coverage in B73_v4CEN [133] for each window and is separated by a purple bar. Unique 101mer coverage was marked in black. The middle and bottom panel were schematic diagrams of corresponding regions in *Zea mays* and *Sorghum bicolor,* respectively. Each contained Gene models (yellow) and blast HSPs (red). Shared blocks of HSPs were connected by black lines. H2A.Z was associated HSP VI and separated from the *Zea mays* gene model Zm00001d015670 (v4 annotation) by an insertion of ~115kb. Additional examples on chromosomes 4, 9, and 10 can be found in *SI Appendix Text SX2, Figure S9*.

3.5 Discussion

3.5.1 Model of neocentromere positioning in CEN5

Factors including H2A.Z, transcription, and proximity to heterochromatin were implicated in neocentromere positioning and strongly suggests that the epigenetic environment is primarily responsible for the positioning of neocentromeres [115-117, 134]. Studies have shown that the loss of heterochromatin can considerably reduce neocentromere formation [135]. However, there are notable examples of neocentromere establishment in heterochromatin free regions of vertebrates, though further investigations have shown neocentromeres that lack pericentromeric heterochromatin associate

with non-pericentromeric heterochromatic regions [68, 116, 119]. Furthermore, irradiation mutagenesis studies in *Drosophila* have reported neocentromere formation only in segments proximal to active centromeres [117]. CEN5L and CEN5R are in the formerly pericentromeric regions of the maize ancestral CEN5M suggesting that pericentromeric heterochromatin likely played a role in the establishment of the neocentromere loci. This may be due to low levels of ectopic cenH3 proximal to the ancestral centromere, which have been observed in chicken DT40 cells and proposed to seed neocentromeres [116]. Neither CEN5L nor CEN5R are perfectly suited for centromeres due to the presence of persistent transcription and H2A.Z, but the low levels of transcription and H2A.Z in pericentromeres are preferred to the gene dense euchromatic regions. Over time these neocentromeres are improved by deletion/silencing of H2A.Z loci and the acquisition of repeats (CR2 insertions) which increase repeat content and potentially push genes and H2A.Z loci out of the active centromeres (see Figure 1 in Topp *et al.* 2006) [136]. It is notable that under rare conditions some transcription is tolerated, as is some H2A.Z. However, the mechanism by which specific centromeric H2A.Z peaks are retained over others is unclear. In the case of CEN5 where transcription is binding the ancestral CEN5M both upstream and downstream, a substantial loss of centC can destabilize the centromere as the cenH3 cannot spread into the surrounding transcribed regions. This resulted in CEN5M jumping to either CEN5L or CEN5R (Presting *et al*. 2018, see Figure 2. [121]). Our results – neocentromeres formed in regions with low transcription, neocentromeres had reduced H2A.Z, deletions of H2A.Z in CEN5L, H2A.Z suppress by cenH3 or be resilient to cenH3 invasion, and CR2s we not enriched for H2A.Z – suggested that neocentromere positioning in maize following disruption of the ancestral centromere was dictated by both the transcriptional and the epigenetic environment proximal to the ancestral centromere that resulted in a neocentromere positioned where it was most stable and least disruptive to the surrounding transcriptional landscape.

CHAPTER 4. Recombination near active centromeres is limited by selection and structural variation

4.1 Abstract

Zea mays was domesticated ~10,000 years ago and coincided with a loss of centC centromeric tandem repeats. This resulted in the centromeric histone cenH3 expanding into formerly euchromatic regions linking genes to the centromere. Here we investigated the relationships between recombination, genes and the neocentromeres. We mapped 104 recombination breakpoints across the 10 centromeres with the closest breakpoint to a centromere being in CML333 ~353kb downstream of CEN5M. Investigation of H2A.Z and recombination indicate H2A.Z was correlated with centromeric recombination. We investigated CEN2 in detail and observed three post domestication centromere haplotypes, and a TIL14 derived domestication locus that contains three transcribed genes flanking the centromere haplotype region. Structural variation in CEN2 included two pericentric, one paracentric, and one centromeric inversion, as well as two deletions and an insertion. The centromeric haplotype region of CEN2 was found to be defined by the CEN2 domestication locus and a haplotype specific insertion. Our results suggested that the selection on the CEN2 domestication locus led to diversity loss in the linked CEN2 resulting in the formation of a preferred centromeric haplotype. Comparison of domestication loci proximal to CEN2 and CEN5 suggested that the closer a selected region was to the active centromere, the less centromeric diversity was observed, likely the result of linkage drag effects on the selected regions.

4.2 Introduction

Maize domestication between 7.5 and 10 thousand years ago from the wild grass *Zea mays* subsp. *parviglumis* in the Balsas River Valley region of Mexico [2-4], led to a loss in genetic diversity relative to its wild counterpart. This loss was due to a genetic bottleneck and selection [9]. Such losses of diversity

related to domestication has been extensively documented in maize and many other systems including wheat, apricot, spurge, and shrimp [7-9, 137-140]. The intense selective pressure over loci during domestication often results in the formation of single haplotype regions that often occur over genes (domestication sweep). Considerable efforts have been expended to uncover and study these domestication loci, often linked to traits with agronomic value [10-14]. Intense selection also has the potential to decrease diversity in regions proximal to domestication loci due to genetic linkage that decreases the observed frequency of recombination flanking the domestication locus (linkage drag) [14, 141, 142]. Meiotic recombination is not only reduced proximal to domestication loci, but near centromeres as well.

Studies of human X chromosome trisomies have reported a correlation between centromere proximal recombination and chromatid non-disjunction [143]. A similar link was observed in *Saccharomyces cerevisiae,* with ~60% of non-disjunction errors being reported associated with a centromere proximal recombination [144]. This evidence suggests that centromere proximal meiotic recombination can be detrimental to organisms and may be selected against due to the correlation with segregation errors during meiosis [145]. Pericentromeric regions play a critical role in suppressing meiotic recombination proximal to centromeres. Studies in *Arabidopsis thaliana* have demonstrated that disruption of H3K9me2 (a histone variant commonly found in pericentromeres) and non-CG methylation increased the rate of centromere proximal recombination [146]. While recombination is limited in centromeres, H2A.Z has been shown to be positively correlated with meiotic recombination, and we have shown in Chapter 3 of this dissertation that H2A.Z loci are relatively common in maize neocentromeres (Chapter 3, [57, 58]). However, the relationship between H2A.Z and centromere proximal recombination has not been reported.

The reduction in recombination around the centromeres in maize manifests as large (>1Mb) recombination-free haplotype regions that include the active centromere but the diversity of haplotypes

varies between centromeres [24]. Analysis of centromeres in maize indicated 7 of the 10 chromosomes displayed reduced genetic diversity (only 1 or 2 haplotypes) among the NAM lines [24]. It is postulated that loss of the tandem repeat centC is due to inbreeding during domestication led to a destabilization of the ancestral centromeres. This resulted in the formation of neocentromere loci in formerly pericentromeric regions that links groups of genes to the centromere due to recombination suppression imposed by the centromere. Further, the observed diversity loss in maize centromeres is the result of selection for the centromere linked genes. This suggests a domestication locus proximal to a region with reduced recombination frequency (such as a centromere) has the potential to reduce diversity or even sweep the entire centromere region.

While traditional meiotic recombination is repressed in centromeres several mechanisms can introduce variation into centromere regions. Previous analysis of *Zea mays* centromere 10 reported a high frequency of double strand break events that have the potential to introduce major structural changes to the centromere indicating structural variation is of major importance to centromere evolution [61]. These breaks can alter tandem repeat arrays and result in large (>1Mb) pericentric inversions observed in maize genomes [106]. Other analyses have reported substantial gene conversion in maize centromeres [59]. Even centromeric lineages diverged by <10 thousand years have been shown to have distinct retrotransposon profiles [24]. Mechanisms like double strand break/repair, illegitimate recombination, retrotransposon insertion, and gene conversion, contribute to the variation in the centromere regions as do rare centromere proximal meiotic recombination events. The recent release of high-quality physical maps for the 26 NAM founder lines [106] presents an opportunity to examine domesticated maize centromeres at unprecedented detail. Specifically, the physical maps permit us to assess both genetic and structural variation as it relates to centromere evolution.

Here we investigated centromere proximal recombination, structural variation, diversity, and divergence in the context of CEN2 of *Zea mays* and compared it to the well documented CEN5. Recombination loci

were generally distant from active centromeres and recombination loci within active centromeres either occurred prior to the centromere occupying that site, or were in a maize line that did not have the active centromere at that locus. H2A.Z loci was found to be correlated with centromere proximal recombination in B73 and we found evidence of considerable structural variation in CEN2 that defines the boundaries of the centromere haplotype region. We also describe a pericentromeric domestication locus containing a domestication gene directly upstream of CEN2. Our results suggested that in the case of maize centromeres, genes under selective pressure that were genetically linked to centromeric regions is the force behind diversity loss and the tighter the linkage between the domestication locus and the centromere, the lower the observed centromere diversity.

4.3 Methods

4.3.1 CenH3, RNAseq and H2AZ mapping

See Chapter 3, section 3.3.2.

4.3.2 Genome Wide 101mer mapping

See Chapter 3, section 3.3.5.

4.3.3 Mauve alignments using whole genomes

Homologous regions between chromosomes were identified using MUMmer (-b –maxmatch –l 1000 –n –F) [107]. Regions with no mummer alignments were sub-aligned using less stringent parameters to check for inversions and distant sequence homology (-b –maxmatch –l 100 –n –F). Inversions were "corrected" to ensure collinearity with B73. The regions were aligned using Mauve, with the "assume sequence collinearity" parameter enabled [147]. Regions that initially did not align were found to be caused by indels and tandem repeats. For certain subregions the seed-family Mauve parameter was

employed to align more distant sequence. Sub-alignments were extracted from the Mauve output that were 1kb overlapping and present in all aligned physical maps moving 500nt per step and allowing up to 10 gaps among the lines. The K2p distance for each alignment was calculated with MegaCC [148]. Sub-alignments were organized into blocks by combining sub-alignments that were <10kb apart. The average K2p of all sub-alignments in each block was then calculated across the aligned region. These distance values were converted to divergence dates using a centromeric conversion factor.

4.3.4 Substitution rate calculation

The substitution rate was calculated using alignments extracted from regions present in 6 physical maps (B73, EP1, F7, PH207, W22, CML247) across the conserved regions of single haplotype centromeres 1, 4 and 6. Three regions previously reported as domestication loci were identified in B73 RefGen_v4 and homologous regions from the other five genomes were aligned with Mauve to determine the maximum distance (K2P) for each locus. In addition, to determine the optimal substitution rate for the centromere, the three centromeres previously reported to also be bottlenecked in domesticated maize [24] were identified and extracted in the other five genomes. The maximum K2P values were used to determine a substitution rate for each centromere based on this region diverging at the time of domestication, 10 ka. The weighted average of the rates was then taken resulting in the substitution rate of $7.3*10^{-8}$ (*SI Appendix Text SX3*). The regions analyzed are upstream of the *tb1* gene [149], a locus on the long arm of chromosome 10 [150] previously shown to be 1.2 Mb with a distant introgressed region in the middle [24] and the region downstream of the centromere of chromosome 5 (CEN5) named “Region D”. The *tb1* gene is at position 270,553,676-270,554,776 on chromosome 1 in the B73_v4 physical map [151]. The locus on the long arm of chromosome 10 was identified by mapping sequences from pzb03527 and pzb03542, the regions of low genetic diversity [150], on the B73 physical map. The “Region D”

domestication locus previously identified [24] was mapped onto the RefGen_v4 physical map (*SI Appendix Table ST9*).

4.3.5 Generating phylogenetic trees

To generate SNP trees, HapMap3 SNP data was obtained from Cyverse (/iplant/home/shared/panzea/hapmap3/hmp321/imputed/uplifted_APGv4). Fasta pseudo-sequence was constructed from the SNP data using a custom Perl script. The lines that contained insertions, deletions, heterozygous positions, and positions with no data, had these positions replaced with "N" in the fasta sequence. Trees were constructed from the fasta pseudo-sequence in MegaX (Neighbor-joining, K2p distance calculation, pairwise deletion). Trees were generated from Mauve physical maps alignments directly in MegaX (Neighbor-joining, K2p distance calculation, pairwise deletion).

4.3.6 HM3 SNP dating

Subregions of the single haplotype regions in CENs 1, 2, 4, and 6, were extracted from the 26 NAM lines using unique markers shared by all lines as anchors (**Figure 4.1***, Table ST8*). The regions from the physical maps were compared to B73 v5 using MUMmer (word length 100) to check for inversions. Inverted sequences were adjusted to reflect the orientation of B73. The adjusted sequences were aligned using Mauve (collinearity parameter enabled). The resulting Mauve alignments were manually edited in Geneious and validated using Junction Viewer annotation [108]. HapMap3 SNP alignments were generated for the regions matching the physical map alignments. K2p values were calculated using MEGAX (pairwise deletion) for the HapMap3 alignments and the physical map alignments which were plotted (*SI Appendix Figure S14*). The data follow a linear trend, and a linear conversion factor was generated weighted on the number of HapMap3 SNPs in each alignment (HM3_K2p/20.27979=APPROX_MAUVE_K2p). Only the CEN2 HAP1 slope was used in the weighted

conversion factor calculation as CEN2 HAP2 does not contain B73 (the HapMap3 reference). The converted HapMap3 K2p values were then converted to estimated divergence date with the centromeric conversion factor. It is notable that discrepancies were observed between HapMap3 SNP data and the physical maps which led to reduced correlation between HapMap3/Physical map divergence (*SI Appendix Text SX11*).

4.4 Results

4.4.1 Distribution of meiotic recombination in maize centromeres

While traditional meiotic recombination is rare near centromeres studies in maize have shown variation is introduced to centromeres predominantly via double strand break induced rearrangement, gene conversion events, or illegitimate recombination [59, 61]. To determine the distribution and extent of recombination events in the centromeres of domesticated maize we evaluated HapMap3 SNP data across the 10 centromeres in 73 maize lines. SNP analysis resulted in the identification of 104 recombination breakpoints proximal to centromeres (**Figure 4.1,** *SI Appendix Text SX11*). Using these recombination breakpoints, we defined the centromeric haplotype regions among domesticated maize (NAM lines and BKN landraces) for the single and two haplotype centromeres observed in Schneider *et al.* 2016 that ranged from 6.2 to 12.3 Mb. Recombination breakpoints were observed in CEN4R and CEN5R, however, investigation of these breakpoints indicated that the recombinations took place prior to the centromere occupying that locus or the recombination was present in lines that do not have the centromere at that locus indicating that we observed no recombination breakpoints within an active centromere (**Figure 4.1**). However, the closest recombination to an active centromere was observed just 426kb upstream of the active CEN5L in M162W. Our data suggested that recombination this close to an active centromere was rare.

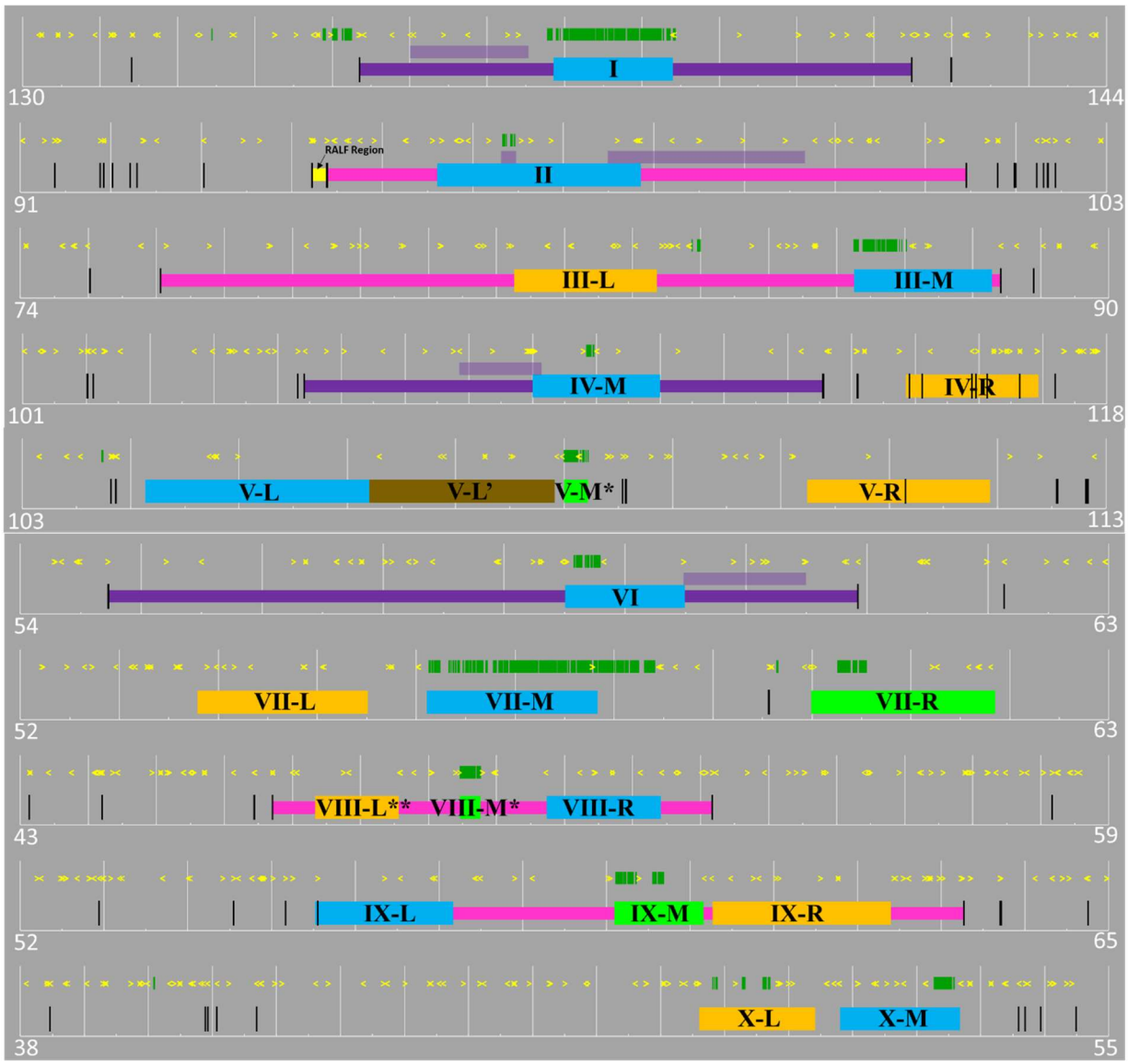


Figure 4.1. Overview of B73 v5 centromeres. The active B73 centromere locus marked light blue, with the alternative centromere loci determined from cenH3 ChIP-seq marked in light green, orange, or brown. Non-B73 centromeres were identified via cenH3 ChIP-seq in their respective NAM lines and MUMmer (Word length 100) was used to identify the B73 v5 coordinates. Recombination intervals identified in HapMap3 and superimposed onto B73 v5 marked with black lines. CentC was denoted with green lines and B73 v5 gene models were marked with a yellow arrow indicating the gene model orientation. Centromeres previously defined as single haplotype among the NAM marked in purple (CEN1, CEN4, CEN6). Solid purple bars denote the single haplotype regions among the NAM and BKN lines. In two, the haplotype centromeres reported in Schneider et al. 2016 (CEN2, CEN3, CEN8 and CEN9,) marked in pink. The yellow bar in CEN2 represents the CEN2 domestication locus. Haplotype boundaries for multi-haplotype centromeres (CEN5, CEN7 and CEN10) were not defined. Regions from

which manually curated Mauve alignments were generated are shaded in purple (CEN1, CEN2, CEN4 and CEN6.

* The active centromere was entirely within the CentC region and was defined as the primary B73 CentC region

** The downstream ~713,469nt of CEN8-L were not present in B73 v5 at MUM100 due to structural variation.

4.4.2 H2A.Z was correlated with centromere proximal recombination in B73

Studies of meiotic recombination in fission yeast has shown that while H2A.Z is not necessary for meiotic recombination, it does promote meiotic double strand break formation by regulating chromatin architecture [57]. In contrast, studies in *Arabidopsis* have shown meiotic recombination hotspots are correlated with H2A.Z and H3K4me3 histone variants [58]. To determine if H2A.Z and recombination were correlated near centromeres, the distribution of H2A.Z in B73 over each of the recombination intervals were identified in the HapMap3 SNP data. The data showed that 87 of the 102 identified breakpoint intervals contained an H2A.Z locus (~85%), with only 1.2% of the centromeric sequence covered by H2A.Z, 10-fold higher than background. It was found that 76 of the breakpoints have near gene models, and only 40 of those gene models have associated RNAseq coverage (root, shoot, ear, and tassel tissues). Furthermore, there were 17 instances of recombination being associated with an H2A.Z locus that lacked a nearby gene model or expression. The data suggested that H2A.Z, not transcription or the presence of gene models, was correlated with centromere proximal recombination.

4.4.3 CEN2 had three post-domestication haplotypes among domesticated maize

To determine the relationships among maize lines within CEN2 we analyzed a 2.17 Mb sub-region of CEN2 defined by a centromeric HAP1 inversion upstream, and a pericentromeric HAP1 inversion downstream denoted as REGION13_SYN. A tree was generated in MegaX [152] from HapMap3 data [153] corresponding to the closest internal SNPs of REGION13_SYN for 73 maize lines (26 NAM parents,

7 improved lines, 23 landraces and 17 teosinte lines) (*SI Appendix Figures S12, S33 and S34, Table ST7*). HapMap3 SNP alignments were dated using a HapMap3 conversion factor to define the array of post domestication haplotypes (*SI Appendix, Text SX3, Figures S14-S15***).** Three post domestication haplotypes resolved among domesticated maize (**Figure 4.2**) differed slightly from the two haplotypes predicted by Schneider *et al*. (2016) due to our inclusion of landraces [24]. Haplotype 3 (diverged from TIL25 ~6.4ka, **Table 4.1**) was found only in 2 landraces (BKN029 and BKN035) and because we lacked a physical map representing this haplotype, we focused primarily on HAP1 and HAP2. Haplotype 1 (diverged from TIL05/07 ~18.0ka) included B73, five other NAM parents and one landrace (BKN025). It was expected the landrace BKN025 was basal to the HAP1 clade, it was instead derived, suggesting BKN025 may have acquired it's CEN2 from a modern line. Haplotype 2 was favored haplotype with 47 lines including 20 NAM parents, 7 Improved lines and 20 landraces. TIL14 was basal to the HAP2 clade indicating the HAP2 CEN2 was likely TIL14 derived and estimated to have diverged ~10.3ka. The HAP2 clade was further subdivided into HAP2A, HAP2B-ni, HAP2B-i and HAP2_BASAL groups. HAP2A was the most frequent CEN2 haplotype among the NAM parents (12 lines) and estimated to have diverged ~5.0ka. The HAP2_BASAL group contained 4 NAM lines (CML52, NC358, Oh43 and CML277) estimated to have diverged ~8.1ka. HAP2B can be subdivided into two subgroups HAP2B-i and HAP2B-ni estimated to have diverged ~5.2ka. HAP2B-i was found to have a pericentric inversion, not present in the HAP2-ni clade. To confirm the divergence estimates obtained from HapMap3 data, a mauve alignment [147] for the 26 NAM lines was generated from REGION13_SYN. Analysis of divergence dates suggested HAP1 and HAP2 diverged by ~163.1ka (*SI Appendix Text SX4, Figure S13*). Additional mauve sub-alignments were generated to determine divergence dates flanking the CEN2 non-recombinant region (**Figure 4.3**). While most of CEN2 was well represented with mauve divergence data, there were four regions across CEN2 for which we had reduced, or no alignment data (**Figure 4.3**). Region 1 had no data and was caused by sequence that was too diverged to align. The absence of data in region 2 was caused by poor alignment

over CentC among the 26 NAM lines. Region 3 corresponded to a HAP1 inversion that required a subalignment with more stringent parameters, and region 4 corresponded to a HAP1 specific indel.

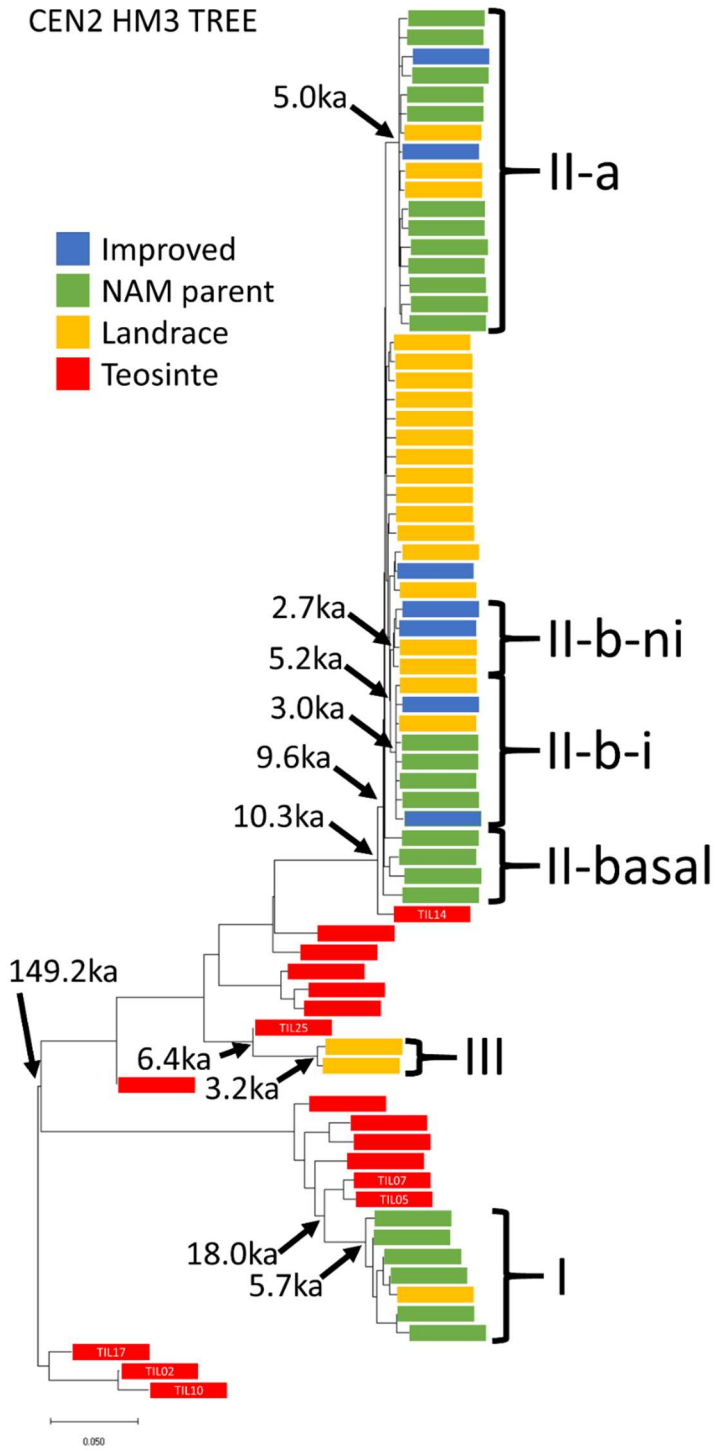


Figure 4.2. CEN2 HapMap3 SNP tree for REGON13_SYN. This tree was generated from Hapmap3 SNPs between 97,329,570 – 99,494,065 (v4). Teosinte, landraces, improved lines and NAM parents marked in red, orange, blue and green, respectively. TILs that gave rise to the CEN2 haplotypes and the rooted TILS are labeled white. Lines corresponding to major haplotypes are bracketed. The estimated divergence (ka) was calculated from HapMap3 SNP alignments using a weighted conversion factor. This tree displaying the line names found in *SI Appendix Figure S12*.

REGION	REGION13_SYN	REGION13_SYN
DATASET	**HM3**	**PHYSICAL_MAP**
HAP2 vs HAP1	149.2	163.1
HAP2 and TIL14	10.3	X
HAP3 and TIL25	6.4	X
HAP1 and TIL05/07	18	X
HAP2	9.6	10.3
HAP2A	5	2.3
HAP2B	5.2	X
HAP2B-i	3	1.1
HAP2B-ni	2.7	X
HAP2_BASAL	8.1	10.3
HAP1	5.7	5.8

Table 4.1. Estimated divergence dates for CEN2 using Mauve and HapMap3 SNP data. An "X" indicated the estimated date was not available for that comparison. Comparisons of HapMap3/Physical map divergence for single haplotype centromeres CEN1, CEN4 and CEN6 can be found in *SI Appendix Table ST18*. All values represent divergence date in thousands of years (ka).

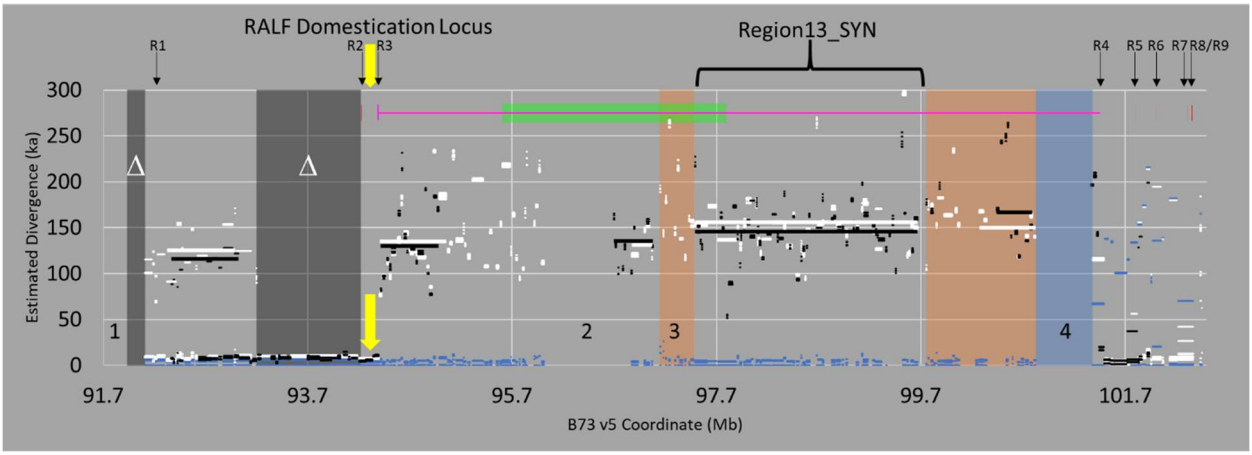


Figure 4.3. An overview of CEN2 divergence among, NAM lines with respect to the B73 v5 physical map. Estimated divergence date for MAUVE alignments shared by all 26 NAM lines across the CEN2 region. Haplotype 1 lines, colored blue and remain closely related to the reference (B73) across the Centromeric haplotype region. Haplotype 2 lines are colored white and are generally far more diverged from the reference through all 26 NAM lines diverged close to domestication in the CEN2 domestication locus (yellow arrow). Black lines represent a secondary mauve alignment superimposed onto B73 v5 coordinates from B73 RefGen_v4CEN, Mo17, and the short read physical maps for CML247, PE0075, PH207, EP1, DK105, and F7. The long thick white and black lines represent average date, weighted on sub-alignment length for the 26 NAM lines and short read physical maps respectively. Deletions in NC358/Oh43 are shaded black (*SI Appendix Text SX5*). Inversions in HAP1 lines with respect to HAP2 lines shaded orange. The active centromere (B73) shaded green. Recombination loci R1-R9 marked as red lines. The numbered regions had no alignment data. The CEN2 haplotype region was defined by the CEN2 domestication locus upstream and by the HAP1 indel downstream (pink line).
Δ) Region deleted in NC358/Oh43. Data shown was only among the remaining 24 NAM lines (seedfam was used, region was not present in NC358/Oh43 genome verified by MUMmer word length 200).
1) Sequence too diverged no matches
2) Poor alignment due to CentC tandem repeat
3) Inverted in the HAP1 lines (seedfam was used)
4) ~416,620nt of sequence present in HAP1 but not present in HAP2 (MUM200 coordinates, B73 v5, 100,823,893 – 101,240,513, Ky21, 103,707,225 – 103,709,236)

4.4.4 There was extensive structural variation in CEN2 among the NAM lines

Between regions 1 and 3 were two deletions shared by an NC358/Oh43 (deltas in **Figure 4.3**) the downstream deletion only contained data from 24 NAM lines excluding Oh43 and NC358 (*SI Appendix Text SX5*). Remarkably these 24 lines were estimated to have a maximum divergence of ~10.9ka. Between this deletion and the start of the centromere haplotype region was a ~164kb locus where all NAM lines were <10ka diverged. This post domestication divergence date suggested this region may be under selective pressure (i.e. a domestication locus). In addition to the deletions and domestication locus, there were four major inversions found among the NAM lines in CEN2 including the HAP2B-i and CML52 pericentric inversions [106], one centromeric HAP1 inversion, and one HAP1 paracentric inversion. The HAP2B-i inversion was by far the largest comprising 26.7Mb (US 93,897,775-93,897,804, DS 120,656,429-120,656,430, B73 v5 coordinates *SI Appendix Table ST14*) in four NAM lines (CML322, Oh7B, P39, Il14H) which was previously observed in Hufford et al. 2021 [106]. Analysis of GSS data mapped to CML322 (HAP2B-i) and Ky21 (HAP2A) (Bowtie, -m 1 -v 1 -X 2000) indicate that modern lines

DK105, and W22 as well as landraces BKN016 and BKN026 contain the inversion. The related HAP2B-ni lines (F7, EP1, BKN027, BKN015) did not contain the inversion. Investigation of the inversion breakpoints revealed a 10.5kb partial duplication of Prem retrotransposon derived sequence just upstream of the 5' breakpoint. Additionally, during the inversion, a 6.5kb partial Cinful element was incorporated onto the 3' end of the inverted sequence, downstream of the duplicated partial Prem. At the downstream breakpoint we observed an incorporation of a ~28kb (27,788nt) cluster of nested Cinful elements in the HAP2B-i lines. Inspection of syntenic genes shared by the NAM lines indicated the HAP2B-i inversion breakpoints occurred in gene depleted regions (*SI Appendix Figure S22*). Analysis of the syntenic relationships with *Sorghum bicolor* indicated these regions corresponded to ancestral centromere loci (*SI Appendix Text SX6, Figure S18*).

The CML52 pericentric inversion was the second largest CEN2 inversion comprising 8.7Mb (94,769,624-103,446,053 CML52 coordinates), also observed in Hufford et al. 2021). Both the HAP2B-i and CML52 inversions have taken place since domestication (<10ka), but there were also two pre-domestication HAP1 specific inversions. The larger paracentric inversion was comprised of 1.1 Mb in the non-recombinant region downstream of the active CEN2 (B73 v5 99,753,945-100,826,519). Adjacent downstream to this inversion was a 553,497nt indel present in HAP1 and absent in HAP2 (B73 v5 100,826,520-101,380,017). The smallest inversion (0.3 Mb) was also HAP1 specific and was within the active B73 CEN2 (B73 v5 97,142,844-97,481,242). Previous studies in teosinte have shown that structural variation (an inversion on chromosome 1) was strongly correlated with altitude suggested potential adaptive evolution [154]. However, the assessment of geographic and bioclimatic data associated with these centromere variants was beyond the scope of these works, thus we cannot report whether any of these structural variants were adaptive.

4.4.5 There was a tight linkage between the CEN2 haplotype region and the CEN2 domestication locus

Analysis of HapMap3 data and structural variation among the physical maps resulted in 22 subregions flanked by either recombination, inversion, or indel breakpoints. Trees were generated for the 73 maize lines in each region, and the closest TIL was determined for the 56 domesticated maize lines. Sharp increases or decreases in divergence date relative to the B73 v5 reference correspond to recombination breakpoints identified in HapMap3 and confirmed using NAM physical maps (**Figure 4.3**). Nine recombination loci existed among the NAM lines in the region from 91.7Mb – 102.5 Mb (**Figure 4.4a***, SI Appendix Figure S20*). The centromeric haplotype region was defined by recombination loci R3 and R4 and contained REGION13_SYN as well as the active centromere. This region was ~7Mb and spanned from 94,391,474 - 101,455,006 in B73 v5 (Bracketed in **Figure 4.4a**). Recombination locus R3, was contained within the B73 v5 gene model Zm00001eb087250 that contained four recombination breakpoints (HAP2_domestication/HAP1_CEN, HAP2_domestication/HAP3_CEN, and a CML277/BKN009 HAP1 introgression). CML277 was found to contain a 305nt introgression of HAP1 sequence over the gene (*SI Appendix Table ST12*). HapMap3 SNPs suggested this introgression was shared by BKN009 that included only 22 SNP positions. BKN009 had a TIL06 recombination on the downstream side of the centromeric haplotype region indicating only the upstream introgression was shared by both BKN009 and CML277. It was remarkable that four recombination events took place within this gene considering there were six B73 v5 gene models between R3 and the active centromere in B73 (*SI Appendix Table ST13*). This suggested R3 may be the only stable recombination site between the domestication locus and the centromere haplotype region indicating a tight linkage.

4.4.6 The CEN2 domestication locus was acquired by HAP1 and HAP3 from HAP2

Just upstream of the non-recombinant region, the divergence dates drop to a maximum of 8.3ka among the NAM parents (**Figure 4.3 and 4.4a** yellow arrows). The locus spanned from 94,224,636 to 94,389,206

in B73 v5 and is 164,570nt defined upstream by the NC358/Oh43 breakpoint and downstream by the HAP1/HPA2 breakpoint. This domestication locus contained two gene models in v5 (Zm00001eb087230 and Zm00001eb087240 B73 v5). Zm00001eb087240 was not found to be expressed in and was also found to have orthologs in only 13 NAM parents. Furthermore, Zm00001eb087240 was not present in v3 or v4 annotations. However, Zm00001eb087230 (Zm00001d004231 B73 v4 annotation) was a full-length, rapid alkalization factor gene homolog that was expressed in multiple tissues (Qteller) and was present in all 26 NAM lines.

The CEN domestication locus was flanked by two genes (Zm00001eb087220 and Zm00001eb087250 in B73 v5, Zm00001d004230 and Zm00001d004234 in B73 v4) and contained the recombination breakpoints that defined the domestication locus region. Zm00001eb087220 and Zm00001eb087250 were both found to contain START_ArGLABRA2 and homeobox domains. Analysis of Zm00001eb087220, Zm00001eb087250 and Zm00001eb087230 indicated all three genes were associated with H2A.Z loci though only Zm00001eb087230 was expressed (root, shoot, ear, and tassel tissues). The upstream region of the active centromere, cenH3 was only found at background levels in the domestication locus (**Figure 4.4b**). Comparison of the region in *Zea mays* to *Sorghum bicolor* indicated considerable conservation between the two genomes for these three genes. Though the alternate order in sorghum suggests either an inversion, or incorrect contig order in the BAC sorghum assembly (**Figure 4.4c, and 4.4d**).

HapMap3 SNPs were identified that correspond to exons for the flanking homeobox genes that were within the CEN2 domestication region. In Zm00001eb087220 exons 4-8 were within the domestication region and in Zm00001eb087250 exon 1 was included in the domestication region. Zm00001eb087250 exon 1 did not contain any part of a functional domain, however Zm00001eb087220 exons 4-8 contained the entire START_ArGLABRA2 domain. Indeed, Zm00001eb087230was the only full-length,

transcribed gene model in the CEN2 domestication region, but it was possible that the Zm00001eb087220 exons 4-8 were under selective pressure.

Investigation of HapMap3 SNPs for 73 maize lines suggested all domesticated maize form a single clade along with TIL14 and TIL10. TIL10 was found to share the domestication locus region and flanking sequence, (*SI Appendix Figure S20* orange shade) but it contained a genetically distant teosinte centromere. Analysis of HapMap3 SNPs in the CEN2 domestication region for TILs 14 and 10 revealed that TIL10 was more closely related to modern lines NC358/Oh43 while TIL14 was more closely related to landraces BKN030 and BKN033 and suggested that TIL14 was the more basal TIL and thus the more likely progenitor of the domestication locus while TIL10 may have acquired the domesticated form of this region from an NC358/Oh43 relative (*SI Appendix Text SX7*). The fact that TIL14 was the progenitor both upstream and downstream of the CEN2 domestication region suggested that HAP2 was the donor of this region to HAP1 and HAP3 acquired it via rare centromere proximal recombination.

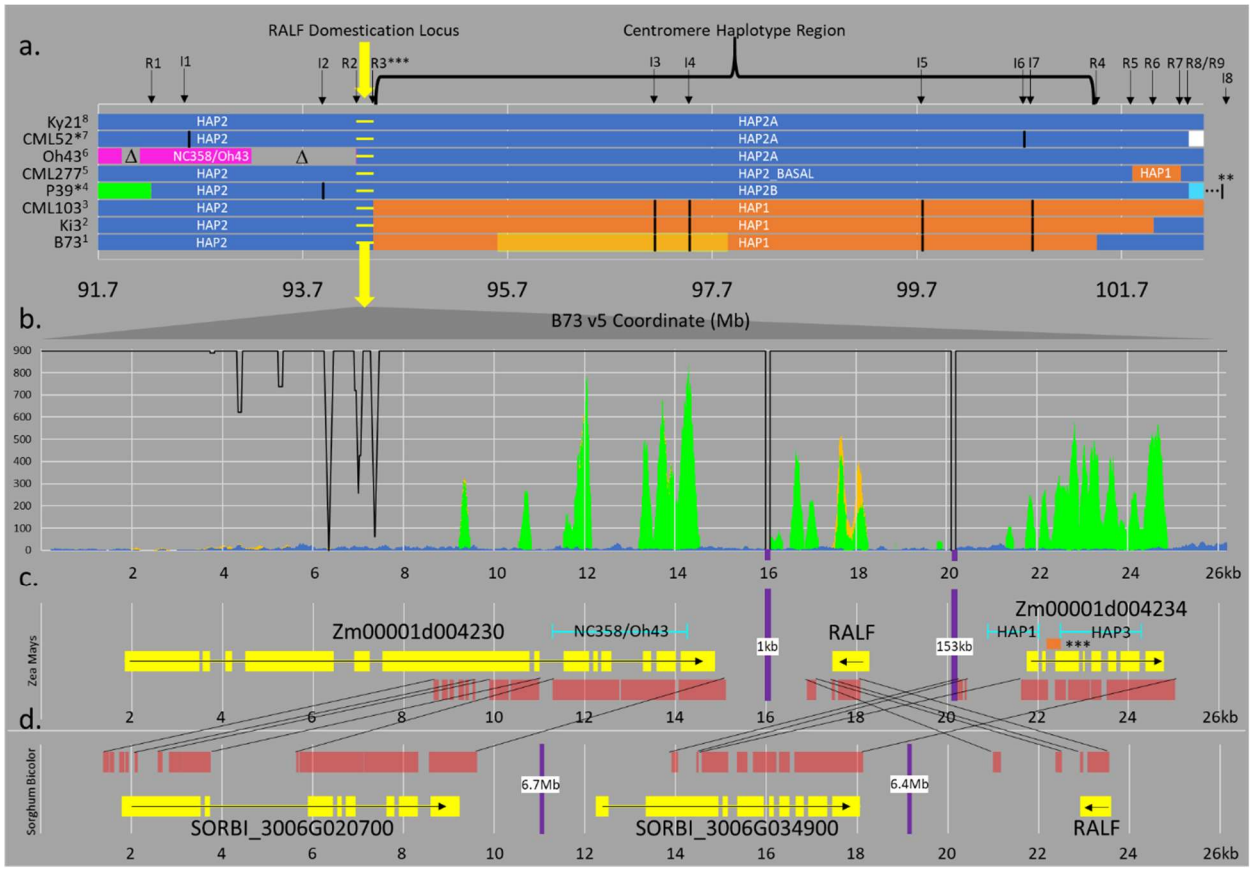


Figure 4.4. An overview of CEN2 recombination among the NAM lines and a comparison of the CEN2 domestication locus and the two flanking homeobox genes in *Zea mays* and *Sorghum bicolor*. Panel **a.** a schematic representation of recombination among the 26 NAM lines superimposed on B73 v5 coordinates. Lines with shared lineage share the same color with only one representative of each type shown, listed below 1-8 are the lines sharing the CEN2 recombination profiles. Recombination loci R1-R9 were marked with black arrows at the top of the chart. Inversion breakpoints (I1-I4) for CML52 and HAP2B are marked with black lines. The active B73 v5 centromere locus shaded green. The CEN2 domestication locus was marked with a yellow arrow. Panels **b**, **c** and **d** comprised of three regions covering the *Zea mays* v4 gene models and their orthologs in Sorghum bicolor. Panel **b.** contains CenH3 (blue), H2A.Z (green) and RNAseq (orange) coverage in B73 for each window separated by a purple bar. Unique 101mer coverage marked in black. Panels **c.** and **d.** panel schematic diagrams of corresponding regions in *Zea mays* and *Sorghum bicolor,* respectively. Each contains Gene models (yellow) and blast HSPs (red). Shared blocks of HSPs were connected by black lines. Panel e. shows the recombination breakpoint windows for the NC358/Oh43, HAP1 and HAP3 recombination breakpoints (light blue lines). A 305nt introgression of HAP1 sequence into CML277/BKN009 (orange bar).

1) B73, B97
2) Ki3, CML228, CML69
3) CML103
4) P39, Il14H, Oh7B, CML322
5) CML277
6) Oh43, NC358
7) CML52

8) Ky21, NC350, Ki11, Mo18W, Tx303, CML333, M37W, Ms71, CML247, Tzi8, HP301, M162W
Lines marked "" were those that contained inversions marked I1-I4
** The 26.7Mb HAP2B inversion extended beyond the ends displayed (120,656,429-120,656,430, B73 v5 coordinates).
***CML277 contained a 305nt introgression of HAP1 sequence between the CEN2 domestication region and non-recombinant region (*SI Appendix Table ST12*).

4.4.7 The CEN2 domestication locus and structural variation defined the CEN2 haplotype region

Remarkably the recombination breakpoints that defined the centromere haplotype region (R3 and R4 **Figure 4.2**) in CEN2 are themselves defined by the domestication locus upstream and the HAP1 specific indel downstream. Investigations of syntenic genes in the domesticated CEN1 indicate the centromeric haplotype region is defined by an inversion found in M162W and Ky21. This suggests that the centromere haplotype regions defined are defined by structural variation, and domestication. It is notable that structural variants do not appear in all haplotype regions among the NAM lines though it was possible that structural variation exists in other maize lines not represented here. Domestication influencing the haplotype region boundaries was not entirely surprising as studies in maize and tomato have shown that lines which have undergone selective sweeps (domestication) have a reduced meiotic recombination rate relative to wild lines proximal to the swept loci (linkage drag) [150, 155]. Moreover, studies in *Arabidopsis,* humans, and *Drosophila* have observed inversion induced linkage disequilibrium resulting in reduced meiotic recombination rates and occasionally a decrease in fitness in heterozygous individuals [156-158]. These results do suggest that centromere haplotypes in maize were positioned where they were due to linkage effects from domestication and/or structural variation that influenced centromere proximal recombination.

4.4.8 Recombination in CEN5 and selection for Region-D

A 25Mb region of chromosome 5 encompassing CEN5L through region D was analyzed for recombination and 28 recombination breakpoints among the NAM and BKN lines broke the window into

27 sub-regions (**Figure 4.4**). Thirteen of these breakpoints occurred between CEN5R and region-D. While recombination was generally suppressed near centromeres, members of the brown lineage (CML277 and CML322) shared a recombination breakpoint 103kb downstream of CEN5M. A second centromere proximal recombination occurred in the CML333 lineage ~353kb downstream of CEN5M and was unexpected given the CML333 active centromere rests in the CEN5M locus. Region-D was found between 118.1 and 123.3Mb. Region-D was interrupted by a 1.03Mb introgression of sequence in the P39 and IL14H lineages, likely acquired from TIL01 or TIL03 about 5.4ka. This region was estimated to have diverged from the domestication clade 72.4ka. The flanking regions were estimated to have diverged between 12.2 and 13.7ka among the NAM and BKN lines. Previous analysis of region-D suggested that rare recombination in region-D became fixed among domesticated maize [24]. Analysis of ancestral TILs across the CEN5 region indicated that an introgression of TIL09 into TIL14 was fixed among the NAM lines downstream of the TIL01/03 introgression. In the multi-haplotype CEN5 we observe a scenario that differed from CEN2. There was selective pressure on the domestication locus region-D, separated from the CEN5L-R region by a recombinant region. The large distance between region-D to the CEN5L-R region (~6Mb) and the resistance of centromeric regions to recombination, suggested that the various CEN5 haplotypes persisted in the germplasm because region-D was not tightly linked to CEN5 (i.e. no selective pressure/linkage drag on CEN5 haplotypes). As the selective pressure was not within or directly adjacent to the CEN5 haplotype regions, we did not observe a major loss of diversity such as observed in CEN2, CEN1 and CEN4.

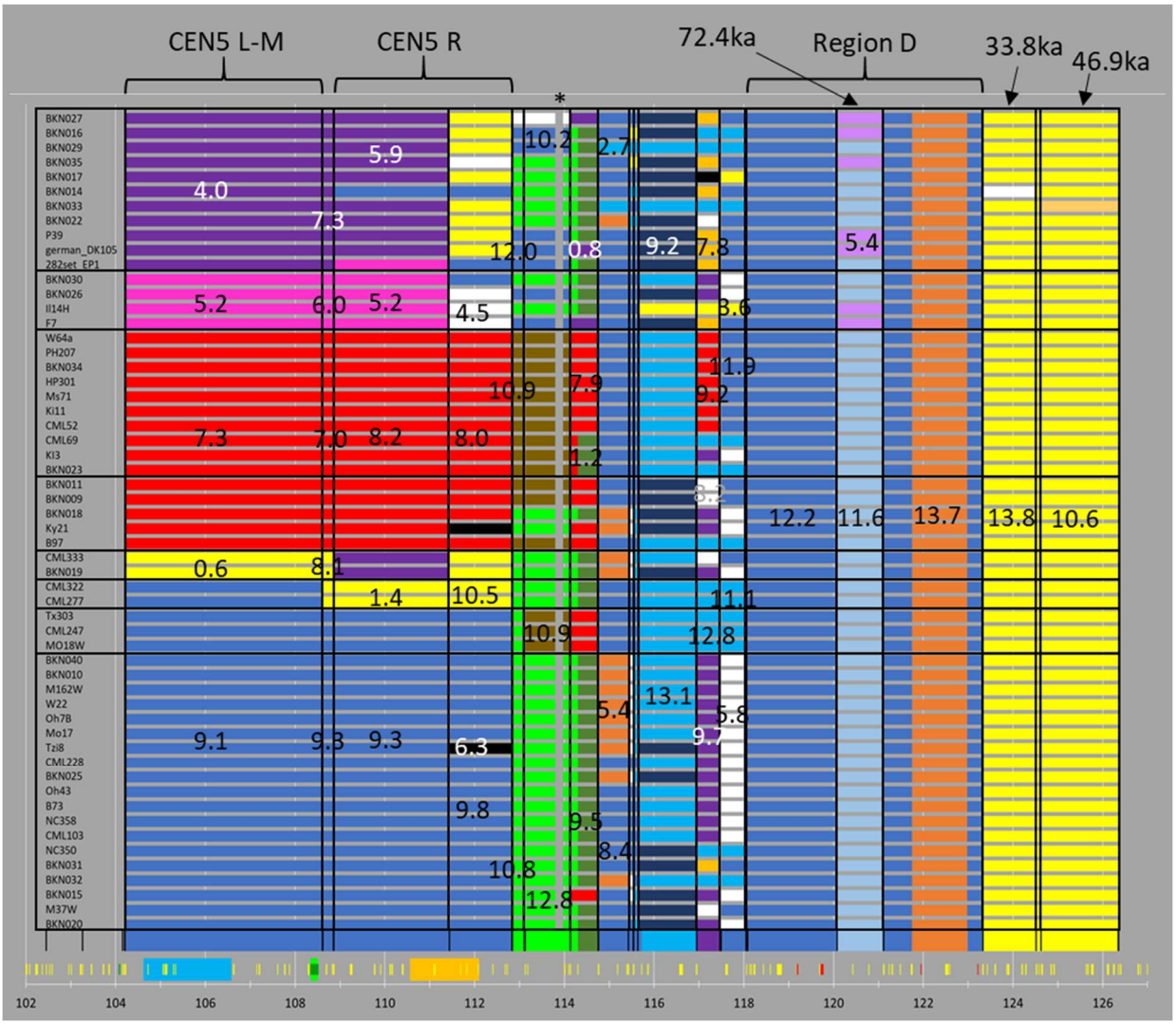


Figure 4.5. Recombination across the multi-haplotype CEN5 and region D among domesticated maize. Recombination was identified using HapMap3 SNP trees generated from regions between breakpoints marked as vertical black lines. Groups of lines corresponding to the blue, orange, brown, CML333, green, red, IL14H and P39, CEN5L-M lineages have been separated into blocks with horizontal black lines. Lines that shared the same ancestral TIL share the same color. The estimated divergence date within each color group marked for 24 major regions across CEN5 and region D. Estimated maximum divergence date between all lines for the introgression in region D and the flanking downstream regions marked above the chart. The bottom row denotes the positions of protein coding genes (yellow lines), genes in region D that were likely under selection (red), and CentC (green lines). The regions shaded light blue, light green and light orange indicated the positions of the active CEN5L, CEN5M and CEN5R respectively. Dates colored white, black, or grey for visibility.

*This region is left blank as it represents a misassembly in B73 RefGen_v4 confirmed by Mo17 [159] and RefGen_v2 (PRJNA10769).

4.4.9 Candidate domestication genes in CENs 1, 4, and Region-D

Gene models from the parts of region-D with divergence dates near 10ka among the NAM and BKN lines were examined to determine the key genes potentially under selection. A total of 31 gene models were evaluated for various criteria suggesting selection including, but not limited to, SNPs among domesticated maize, CIS regulatory effects, differential expression, previously identified domestication genes, possible promoter changes, and paralogs. In the CEN5 domesticated region-D upstream of the TIL01/03 introgression three genes on the downstream side were likely candidates (Zm00001d015785, Zm00001d015788, Zm00001d015789). In the domestication region downstream of theTIL01/03 introgression we found four gene models that appeared to be under strong selection (Zm00001d015809, Zm00001d015810, Zm00001d015811, Zm00001d015814). Trees obtained from the grammene database (http://www.gramene.org/) indicated Zm00001d015810 and Zm00001d015811 were a single gene coding for an alpha/beta-Hydrolase. Interestingly, the gene models that flanked Zm00001d015788 and Zm00001d015810-11 contained a high concentration of domestication SNPs though Zm00001d015788 (proteasome component4) and Zm00001d015810-11 and the flanking 10kb contain very few domestication SNPs. The flanking genes all contained >10 domestication SNPs within 10kb of their 5' start, and all but Zm00001d015809 were shown to have possible promoter changes, but only the internal Zm00001d015788 and Zm00001d015810-11 were previously called as domestication genes that had evidence of CIS regulatory effects. Zm00001d015788 and Zm00001d015810-11 may be the key genes driving selection of region-D. Additionally, the Zm00001d015788 and Zm00001d015810-11 genes were acquired from TIL14 and TIL09 respectively suggesting that a rare recombination between the two TILs became fixed in domesticated maize. Potential candidate genes were also identified in CEN1 and CEN4. In CEN1 we observe three gene models (Zm00001d030468 and Zm00001d030470/71) that contained domesticated SNPs and were highly expressed in anthers and embryo tissues respectively. Moreover Zm00001d030470/71 was previously observed to have CIS

regulatory effects [160]. In CEN4 Gene models Zm00001d050635 and Zm00001d050649 (ZCN2) were genes of interest. Zm00001d050635 has been reported to be differentially expressed between teosinte and maize, had no paralog and was highly expressed in embryo tissue. ZCN2 also had no paralog and was expressed at low levels in various tissues but overexpression studies in maize had shown ZCN2 affects flowering time and inflorescence architecture, both key domestication traits [161]. Validation of the candidate genes would require mutagenesis experiments (i.e. CRISPR knockouts, knockdown, or overexpression using a reverse genetics approach) to evaluate phenotypic effects. However, these experiments were beyond the scope of these investigations.

4.5 Discussion

4.5.1 Model of CEN2 evolution

Our analysis suggests that CEN2 HAP2 was the ancestral haplotype containing the linked CEN2 domestication locus. Selection on this locus during domestication resulted in the formation of the preferred CEN2 haplotype. Simultaneously, rare recombinations between HAP2/HAP1 and HAP2/HAP3 fixed this locus among domesticated maize. The CEN2 haplotype region itself was a result of reduced recombination due to selection on the domestication locus upstream, the two HAP1 specific inversions, and the HAP1 indel downstream of CEN2. Only one recombination locus (R3) was present between the CEN2 domestication locus and the CEN2 haplotype region, suggesting tight linkage between the two blocks. The selective pressure on the CEN2 domestication locus coupled with the linkage to CEN2 resulted in the formation of a preferred CEN2 (HAP2) and an overall decrease in CEN2 diversity.

4.5.2 Model of centromere haplotype formation in maize

Our results i) H2A.Z was correlated with centromeric recombination, ii) there was a reduction in diversity and iii) a preferred haplotype in CEN2, The CEN2 domestication region was selected and tightly

linked to the centromere, and the extensive recombination between the domestication locus Region-D and the CEN5 centromere region, suggested that domestication and linkage drag have heavily reduced the diversity in maize CEN2. However, these forces were likely playing a role in the single haplotype domesticated centromeres 1, 4, and 6. It is possible that in a single haplotype centromeres such as these, that genes under selection were inseparable from the centromere region due to 1) Close physical proximity of the selected gene/genes to the active centromere, and, 2) limited recombination caused by the centromere itself, and/or structural variation. In these cases, selection for the linked allele would also select for a centromere haplotype, resulting in a sweep of the allele and centromere resulting in the observed domesticated, single haplotype centromeres. The distance of the domestication locus to the centromere impacts the level of observed diversity loss due to linkage drag effects. This was observed with the multi-haplotype CEN5 and the domestication locus Region-D. The domestication locus was ~6Mb from the closest active centromere (CEN5R). The increased physical and genetic distance and lack of major (>1Mb) inversions permitted more recombination between the centromere and the domestication locus and greater centromere diversity.

4.5.3 Potential consequences centromere haplotype formation

Given centromeric haplotype regions in *Zea mays* were often multiple Mb in size, many genes become trapped in the centromere haplotype regions and once genes become linked, they were unlikely to recombine. In single haplotype centromeres at least one of the genes was under strong enough selection that resulted in a sweep in domesticated maize. Though it was possible that slightly deleterious alleles may have segregated with the domesticated centromere due to linkage drag. These alleles would be protected from purifying selection due to linkage with the domesticated gene. However further investigation is needed to determine the nature of specific centromeric genes.

4.5.4 Future directions

Population bottlenecks and selection during domestication have been reported to lead to an increase in the abundance of deleterious alleles in maize [162]. When a gene or genes were under selection in the centromere haplotype region, it can sweep all the genes in the region fixing between 6.2 to 12.3 Mb of sequence. While the centromere region was under selection due to relatively few genes, many additional genes and pseudogenes can be swept along with the centromere due to the inhibition of meiotic recombination in the region. It was likely that some of these genes were slightly detrimental to the organism. An understanding of the genes within the centromere regions would be valuable for our understanding of how the centromeres of maize evolved and could potentially uncover targets to further improve maize. Bioinformatic analysis of the genes in the centromere haplotypes can reveal basic properties about annotated gene models and centromere structure, but further experimentation was required to understand the physiological and phenotypic impacts of centromere domestication on the organism. Specifically, CRISPR knockouts/knockdown lines could be generated for each of the gene models in the active centromeres 1, 4, and 6 to elucidate phenotypic effects (if any). This could provide additional data that would help pinpoint the specific centromere linked genes responsible for the domestication of centromeres, however these experiments were beyond the scope of this works.

CHAPTER 5. The expansion and evolution of the CR2 subfamily in domesticated maize

5.1 Abstract

Zea mays was domesticated ~10 thousand years ago. The genetic bottlenecks and inbreeding resulting from domestication disrupted the ancestral centromeres and resulted in the formation of *de novo* neocentromeres. These neocentromeres were subsequently colonized by a group of centromere targeting retrotransposons (CR1 and CR2). Both CR1 and CR2 have been actively transposing in maize

centromeres following domestication. CR2 specifically underwent a post domestication burst of activity. Furthermore, investigation of teosinte, landraces, and improved maize lines suggest maize lost CR1 and gained CR2 during domestication. Unlike CR1, CR2 did not expand as the result of a preferred recombination. However, there is extensive evidence of recombination among the seven parental CR2 subtypes. SNP analysis of the CR2 lineage suggests modern CR2s and *Zea luzurians* are both enriched for a set of six SNPs suggesting potential horizontal transfer. However, no evidence of such a transfer remains in the B73 genome. Investigation of unmethylated regions in CR1 and CR2 are correlated with active centromeres, though unmethylated CR elements are not shared across even closely related genomes. Unmethylated CR2 elements are also observed proximal to the recently shifted CEN5L' in CML277 suggesting CR elements can remain unmethylated following centromere movement. These data suggest that CR2 may have expanded due to an acquisition of a foreign CR2 acquired during a horizonal transfer, and/or CR2 expanded because of epigenetic changes in CR2s in the ancestral centromere following neocentromere formation.

5.2 Introduction

Modern *Zea mays* ssp. *mays* was domesticated from the wild teosinte *Zea mays* ssp. *parviglumis* between 7.5 and 10 thousand years ago (ka) in the Balsas River valley region of Mexico [2-4]. The domestication of maize resulted in profound morphological and genetic changes including a major loss in diversity over the centromeres [24]. Studies of the centromeres of modern maize inbreds uncovered genetic bottlenecks and inbreeding during domestication that led to the loss of the maize centromeric tandem repeat centC. The loss of centC destabilized the ancestral centromeres leading to *de-novo* neocentromere formation. Following the formation of the neocentromeres, they were colonized by Ty3/*Gypsy* centromeric retrotransposons (CR) [24]. Six subfamilies of CR elements have been described (CR1 – CR6) that are widely distributed in the *Poaceae* family (grasses), though only CR1, CR2 and CR3

are shown to target active centromeres [111]. Detailed studies of CR1 reveal two parental elements of CR1 (CR1-A and CR1-B), along with five major recombinants (CR1-R1 through CR1-R5) with many unique recombinants [82]. The recombinants are reported to have formed via intrastrand homologous recombination and all major CR1 recombinants shared a region of the RNAseH domain from the CR1-A parental element. It is postulated that this recombination is correlated to the expansion of the recombinant CR1 subtypes.

Many factors have been attributed to the expansion or activation of retrotransposon families including heat stress, polyploidization, and events that alter the epigenetic state (methylation) of the retrotransposons [163-165]. However, retrotransposons can be deleterious to the host if copy number increases unchecked as insertions into promoters or genes can alter expression and thus alter plant development. The study of these phenotypic differences are what led to the initial discovery of TEs by Barbara McClintock [166]. Due to the potentially damaging effects of unchecked TE expansion, TEs are frequently silenced to suppress their transcription and therefore, mobility and amplification. Numerous studies have shown siRNAs actively target TEs [167-169]. The siRNAs generated from the elements can also modify chromatin to further suppress TE transcription through siRNA mediated DNA methylation [170]. There is currently little information regarding the regulation/methylation of CR elements in the literature, though previous studies of CRR2 (the rice CR2 ortholog) reported transcription of CRR2 is low and derived from relatively few CRR2 elements [124]. This limited transcription suggests CR2 elements are potentially being silenced by these same mechanisms.

While considerable research has been conducted on the CR1 lineage little is known about the CR2 lineage. Efforts to evaluate the evolution of CR2 were constrained by the limited numbers of available full-length elements. The release of high-quality physical maps for the NAM founder lines with many completely assembled centromeres and supplemented by epigenetic data, allows for further studies of the CR2 lineage and its expansion in a diverse array of maize inbreds [171]. Here we examine the lineage

of the CR2 subfamily as it relates to CR1 and describe the amplification, evolution, and regulation of the CR2 haplotypes and propose a mechanism for CR2 expansion during domestication.

5.3 Methods

5.3.1 CR2 extraction

CR2 elements were identified with blastn using a post domestication CR2 representative (97243919.97236347.chr2.B73.GAAGA.GAAGA.R). Blast results were parsed and CR2 elements extracted from the genome using a custom Perl script. The resulting elements were then verified visually with Junction Viewer [108].

5.3.2 CR2 phylogenetic analysis, haplotyping

A full-length alignment of all extracted CR2s was constructed using Mafft [172]. The alignment was manually edited in MegaX [152]. MegaX was used to generate phylogenetic trees (K2p distance, pairwise deletion). Haplotypes were defined using the B73 CR2 tree. To haplotype the CR2s in the other NAM lines B73 CR2s were combined with one NAM line at a time and trees were generated from the alignments of all CR2s in both lines. The haplotyped B73 elements were used as a guide to haplotype the non-B73 elements.

5.3.3 CR2 recombination and consensus sequence generation

Using the list of all CR2s we matched CRs in the same orientation on the same chromosome with the same TSDs to determine a list of unique CR2 insertions and to evaluate phylogenetic relationships among CR2s in the NAM lines. The dataset was manually checked to find any CR2 that was not properly matched due to an inversion. CR2s inserted directly into CentC repeat on the same chromosome with the same TSD in opposite orientation were not matched as we could not be sure if they were truly

unique insertions. Majority consensus sequences were generated from alignments of all unique insertions in the NAM lines from a clade using Geneious. These consensus sequences were aligned using muscle in MegaX and all variable sites were extracted to generate a set of 205 SNPs used to define recombination breakpoints among the clades. A subset of the 205 SNPs shared between at least two clade consensus sequences (91 SNPs) was used for the enrichment analysis.

5.3.4 CR1 and CR2 fractions for maize lineages

Short read data was combined for 80 datasets (26 NAM and two additional B73 data sets W22 and Mo17, 17 TIL, 23 BKN, and 10 *Zea* datasets, obtained from PRJNA384363, PRJNA389800, [173], [174], [175]) was mapped to the HAP_III consensus CR2 and the CR1_R5 consensus using Bowtie2 [176]. CR1 and CR2 fractions were calculated as the number of reads mapped to the representative out of the total reads in the data set.

5.3.5 SNP identification via short read mapping and SNP enrichment

Short read data for 17 TIL lines, 23 landraces and 28 domesticated lines were mapped to a reference B73 v5 "G" element (97243919.97236347.chr2.B73.GAAGA.GAAGA.R) with Bowtie2 (--local --no-unal). The resulting BAM files were converted to VCF files with BCFtools mpileup and call (call -Ov -m -A -V indels) [41]. The VCF files were then parsed with a custom Perl script that calculates the percentage of reads matching the reference for every position. These percentages were averaged at each position across the element among, domesticated, landraces, teosinte and *Zea* datasets.

5.3.6 UMR analysis

UMR data was obtained from Cyverse (/iplant/home/shared/NAM/NAM_genome_and_annotation_Jan2021_release). Additional data on how

UMRs were defined can be found in the supplemental text of Hufford et al. 2021 [1]. The UMR bed files were compared against CR2 coordinates using a custom Perl script that reports any UMR that occurs between the terminal coordinates of a CR2. The Fasta sequence for all CR2 associated UMRs was then extracted from the appropriate NAM line physical map and aligned to the seven CR2 parental consensus sequences using blastn. Partial Blast hits from UMRs split over CR2 and other sequences were retained. Per-nucleotide coverage was then calculated across the CR2 consensus sequences. CenH3 (SRR3018834) and H2A.Z were mapped to the parental CR2s using Bowtie (--best --strata -S -X 2000 -p 10 -v 2 -m 2). The resulting bam files were converted to vcf files with bcftools from which raw coverage was extracted.

5.3.7 Identification of the oldest element in the CR2 Clades

Initially the unique list of CR2 elements was sorted by haplotype and date. Only elements with estimated LTR insertion dates >10ka were examined. All elements with the same TSD as the CR2 in question were checked with JV images to verify that the element was not truly shared. Elements in CentC were excluded as the high likelihood of repeated TSDs obscures the analysis of whether an element is in fact shared. If elements found in a single line were not in CentC, the archaeological lines were searched to see if the element was present. If it was found in at least 1 archeological line the CR2 was retained for further analysis. In cases of two or more lines sharing a CR2 element, the lines containing the insertion were compared to HapMap3 centromere trees to ensure continuity. All individual insertions from each line containing the CR2 were extracted and aligned to verify the estimated LTR insertion date and provide a higher resolution estimation of insertion date. The oldest date was defined as the element with the highest confirmed insertion date estimate.

5.3.8 Verification of shared elements in the unique list of CR2s

Investigations of CR2 elements in chr1 (137050019_137057586_chr1_B73_GTTAG and orthologs in NAM) indicated that inversions and mis-assembly can cause shared elements to be called as unique. To evaluate the pervasiveness of this in the list of unique CR2 elements, all elements sharing a TSD and haplotype were grouped. To verify the context of n elements insertion, a 1kb region flanking the 5' and 3' of each CR2 in each group was extracted. All 5' flanking sequences were aligned with blastn to all other 5' sequences in the group and the same was done for the 3' sequences. Any HSPs with a length >900 on both the 5' and 3' ends were considered to share the same context and were combined as a single element. We characterized the context of all CR2 insertions sharing a TSD and haplotype to determine how many in the unique list should be combined. Of the 3997 unique insertions belonging to one of the 18 clades only 23 needed to be combined. Of these 23, 19 were due to the inversion on chr1, 3 were due to the mis-assembly on chr6 and one CR2, far outside the centromere may be due to either a mis-assembly or a translocation between chr10 and chr9.

5.4 Results

5.4.1 Inventory of the CR family members in B73

A survey of complete and interrupted (up to 42 kb insertion permitted) CR elements in B73_v4CEN (**Chapter 2**) revealed that CR2 elements (410 total) had been the most active of the six CR subfamilies and their non-autonomous counterparts since domestication (**Figure 5.1**). The CR1 subfamily contains the largest number of elements (614 total), some of which were inserted post-domestication, though the bulk were inserted pre-domestication. The third CR subfamily that targets centromeres, CR3, is represented by only 13 full-length elements and 30 non-autonomous recombinant CR3 derivatives (CentA) of which only one was inserted post domestication. In contrast, no post-domestication insertions were detected for the three CR subfamilies that do not target centromeres or their non-autonomous derivatives (CR4, CR5, and CR6).

Non-autonomous derivatives were detected for CR3 (CentA), CR4 (noaCR4) and CR5 (noaCR5) and except for CentA, the autonomous parents outnumbered their non-autonomous derivatives. Two different deletion variants of CR2 were detected missing 854 nt over the integrase core domain or 533 nt over the reverse transcriptase core domain that occurred in small numbers as independent integrations (*SI Appendix Table ST19*). The older CR2 Integrase deletion variant had not transposed for 75 thousand years (ka) while the newer reverse transcriptase deletion variant formed post-domestication (0 ka) and had transposed twice. Retrotransposon genomes are known to be co-packaged in virus-like particles [177]. As the CR2 deletion variants were missing large sections of core functional domains, we postulate that the truncated elements were co-packaged with and integrated by full-length/fully functional elements. It iwa feasible that these variants may represent evolving non-autonomous derivatives.

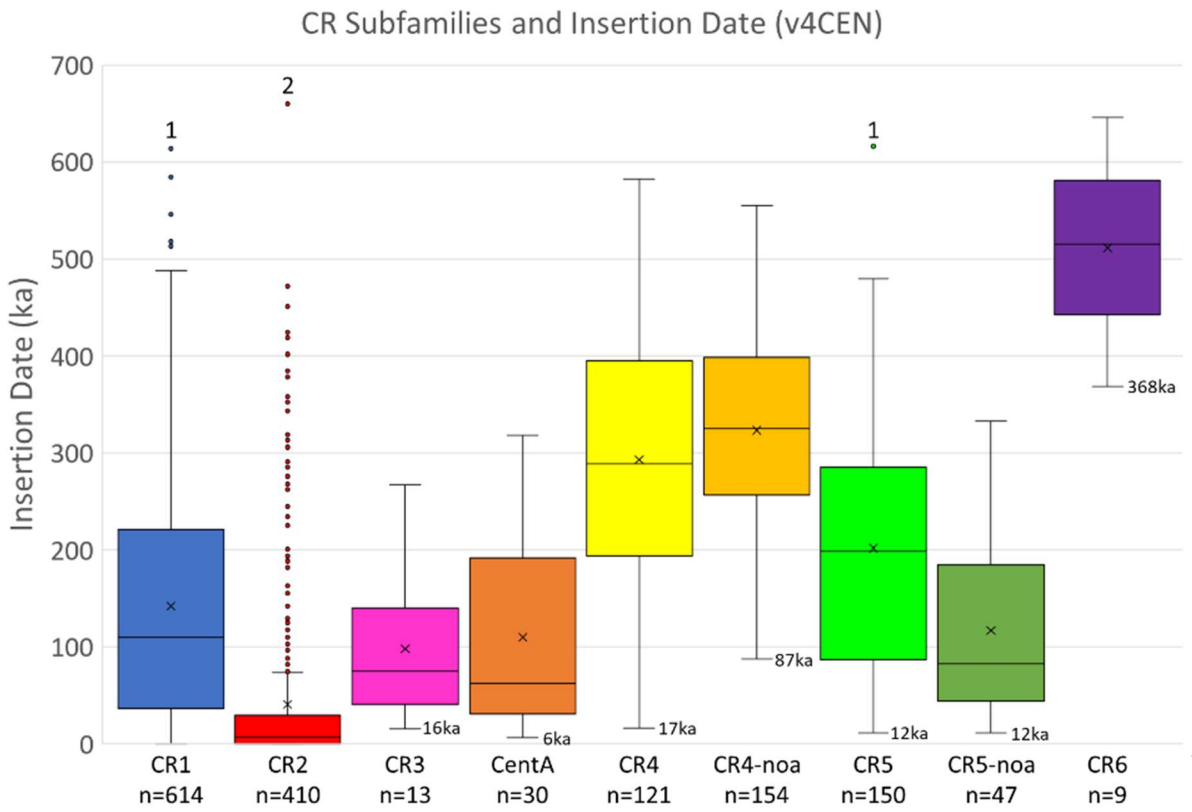


Figure 5.1. Insertion time distribution for CR elements of the B73 genome. Boxes represent the second and third quartile bisected by the median insertion date. X marks the mean insertion date for each subtype. Whiskers represent the local maxima and minima for each type (5^{th} and 95^{th} percentile). Outliers marked as points outside the whiskers. The CR4 group contains members of both the CR4-A and CR4-A-DV (LTR deletion variant of CR4-A), CR4-B and CR4-R1 subtypes both described in Sharma *et al*. 2008. Three CR3-Bs were discovered and not shown (A CR3 PP and CentA like LTRs and UTR). The CentA group also included 22 CentA and 9 CentA variants that had extended LTRs. The CR1 block includes both CR1-A and B parental types as well as recombinants R1-R5. The total number of elements with matching TSDs included in each subtype shown beneath the subtype name. CR6 was the oldest and most rare subfamily and was not been active for ~368ka. The great age and small number of CR6 elements suggested they are either too fragmented to detect or have been lost from the genome. Outliers were numbered and characterized: **1,** Date and CR insertion were confirmed with B73 RefGen_v2. **2,** Absent in V2, but date was confirmed in B73 RefGen_v5 and other NAM lines (B73, Oh43, Oh7B, Tx303, Tzi8 and 660ka +/- 173ka). Data were obtained using B73 RefGen_v4CEN (and confirmed using V2 and V5).

5.4.2 The CR2 subfamily expanded in a post-domestication burst

Analysis of total CR1 and CR2 content in short read datasets representing 74 maize cultivars show an inverse relationship between CR1 and CR2 when comparing modern maize inbreds to teosinte inbred lines (TIL) which was consistent with previous analysis (**Figure 5.2**) [14, 24]. TIL lines displayed a high CR1 content and low CR2 content. Notably, TILs derived from *Zea mays* ssp. *mexicana* displayed a CR1 fraction higher than their *Zea mays* ssp. *parviglumis* counterparts. Modern maize inbreds conversely, show high CR2 and low CR1 fractions and domesticated landraces (BKN) fall between the TILs and modern inbreds. Investigation of two *Zea mays* ssp. *huehuetenangensis, Zea diploperennis* and *Zea luxurians* lines indicate variable CR1 fractions, and except for *Zea diploperennis,* CR2 fractions were roughly consistent with teosinte. *Zea diploperennis* contained the smallest fraction of CR1 and CR2 reads of any line. The trend suggested that an expansion of CR2 elements occurred during domestication and was accompanied by a loss of CR1.

To prove that CR2 underwent a post-domestication expansion, cumulative counts of CR1 and CR2 elements among the NAM lines were analyzed. It was found that for much of the past 600ka CR2 and CR1 insertion rates were relatively constant, though CR2 showed a sharp increase in insertion rate within the last ~15ka that roughly corresponds with domestication given our resolution (each CR2 LTR

SNP represents ~7.5ka, **Figure 5.3a**). This showed that CR2 expanded during domestication more rapidly than historically. Overall, the CR1 subfamily did not show the same increase in insertion rate during domestication. However, when individual CR1 subtypes were investigated, we observed variable insertion rates between subtypes (**Figure 5.3b**). Four subtypes (CR1-A, B, R2 and R3) did not expand through domestication. Though other subtypes such as CR1-R1, R4, and R5 have been expanding since their formation. It is notable that the rate of insertions in CR1-R1 and CR1-R4 appear to be decreasing while the insertion rate of CR1-R5 has remained consistent through domestication. It is important to note that none of the CR1 subtypes are expanding at the rate of CR2. This data suggested that CR2 expanded in a burst event during domestication. Further, certain CR1 subtypes did expand through domestication though overall there was no change in the overall insertion rate of the CR1 subfamily. This was remarkable given that both CR1 and CR2 were centromere targeting and both had subtypes that were active during domestication.

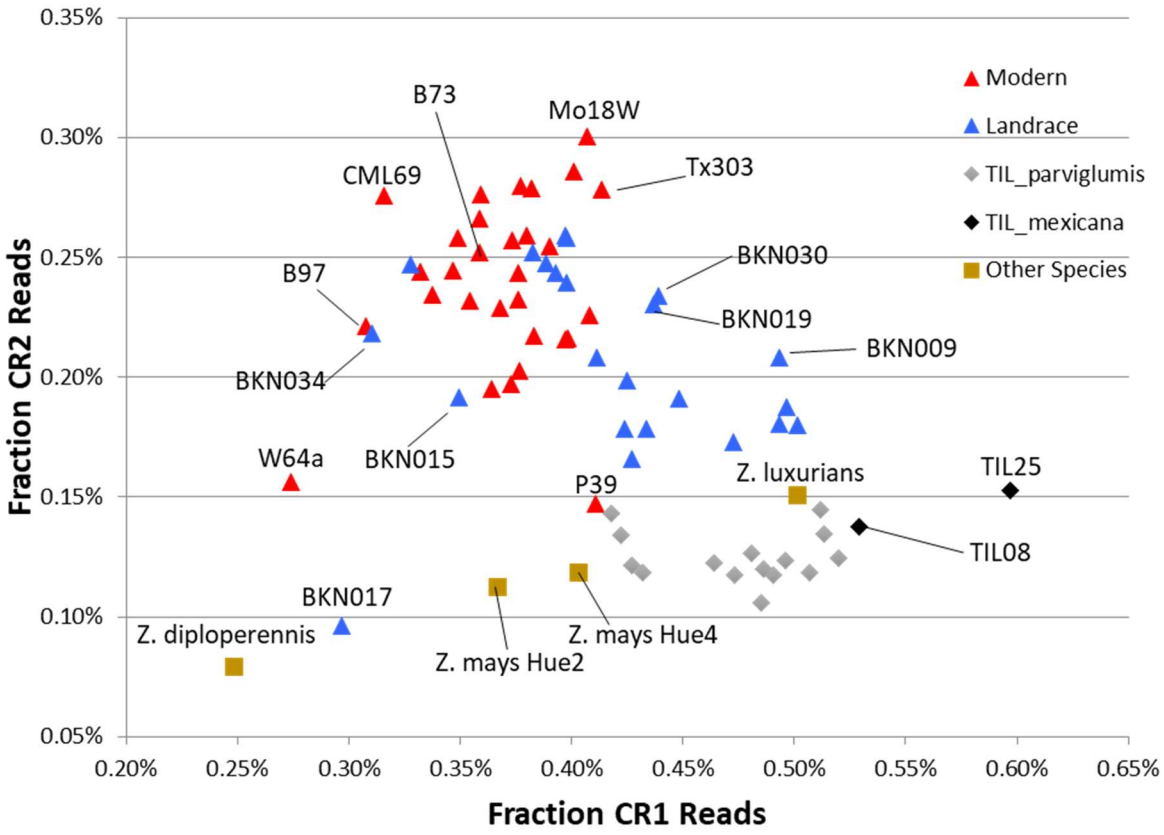


Figure 5.2. An increase in CR2 content in the genome of domesticated maize associated with a reduction in CR1. The CR2 content of different *Zea* genomes was determined from the proportion of short read sequences with homology to a reference modern CR2 element and CR1-CR6 using competitive Bowtie. Teosinte inbred genomes (*Zea mays* ssp*. parviglumis* or *Zea mays* ssp*. mexicana* (grey and black diamonds, respectively), represented the genetically diverse ancestors of domesticated maize, contain between 0.42%-0.6% CR1 and 0.11%-0.15% CR2. Many landraces of maize (blue triangles) contain a similar proportion of CR1 but higher amounts of CR2. Some landraces have CR1/CR2 contents like modern maize (red triangles), with CR1 content of 0.27%-0.41% and CR2 contents of 0.15%-0.30%. Genomes of *Zea mays* ssp*. huehuetenangensis* and two additional *Zea* species contain CR2 levels like teosinte but highly variable CR1 amounts. The outlier P39 was likely grown in isolation resulting in a lower CR2 fraction. *Zea luxurians* contains one of the highest CR2 fractions among non-domesticated maize.

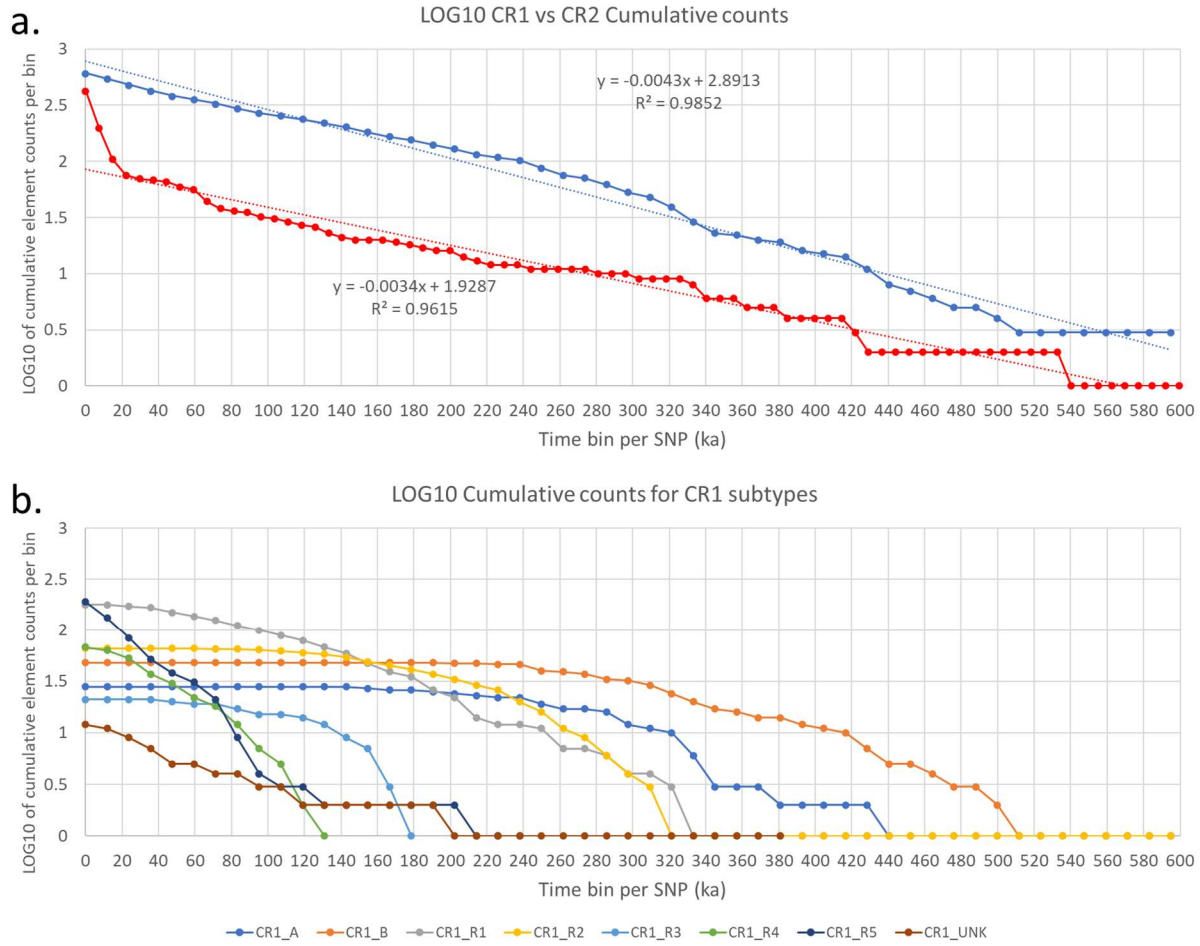


Figure 5.3. Log10 transformation of cumulative insertions in B73 over the last 600ka. Each point across the x axis corresponds to the addition of 1 SNP in the LTR in either CR1 or CR2. The shorter length of the CR1 LTR leads to decreased temporal resolution seen as a larger range between time points. The y axis is the log10 transformation of cumulative element counts through each successive timepoint. **a.** The expansion rate for CR1 and CR2 overall are roughly linear along most of their length but a large increase in CR2 insertion rate can be observed starting ~15 ka which continues through to 0ka corresponding to domestication. **b.** The insertion rates among the CR1 subtypes are variable, though CR1-R5 has been steadily expanding for the past ~100 ka. The insertion rate of CR1-R4, which is also active post domestication, may be leveling off suggesting a decrease in insertion rate.

5.4.3 A preferred recombination was not responsible for CR2 expansion

Previous studies were unable to explore the diversity and evolution of the CR2 subfamily due to the small number of full-length CR2 elements available. This problem was compounded by the relatively close genetic distance between CR2s (Max distance 105 ka between CR2 subtypes “A” and “G”). In

contrasted with the genetic distance between CR1 “A” and “B” parents of 873 ka indicating that CR1 was far more diverged (*SI Appendix Table ST20*). The release of the higher quality NAM physical maps with many completely assembled centromeres has allowed for a more comprehensive inventory of the CR2 subfamily across multiple lineages, permitting detailed study (4,134 unique CR2 insertions among the NAM lines). There was no evidence of sequence specific insertion (i.e., overrepresented TSDs) among the extracted CR2s. However, phylogenetic and SNP analysis among these CR2s revealed evidence of extensive recombination. The CR2 subfamily is represented by seven parental clades (“A” – “G”) and 11 recombinant clades (**Figure 5.4**). Further analysis of the parental clades indicated “C”,” D”,”E”, “F”, and “G” clades contain subclades which correspond to sub-recombinants. Sub-recombinants were more prevalent in the clades with high post-domestication activity (“F” and “G”) (*SI Appendix Figure S40*) indicated a correlation between the period of activity (domestication) and CR2 recombination frequency.

The parental CR2 clades do not share the same periods of activity. The ancestral clades (“A” and “B”) were no longer active having last being inserted 183 and 29 ka respectively, though they could still potentially be transposed as the “A” and “B” clades contain 2 and 28 uninterrupted insertions, respectively. The pre-domestication “C”, “D” and “E” have transposed since domestication though the distribution of their insertion dates indicated they have decreased in activity. The “F” clade represents a unique group that has been present in the genome for 101 ka at low frequency, though the bulk of its expansion occurred post-domestication, resulting in a skewed insertion date distribution similar to the “G” clade. The “G” clade was the most recently formed parental clade (~14ka) and has primarily expanded post-domestication. It was also the most prolific of any CR2 clade containing 987 elements. All 11 recombinant clades were at least partially derived from “F” and/or “G” elements either as direct parents of the recombinant clade, or parents of the parents. Analysis of recombinations among parental,

and recombinant CR2 clades showed there was no recombination or SNP shared by all the clades that expanded post-domestication. This contrasted with CR1 recombinants R1-R5 reported to contain a shared recombination over the RNAseH domain inherited from the CR1-A parental element [82]. This indicated that a specific recombination was not responsible for the expansion of the CR2 subfamily.

Figure 5.4. Date distributions and recombination among the clades of the CR2 subfamily. Clades were initially determined within B73 based on their position in a phylogenetic tree generated from a full-length alignment of all CR2s. Insertions in the NAM lines were haplotyped using phylogenetic trees and the haplotyped B73 elements as a reference. Date distributions shown in the violin plots. Violin plots generated from unique elements with inversion duplicates and any elements with LTR insertion dates greater than that of the oldest confirmed insertion removed. Plots ordered from oldest to youngest (not

including rare elements (black)) based on the oldest insertion date estimate for each clade in B73 (full-length alignments of the shared oldest element). While clade Gx(FxE) was derived from FxE the oldest element in FxE was younger than clade Gx(FxE) by 900 years. As such FxE existed before Gx(FxE) but the ancestral element was either not detected or deleted. The average LTR insertion date of the oldest confirmed element in each clade was marked as a red line. The estimated divergence date between full-length copies from NAM lines containing the oldest element marked as a green line. Schematics to the right of each violin plot show recombinant elements to scale. The schematic above the ancestral (A/B) clades represents a typical CR2 including major regions and domains. Recombination was prevalent within and between clades that confounds relationships. The number of sub-recombinants in the ancient and ancestral clade groups shown as a fraction of the total (N) beneath the clade name. The number of uninterrupted elements in each clade marked in blue as a fraction of the total (N).

5.4.4 SNPs in recently active CR2 clades in domesticated maize were enriched in *Zea luxurians*

Analysis of CR2 clade consensus sequences generated from unique insertions resulted in 205 variable sites of which 91 were present in at least two consensus sequences. Investigation of the frequency of post-domestication alleles ("G" clade enrichment) at the 91 positions between modern inbreds, TILs, and domesticated landraces, indicated 44 SNPs were highly enriched (>=95th percentile for TIL enrichment) in domesticated maize (modern inbreds and landraces) relative to teosinte (TILs) (*SI Appendix Figure S41*) (**Figure 5.5a**). Of the 44 SNPs only 10 were non-synonymous substitutions. The increase in these allele frequencies through domestication was remarkable, though it was unknown whether the shift in frequencies was the result of adaptive evolution, or genetic drift.

Inspection of the TILs revealed enrichment of post-domestication alleles in TIL03 and TIL11 (**Figure 5.5b**). Shared junction analysis revealed 4 and 2 CR2 junction pairs (each junction pair covers 5' and 3' junctions from a single element) in TIL03 and TIL11 respectively shared with modern inbreds dated between 0 and 2 ka. Further, phylogenetic analysis of HapMap3.2.1 SNP data of CEN3 indicate TIL03, and TIL11 are within a clade that diverged post-domestication (7.8 ka) rather than basal group which would be expected if these TILs were the donor for the domesticated CEN3. The post domestication dates of CR2s shared with the TILs, and phylogenetic placement of the TILs among domesticated maize,

suggested enrichment was the result of a domesticated haplotype-1 (B73-like) CEN3 escaping back into teosinte.

Evaluation of SNP enrichment for the *Zea* lines indicated a set of six SNPs highly enriched in *Zea luxurians* and to a lesser extent, *Zea mays* ssp*. huehuetenangensis* relative to the TILs (**Figure 5.5c**). The SNPs were split into two groups with two SNPs upstream (GT) and four downstream (ACTA) of the core GAG domain. Investigation of these six SNPs revealed two non-synonymous substitutions (one in the upstream group and one downstream group). These SNPs were found to be specific to the "F" and "G" parents and their recombinants. Comparisons of the *Zea luxurians* genomic consensus CR2 to the clades in domesticated maize over the 91 SNPs indicated CR2 was closely related to the "F" subtype. This suggested that an "F" clade relative may be the most prevalent CR2 subtype in *Zea luxurians*. The "F" clade has persisted for 101 ka though it only expanded following domestication in maize. Given the divergence between *Zea luxurians* and *Zea mays* ssp. *parviglumis* (the ancestor of domesticated maize) was estimated at ~140 ka [178] the prevalence of these SNPs in *Zea luxurians* suggested a potential horizontal transfer (HT) of the "F" subtype CR2 between the two species. While canonically recombination frequency was low in and near the centromere, there was evidence of a distantly related (~500ka) introgression near the active centromere of chromosome 9 (*SI Appendix Text, SX8, SX9, Figures S23-S25*) that suggested that such a transfer could occur. Analysis of shared CR2 junction pairs between *Zea luxurians* and the NAM lines indicated putative shared CR2s on CENs 5, 6, 8 and 9 though analysis of SNP density using *Zea luxurians* as a reference suggested these centromeres did not contain introgressions derived from *Zea luxurians* (*SI Appendix Figure S42, Table ST21*). This data suggested that if there was an introgression, it may have been transient, existing just long enough for CR2s to escape into the genome.

Figure 5.5. Enrichment of 91 SNPs to the "G" consensus CR2 in domesticated maize teosinte and Zea lines. Each position on the x axis represents one of 91 SNPs. The enrichment relative to teosinte was calculated for each line as the ratio of the allele percentage of reads corresponding to the post-domestication "G" allele for the line and the average of the TIL allele percentages for the SNP position. Groups of lines we calculated as the ratio between the mean for the groups (NAM, Landrace, or TIL) over the position. The black line across each plot represents the 95th percentile for teosinte enrichment data. **a.** This panel shows illustrates clear enrichment for the 44 SNPs (>95 percentile of TIL enrichment) among both NAM and Landraces (blue and orange respectively) compared to individual TILs indicating potential selection for these alleles in domesticated maize. **b.** TILs 03, and 11 are slightly enriched for post domestication SNPs. Analysis of HapMap3.2.1 SNP trees over the centromeres suggest that this is due to each of these TILs sharing a post-domestication centromere with modern maize. TIL25 is shown as a control and contains no escaped domesticated centromeres and TIL14 was a control showing a centromere donor (i.e. The basal TIL for CEN2). **c.** *Zea luxurians* shows clear enrichment for various "G" SNPs at a magnitude matching the NAM/Landraces (red). Analysis of the 91 SNPs extracted from CR2 haplotype consensus sequences suggested *Zea luxurians* was enriched for the "F" clade specifically with only the blue ("G") SNPs being enriched. A set of 6 SNPs (red boxes) were specific to F/G clades and their

recombinants. *Zea diploperennis* conversely was depleted for most SNPs suggesting it contained either only ancestral elements, or an entirely different *diploperennis* specific CR2 clade.

5.5.5 Unmethylated regions in CR1 and CR2 are correlated with the active centromere

The methylation status of CR2 elements among the NAM lines was evaluated using annotated sets of unmethylated regions (UMRs) presented in Hufford *et al.* 2021 [1] and account for all three contexts of methylation found in plants (CG, CHG, and CHH). Comparisons of UMRs, H2A.Z, and cenH3 distribution across CR element subtypes that have been active post-domestication (CR1-R5, and CR2-G) highlight the key epigenetic differences between the two CR subfamilies (**Figure 5.6**). Analysis of cenH3 distribution indicated both elements were covered almost entirely by cenH3 and expected because the centromere targeting nature of these elements. We had observed that 72% of CR2-G elements and 47% of CR1-R5 elements were found in an active centromere and account for the observed cenH3 coverage variability in CR1 vs CR2. It was notable that currently active CR lineages had higher representation in the active centromere (*SI Appendix Table ST22*). This suggested transposition favors recently active elements. Distribution of the transcription associated histone H2A.Z was relatively consistent between the two subfamilies with peaks localizing to the LTRs and between the major functional domains in the PP (*SI Appendix Text SX13*). The remarkably similar H2A.Z distribution between CR1 and CR2 suggested that it was likely not correlated with the observed variable expansion between the two subfamilies. The distribution of UMRs conversely, differs considerably between CR1 and CR2. In CR1 UMRs localized to the UTR and PP regions of the element, though CR2 UMRs localized to the LTR and UTR (*SI Appendix Figures S38, S39*). Remarkably, The CR2 UMR peak in the LTR is ~100nt upstream of the TATA box suggesting it may affect CR2 regulation. Analysis of the parental subtypes in CR2 and CR1 indicated the distribution of UMRs was consistent across all subtypes. Furthermore, analysis of UMRs among the individual NAM line insertions show all CR2 clades had at least one representative containing a UMR, suggested UMRs were not subtype specific (*SI Appendix Figures S38, S39*). This data suggested variability

in methylation levels (UMRs) were more likely to be affecting CR1/CR2 retrotransposition than cenH3 or H2A.Z.

Investigation of the distribution of UMRs in B73 v5 and across the various centromere lineages in CEN5 demonstrate that CR1/CR2 elements with UMRs were highly correlated with the active centromere (*SI Appendix Figures S36, S37*). The lack of subfamily/subtype specificity among UMRs, and the tight correlation of CR1/CR2 elements with UMRs to the active centromeres suggest CR UMRs were a property of active centromeres rather than a property specific to retrotransposon subfamilies. Evaluation of CEN5 indicated representatives of the brown lineage (CML322 and CML277) contained a dense cluster of recently inserted CR2s not covered by cenH3 suggesting a post-domestication neocentromere movement from the CEN5L locus to CEN5L' (**Figure 5.7**). Remarkably, two CR2s retained their UMRs upstream of the active centromere in CML277 suggesting UMRs may persist for a time after centromere repositioning. The CR2s in the cluster were not shared with other CEN5L lineages (Blue/Green/Il14H) [4] indicating the brown lineage was an independent neocentromere formation. Furthermore, the reported divergence estimates for the Blue/Orange/Brown CEN5 lineages predated the CR2 insertions shared by both brown lines indicating the CR2s followed the neocentromere rather than seed it. Only a subset of CR2s within the active centromere contained UMRs and comparisons of CR2 UMRs between CEN5 Brown lineages indicated little UMR conservation, even between closely related (~1.2 ka) haplotypes thus obfuscated phylogenetic comparisons.

While CR elements were repetitive, the LTR UMR signal found in CR2 was likely not an artifact of mapping for two reasons: 1st the mapping algorithm (BSseeker2) does not retain repetitive sequence using the reported mapping parameters, thus the "uniqueness" was dependent on the methylation patterns of the read and methylation context of the genome that allowed repeat derived reads to map uniquely. 2nd, the LTR UMRs were within the reported 1kb mapping distance of the unique CR2 insertion junctions allowing reads to be mapped to specific CR2s. Due to the lack of consistently unique sequence

within 1kb of the UTR UMR peaks in CR1/CR2 or the PP peak in CR1 we cannot be entirely certain that the UMRs located in the UTR/PP were not artifactual.

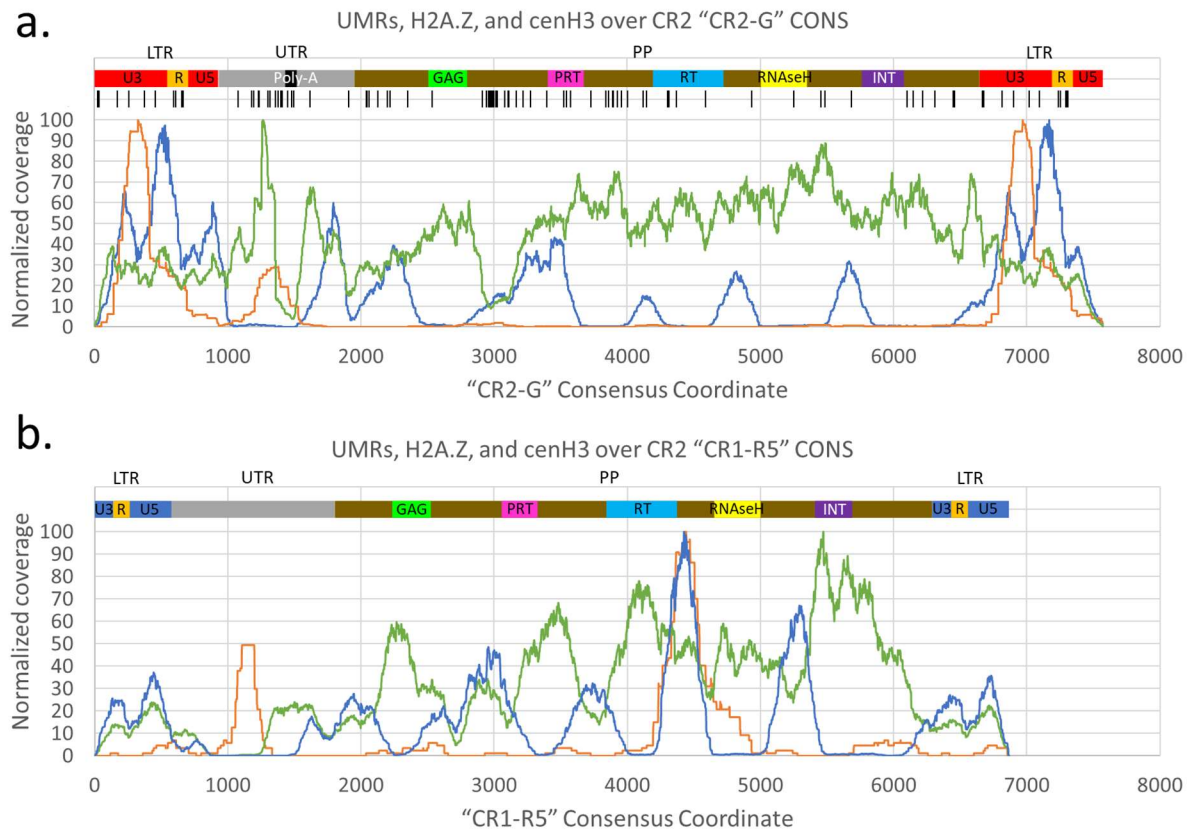


Figure 5.6. UMR, H2A.Z and cenH3 coverage for CR2-G (a.) and CR1-R5 (b.). A schematic of major functional domains for CR1 and CR2 were shown above the respective charts. The 91 CR2 variable sites marked with black lines. UMRs found within the terminal coordinates of CR2s across the NAM lines were extracted from their respective genomes and blasted to a consensus post-domestication CR, either "G" for CR2 or "R5" for CR1. The number of UMRs aligned to each position was then counted and normalized to 100. UMR coverage (orange) displays three peaks one in each LTR and one in the UTR in CR2 and two peaks in CR1 (UTR and PP). H2A.Z and cenH3 coverage (B73) were obtained via Bowtie mapping to the "G" and "R5" consensus sequences and normalized to 100. H2A.Z peaks (blue) found in the LTRs and across the polyprotein in both CR1 and CR2. CenH3 variable coverage (green) though most of the element was covered in both CR1 and CR2.

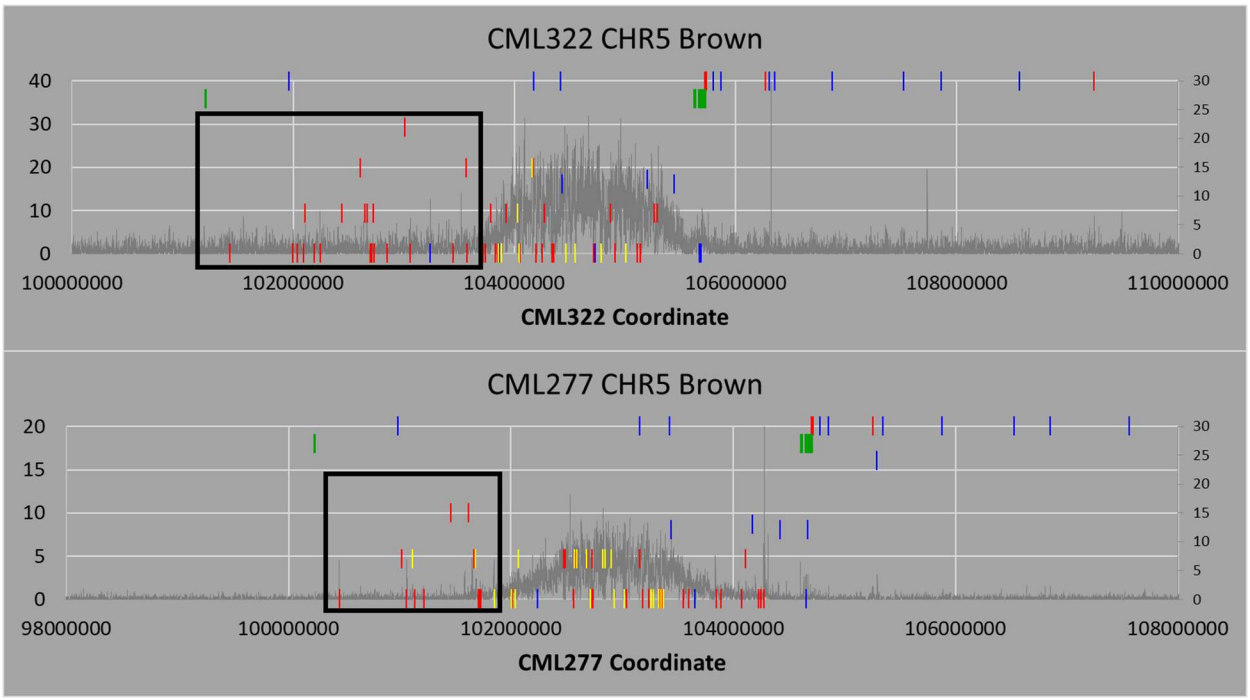


Figure 5.7. Distribution of UMRs across the CEN5 brown lineage (CML277 and CML322). The position and insertion date of CR2 elements and CR1 elements were marked in red and blue respectively on the secondary Y axis (right). Any element with an insertion date >=30ka appears at the top of the chart. Those CR2s containing a UMR marked in yellow. The centC repeat was marked in green and cenH3 raw coverage marked as a grey line on the primary y axis (left). The brown lineage showed a recent movement of the neocentromere downstream characterized by many recent CR2 insertions not within the cenH3 domain (black box). This movement resulted in non-cenH3 covered CR2s with UMRs in CML277. The upstream CR2s were not shared with other CEN5L lineages indicating it was an independent neocentromere formation.

5.5 Discussion

5.5.1 CentC loss during domestication led to the decrease of CR1 in modern inbreds

The relative expansion of CR2 and loss of CR1 through domestication was partially explained (Schneider *et al.* 2016) that the CentC locus represents the ancestral centromere positions. The majority of ancestral CR1 insertions in CEN5 occur in or around the ancestral CEN5M region. Further the *de-novo* neocentromere formations observed in maize is caused by CentC loss in ancestral centromeres. Other investigations into centC loss during inbreeding were inconclusive (*SI Appendix Text SX12, Figure S43, Table ST24*). Assuming centC is indeed lost during domestication, the loss of sequence in the ancestral

centromere would likely cause a decrease in CR1. Furthermore, the colonization of the neocentromeres with CR2 after neocentromere formation would increase the relative abundance of CR2 sequence. The CentC loss and successive colonization resulted in the observed decrease in CR1 and increase in CR2 through domestication. Our data suggested two potential mechanisms by which CR2 may have expanded during domestication. Either CR2 expansion was triggered by acquisition of a horizontally transferred CR2, or CR2 expansion was a consequence of neocentromere formation.

5.5.2 Expansion of CR2 via horizontal transfer of a foreign CR2

The CR2 expansion during maize domesticated was potentially triggered by horizontal transfer (HT) of a foreign CR2 element. How rapidly ancestral centromeres were methylated after neocentromere formation is unknown. If the ancestral CENs were rapidly methylated then the ancestral CR2s therein would be unable to transpose due to silencing. The introduction of a foreign CR2 that could be transpose into the active centromeres following neocentromere formation may have resulted in the observed expansion. Previous studies of HT of transposable elements in plants reported Maize/Sorghum and Maize/Millet transfers resulted in significant increases in copy number of the transferred TE [179]. Additionally, analyses conducted on CR elements across the grasses found evidence that horizontally transferred foreign CRs have recombined with native CR elements and the subsequent recombinants had expanded in copy number [111]. This evidence coupled with the observed six CR2 alleles enriched in domesticated maize and *Zea luxurians* (from an "F" like ancestor) suggested that HT of a foreign CR2 could have been responsible for CR2 expansion. However, we did not observe a large influx of novel allele variants in any of the parental CR2 elements expected if a foreign element was introduced. Additionally, we were unable to locate any centromeric/pericentromeric introgressions shared between *Zea luxurians* and the NAM lines. We were also unable to locate any confirmed shared CR2 between *Zea luxurians* and domesticated maize. This suggested that if there was horizontal transfer, it would have

been transient, existing long enough for the foreign CR2 to escape in domesticated maize. Because of the possibility of a transient HT, and the lack of a full-length CR2 inventory from *Zea luxurians*, we cannot rule out HT as a potential trigger for CR2 expansion.

5.5.3 Expansion CR2 as a consequence of neocentromere formation

Methylation in plants occurs in three contexts CG, CHG, and CHH where H is any nucleotide other than G. Studies in plants have shown that retrotransposons are commonly regulated via siRNA directed DNA methylation [169, 170] and studies in mammalian hybrids have shown that decreases in methylation are associated with retrotransposon activation [164]. This makes the localization of CR associated UMRs to active centromeres and the proximity of the CR2 LTR UMR to the promoter, especially intriguing. Previous studies of the epigenetics of centromeres using 5-methyl-cytosine (5mC) immunostaining have reported that centC and CR elements are hypomethylated in the active centromere but centromeric repeats outside of the active centromere, were hypermethylated [180]. More recent bisulfite sequencing analysis of methylation in the centromeres of maize reported a gradual increase in methylation as one approaches the centromere. Though notably, this analysis did detect lower CHG methylation levels in CR2 (and centC) compared to the unique centromere mean methylation [46]. This suggests that the variability in centromeric repeat methylation levels among the three methylation contexts, is partially due to the previously observed hypomethylation in active maize centromeres (i.e. poor 5mC immunostaining). The presence of UMRs in CR1 and CR2 are evidence of methylation differences between the centromere as a whole and the centromeric repeats (CR and CentC) therein. Centromeres are heterochromatic and usually condensed with cenH3 that limits transcription. Also, the formation of neocentromeres during domestication could shift the cenH3 away from the ancestral centromere and opened up CR2 elements in the ancestral centromere to transcription resulting in the observed post-domestication expansion. There are three lines of evidence supporting this hypothesis.

First, the CR UMRs were present in both CR1 and CR2, and CR UMRs are highly correlated with active centromeres. This suggests low methylation in CRs is a centromeric property, rather than a property specific to CR1 or CR2 elements. Second, our observations suggested that CR2s followed the neocentromere rather than seed it. This indicated CR2s expanded after neocentromere formation as would be expected if neocentromere formation was the trigger for expansion. Third, we detected UMRs outside of a recently moved neocentromere (CEN5 Brown line CML277). It would be necessary for CR2s in the ancestral centromere to retain UMRs like those observed in CML277 after the neocentromere formed. Following neocentromere formation these CR2s would be unmethylated, potentially allowing for increased transcription rates and expansion. Overall, this suggested that the expansion of CR2 may have resulted from epigenetic changes to the ancestral centromere caused by neocentromere formation.

5.5.4 Model of CR2 Expansion

Though CR1 and CR2 transposition rates have been stable for much of their observable history, CR2 underwent a transposition burst during domestication. CR1-R1, R4 and R5 subtypes have increased in copy number during domestication as well, though not to the same extent as CR2. The expansion of CR2 was accompanied by a flurry of recombination resulting in 7 parental and 11 recombinant subtypes though no recombination was shared by all post-domestication subtypes indicating a preferred variant was not responsible for expansion. Analysis of CR2 SNP frequencies showed dramatic shifts between domesticated maize and teosinte though it is unknown if this is due to adaptation or genetic drift. Post-domestication CR2 alleles enriched in *Zea luxurians* suggested potential HT of CR2 elements between the two species though no remaining evidence of a transfer was uncovered. Unmethylated regions in CR elements tightly correlated with the active centromeres show centromeres alter the epigenetic state of the CR elements therein. We concluded that there were two potential mechanisms to explain the

expansion of CR2 during domestication. Either a transient horizontal transfer of a foreign CR2 element coupled with selection against CR2s in the ancestral centromeres triggered CR2 expansion, or changes in the epigenetic state of CR2s in the ancestral centromere followed by neocentromere formation triggered the expansion of the CR2 subfamily. We currently cannot rule out HT of a foreign CR2 though a comparable inventory of full-length CR2 elements from *Zea luxurians* could answer this question.

CHAPTER 6. Summary and Conclusions

6.1 Summary of major findings

This work has advanced our current understanding of centromere evolution by providing insight into three fundamental questions regarding maize centromere evolution.

1) Why do the domesticated maize neocentromeres form where they do?

2) Why are the centromere haplotype regions positioned where they are?

3) What is the evolutionary history of the CR2 retrotransposon lineage and why did it expand?

To answer the first question, epigenetics and transcription were investigated among the NAM founder lines and it was determined that both transcription and H2A.Z were involved in the positioning of neocentromeres. Transcription and H2A.Z were found to be negatively correlated with cenH3 suggesting transcription and H2A.Z are disruptive of cenH3. Formerly pericentromeric regions were found to be ideal for neocentromere loci due to the relatively low transcriptional activity and diffuse H2A.Z density relative to the euchromatin. Neocentromeres were further stabilized by H2A.Z loss following neocentromere formation via deletion and cenH3 invasion of former H2A.Z loci. This resulted in

neocentromeres positioned where they are least disrupted, and least disruptive to the surrounding genome.

To answer the second question, we investigated the extent of centromere-proximal meiotic recombination among the NAM founder lines. We found the closest recombination to an active centromere was just 353kb downstream of the active CEN5M and our data suggest that a recombination this close to an active centromere is rare. Investigation of H2A.Z, transcription and gene annotations suggest that centromeric meiotic recombination is correlated with H2A.Z, and not transcription or the presence of gene models. We also observed extensive structural variation in the centromeres and a domestication locus containing three genes tightly linked to CEN2. This locus and the structural variation in CEN2 were found to define the non-recombinant region in CEN2. We also investigated recombination proximal to CEN5 and found the various CEN5 haplotypes persist in the germplasm due to the large physical distance between CEN5 and the domestication locus region-D. As the selective pressure was not within or directly adjacent to the CEN5 haplotype regions (i.e. Region-D not tightly linked to the centromere), we did not observe a major loss of diversity as observed in CEN2, CEN1 and CEN4. Comparison of the tightly linked domestication locus and CEN2, and the extensive recombination between Region-D and CEN5, suggested that selection and linkage drag were reduced in CEN2 diversity while the diversity at CEN5 was retained.

To answer the third question, we turned to the NAM founder line physical maps and extracted all full-length CR2 elements. We observed that the CR2 family expanded in a post-domestication burst. There was evidence of extensive recombination between CR2 elements though a preferred recombinant was not the trigger for expansion of the CR2 lineage. Investigation of SNPs among the CR2 lineage uncovered a set of 6 SNPs in recently active CR2 clades in domesticated maize are enriched in *Zea luxurians* but we

observed no definitive evidence of horizontal transfer remaining in B73. We also observed a correlation between CR associated unmethylated regions and the active centromere, though these unmethylated regions were not found to be consistent between even closely related NAM lines. This led to two potential explanations for why the CR2 lineage expanded in domesticated maize; either 1) The CR2 lineage expanded via horizontal transfer of a foreign CR2, or, 2) CR2 expanded due to "genomic shock" caused by epigenetic changes associated with neocentromere formation.

6.2 Interplay of epigenetics and transcription in the centromere

Our investigations into the relationship between the epigenetics of the centromere (cenH3) and the histone variant H2A.Z and their relation to transcription strongly suggested that centromere evolution was profoundly impacted by the epigenetic and physical landscape proximal to the centromere. Both transcription and H2A.Z were found to be negatively correlated with cenH3 suggesting both factors were destabilizing to neocentromeres. Neocentromeres in CEN5 were found to have formed in pericentromeric transcriptional voids but transcription itself was insufficient for predicting whether a neocentromere will be established in CEN5L or CEN5R. Investigation of H2A.Z indicated that H2A.Z was lost from neocentromeres by deletion of H2A.Z loci or by invasion of H2A.Z loci by cenH3 following neocentromere formation. This explanation was consistent with previous studies in yeast that showed that H2A.Z loss following neocentromere formation stabilizes neocentromeres [115]. However, neither transcription nor H2A.Z were definitive predictors of whether a neocentromere will form in CEN5L or CEN5R. It was possible that H2A.Z and transcriptional data from the TILs (teosinte) could further elucidate why certain lines formed neocentromeres in either CEN5L or CEN5R.

6.3 The impact of genes and structural variation on centromere evolution

Our investigations have shown that cenH3 and transcription were antagonistic and that transcribed genes in active centromeres were generally devoid of cenH3 (*SI Appendix Figure S2*) suggesting transcription was destabilizing to centromeres. Centromeres were a "rough neighborhood" for genes given the large number of retrotransposon insertions (see **Chapter 5**), frequent double strand break events [61], and epigenetic environment (methylated and condensed by cenH3) [46]. Regardless, we observed that genes persisted and are transcribed within centromeres, suggesting that low levels of transcription can be tolerated. Previous studies in maize CEN10 have shown that euchromatic genes can become linked to centromeres via inversions [24]. However, linkage of genes to centromeres alone is not sufficient to impact centromere evolution. We instead observed that a combination of linkage and selection heavily reduce centromere haplotype diversity. Our exploration of CEN5, CEN2 and CEN1 indicated that the tightness of linkage between the selected gene and the haplotype region correlated with the observed diversity loss. The tighter the linkage on the selected gene the lower the haplotype diversity. We also observed that a combination of selection and structural variation in the form of inversions and indels were the factors that define the non-recombining centromeric haplotype regions in CEN2 and CEN1.

6.4 Functional implications of CR2 elements in neocentromeres

Investigations of retrotransposon content in orthologous regions of rice, maize, and sorghum have shown that all three genomes independently experienced recent retrotransposon activity of variable intensity [181]. These observations are consistent with previous studies that have reported panicoid genome evolution is punctuated by periods of retrotransposon activity often associated with "genomic shock" events (i.e. polyploidization and/or speciation events) [182]. One might speculate epigenetic changes resulting from domestication (i.e. neocentromere formation) in *Zea mays* may have also led to

"genomic shock" resulting in the expansion of the CR2 lineage. However, we also cannot rule out the possibility of horizontal transfer of a foreign CR2 as the trigger for CE2 expansion. Regardless, the expansion of CR2 in domesticated maize may have been beneficial to both the retrotransposon and the host. It is also possible that CR elements help to stabilize neocentromeres that jumped from their ancestral position (CEN5), which at the time of their formation contained no tandem repeat (CentC). It may be that CR elements (CR2 in the case of maize), assist in neocentromere stabilization by two mechanisms.

1) CR2 insertions physically enlarge the centromeric DNA and with enough insertions physically pushed "sensitive" DNA (active genes, promoters, enhancers, etc.) out of the active centromere thus preserving transcription and regulation patterns.

2) CR2 insertions increased the repeat content of neocentromeres. Centromeres are under physical stress during cell division and this often leads to double strand break (DSB) events [61]. The increase in CR2 repeat content can facilitate the efficient repair of these DSB events via homologous recombination, thus maintaining chromosome integrity. In addition to increases in copy number for full-length elements, both CR1 and CR4 have been observed to form tandem repeat arrays. In fact, CR1 has been observed to form tandem repeat arrays in both CEN9 [183], and independently on CEN2 (*SI Appendix Text SX10*). One might speculate that the centromeric repeats (CRs, CR derived tandem repeats, and CentC) act as a defined point of failure (much like the crumple zone in modern cars) thus preventing damage to sensitive regions of the chromosome.

Both mechanisms would, in theory, assist in the stabilization of neocentromere loci thus helping the host, and increase the copy number of the retrotransposon. While these lines of reasoning were not

definitive evidence of the functional importance of CR elements, they do suggest that CR elements play a pivotal role in centromere evolution and neocentromere stability.

6.5 Model of centromere evolution in domesticated maize

Neocentromere formation is a severe genomic event that can result in major changes to local epigenetic environment as well as global genomic changes such as chromosome breakage, fusion, and inversion [184, 185]. Stability is of key importance for neocentromeres. To remain functional, centromeres must maintain a minimum level of cenH3 and if the centromere is simultaneously accumulating deletions and constrained by flanking transcribed genes the centromere will become unstable. Previous studies have shown that centromere instability resulted from the loss of tandem repeats at the ancestral centromere locus [24]. We add to this by showing transcription of flanking genes prevented cenH3 expansion and resulted in pericentromeric neocentromere formation. The fact that maize CEN5 neocentromeres formed in formerly pericentromeric regions was not unexpected as pericentromeric DNA is highly methylated, and the highly repetitive (high retrotransposon content) sequence is packaged with post translationally modified histones that suppress transcription making them relatively ideal for neocentromere formation [41-43, 46]. The neocentromeres were further stabilized over time by the addition of repeats (CR elements) that can push sensitive DNA out of centromeres and the loss of H2A.Z via deletion and suppression by cenH3 being proposed as an intermediate state in the stabilization of neocentromeres [115]. Previous studies have also shown that centromere linked genes can reduce diversity at centromeres due to linkage drag effects. We added to this by exploring domestication loci proximal to two-haplotype (CEN2) and multi-haplotype (CEN5) centromeres which we compared to the single haplotype (CEN1), and exploring the prevalence of recombination near centromeres. We showed that centromeric haplotype regions were defined by selection and structural variation and that the tighter the linkage of a domesticated region to the active centromere, the less centromeric diversity was

observed. After the formation of neocentromeres they were colonized by CR elements. Sharma et al (PNAS paper) detailed the evolution of CR1 and suggested preferred recombinations that led to expansion of CR1 recombinant subtypes. We showed that the CR2 lineage expanded in a burst event during domestication not attributed to a preferred recombination. Our investigations suggested that the expansion of CR2 may have been the result of either a horizontal transfer of a foreign CR2 from a different *Zea* species or the result of epigenetic changes caused by movement of the centromere from its ancestral position. Regardless of the trigger for CR2 expansion, the insertions of CR2 helped to further stabilize the neocentromeres by increasing the repeat content and pushing genes/H2A.Z loci out of the active centromeres. Together these conclusions suggested that centromere evolution in domesticated maize was a story of stability. The process of domestication destabilized centromeres, and centromeres subsequently evolved in such a way to regain that stability in a new genomic context.

6.6 Conclusion

The works presented here can be distilled down to the following statement…

Neocentromere positioning in maize was dictated by the transcriptional and epigenetic landscape around the ancestral centromere and regions of centromeric recombination suppression were defined by structural variation and selection on linked regions, newly formed neocentromeres were stabilized post-domestication by a loss of H2A.Z and insertions of CR2 retrotransposons.

6.7 Future directions

The work presented here, expand the current body of knowledge, and answer some key questions regarding centromere biology and evolution but also present opportunities for further study. Three specific experiments fell outside the scope of this work but would be valuable to the advancement of

the field. First and foremost, would be the investigation of single haplotype centromeres to determine the precise genes that are under selection. This would require knock out or knock down experiments of genes in the centromeric haplotype regions. Specifically, CRISPR could be employed to deliver targeted mutations to genes in the centromeric haplotype regions to elucidate the functional importance of candidate genes presented in supplemental data, or for all the gene models in the region. If pseudogenes or mildly deleterious alleles were uncovered, multiplexed CRISPR may be used to knock out several genes simultaneously to see if the loss of the alleles confers a fitness advantage to the plant.

A second experiment that would be of value is the investigation of the geographic origins of the centromere haplotypes. A previous study had used SNPs and other markers across a slew of landraces, and teosinte from the Americas to determine the geographic origin of maize domestication [186]. From the geographic data and SNPs, they were able to determine the center of maize domestication. Determining the physical origins of the centromere haplotypes would add context to the story of domesticated maize centromere evolution by potentially explaining why certain maize centromere haplotypes were maintained or lost from the domesticated germplasm. I attempted this experiment using the genomic SNP data available for landraces and ancient maize lines deposited in the NCBI database, however I quickly realized that the marker density was too low, and quality too variable across datasets to reliably haplotype the centromere regions. To complete this experiment high-depth genomic sequence and geographic collection data, as well as a reliable centromere haplotyping pipeline would be required. This experiment may provide context to the single/two haplotype centromeres as their geographic origins (tropical/temperate/highland/lowland) may inform applied research (i.e. which centromere haplotype is “best” for a given region or heterotic group). However, it would be challenging to untangle human migration from the phylogenetics and geographic data.

The final direction worth exploring is an investigation of the ancestral relatives of *Zea mays*. Previous studies have reported *Zea mays* ssp. *mays* was domesticated from the wild ancestor *Zea mays* ssp. *parviglumis*. However, other ancestral species have contributed to mays germplasm including *Zea mays* ssp. *mexicana* and potentially *Zea luxurians,* and *Zea diploperennis.* The data herein have shown that structural variation is of key importance for maize centromere evolution and along with selection, defines the borders of the centromere haplotype regions. The physical maps of *parviglumis*, *mexicana* and *luxurians* would be of great value to the study of centromere evolution and to the overall maize genetics community. Our lab is currently not equipped to handle such large sequencing projects but the release of the physical maps for the NAM line founders suggests the groundwork has already been laid to undertake such an endeavor.

These three unexplored directions would provide additional context to the conclusions found in these works and may provide specific gene targets to apply the findings in this dissertation to maize germplasm improvement.

REFERENCES

1. Smith, M.R., et al., *Global Expanded Nutrient Supply (GENuS) Model: A New Method for Estimating the Global Dietary Supply of Nutrients.* PLOS ONE, 2016. **11**(1): p. e0146976.
2. Piperno, D.R. and K.V. Flannery, *The earliest archaeological maize (Zea mays L.) from highland Mexico: new accelerator mass spectrometry dates and their implications.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(4): p. 2101-2103.
3. Matsuoka, Y., et al., *A single domestication for maize shown by multilocus microsatellite genotyping.* Proceedings of the National Academy of Sciences, 2002. **99**(9): p. 6080.
4. Bennetzen, J., et al., *Genetic Evidence and the Origin of Maize.* Latin American Antiquity, 2001. **12**(1): p. 84-86.
5. Eyre-Walker, A., et al., *Investigation of the bottleneck leading to the domestication of maize.* Proceedings of the National Academy of Sciences, 1998. **95**(8): p. 4441.
6. Hilton, H. and B.S. Gaut, *Speciation and domestication in maize and its wild relatives: evidence from the globulin-1 gene.* Genetics, 1998. **150**(2): p. 863-872.
7. Bourguiba, H., et al., *Loss of genetic diversity as a signature of apricot domestication and diffusion into the Mediterranean Basin.* BMC plant biology, 2012. **12**: p. 49-49.
8. Reif, J.C., et al., *Wheat genetic diversity trends during domestication and breeding.* Theoretical and Applied Genetics, 2005. **110**(5): p. 859-864.
9. Tenaillon, M.I., et al., *Selection Versus Demography: A Multilocus Investigation of the Domestication Process in Maize.* Molecular Biology and Evolution, 2004. **21**(7): p. 1214-1225.
10. Doebley, J. and A. Stec, *Genetic analysis of the morphological differences between maize and teosinte.* Genetics, 1991. **129**(1): p. 285-295.
11. Doebley, J. and A. Stec, *Inheritance of the morphological differences between maize and teosinte: comparison of results for two F2 populations.* Genetics, 1993. **134**(2): p. 559-570.
12. Doebley, J., et al., *Genetic and morphological analysis of a maize-teosinte F2 population: implications for the origin of maize.* Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(24): p. 9888-9892.
13. Dorweiler, J., et al., *Teosinte glume architecture 1: A Genetic Locus Controlling a Key Step in Maize Evolution.* Science, 1993. **262**(5131): p. 233.
14. Hufford, M.B., et al., *Comparative population genomics of maize domestication and improvement.* Nature Genetics, 2012. **44**(7): p. 808-811.
15. Mahtani, M.M. and H.F. Willard, *Physical and genetic mapping of the human X chromosome centromere: repression of recombination.* (1088-9051 (Print)).
16. Stitzer, M.C. and J. Ross-Ibarra, *Maize domestication and gene interaction.* New Phytologist, 2018. **220**(2): p. 395-408.
17. Nannas, N.J. and R.K. Dawe, *Genetic and genomic toolbox of Zea mays.* Genetics, 2015. **199**(3): p. 655-669.
18. G.W., B., *A Possible Influence of the Spindle Fibre on Crossing-Over in Drosophila.* PNAS, A Possible Influence of the Spindle Fibre on Crossing-Over in Drosophila. **18**: p. 160-165.
19. ALBERT LEVAN, K.F.a.A.A.S., *NOMENCLATURE FOR CENTROMERIC POSITION ON CHROMOSOMES*, L. INSTITUTE OF GENETICS, SWEDEN, AND ROSWELL PARK MEMORIAL INSTITUTE, Editor. 1964: BUFFALO, N.Y., U.S.A.
20. Neumann, P., et al., *Stretching the Rules: Monocentric Chromosomes with Multiple Centromere Domains.* PLoS Genet, 2012. **8**(6): p. 1 -13
21. Kiyotaka Nagaki, K.K., and Minoru Murata, *Visualization of Diffuse Centromeres with Centromere-Specific Histone H3 in the Holocentric Plant Luzula nivea.* the Plant Cell, 2005. **17**: p. 1886-1893.

22. Zachary Lippman1*, A.-V.r.G., Michael Black3†,, et al., *Role of transposable elements in heterochromatin and epigenetic control.* Nature, 2004. **430**(6998): p. 467-71.
23. Andreas Houben, T.W., Rieko Furushima-, G.P. Shimogawara, Gottfried KuÈ nzel, Ingo, and S.a.K.c. Fukui, *The cell cycle dependant phosphorylation of histone H3 is correlated with the condensation of plant mitotic chromosomes.* The Plant Journal, 1999. **16**(6): p. 675 -679.
24. Schneider, K.L., et al., *Inbreeding drives maize centromere evolution.* Proc Natl Acad Sci U S A, 2016. **113**(8): p. E987-96.
25. Jiming Jiang, J.A.B., Wayne A. Parrott and R. Kelly Dawe, *A molecular view of plant centromeres.* Trends in Plant Science, 2003. **18**(12): p. 570-575.
26. Bodor, D.L., et al., *The quantitative architecture of centromeric chromatin.* eLife, 2014. **3**: p. e02137.
27. Tachiwana, H., et al., *Crystal structure of the human centromeric nucleosome containing CENP-A.* Nature, 2011. **476**: p. 232.
28. Dalal, Y., et al., *Structure, dynamics, and evolution of centromeric nucleosomes.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(41): p. 15974-15981.
29. Henikoff, S. and T. Furuyama, *The unconventional structure of centromeric nucleosomes.* Chromosoma, 2012. **121**(4): p. 341-352.
30. Black, B.E. and D.W. Cleveland, *Epigenetic Centromere Propagation and the Nature of CENP-A Nucleosomes.* Cell, 2011. **144**(4): p. 471-479.
31. Smith, M.M., *Centromeres and variant histones: what, where, when and why?* Current Opinion in Cell Biology, 2002. **14**(279-285).
32. Hasson, D., et al., *The octamer is the major form of CENP-A nucleosomes at human centromeres.* Nature structural & molecular biology, 2013. **20**(6): p. 687-695.
33. Annika Guse, C.W.C., Ben Moree, Colin J. Fuller & Aaron F. Straight, *In vitro centromere and kinetochore assembly on defined chromatin templates.* Nature, 2011. **477**(15): p. 354-358.
34. Fukagawa, T. and William C. Earnshaw, *The Centromere: Chromatin Foundation for the Kinetochore Machinery.* Developmental Cell, 2014. **30**(5): p. 496-508.
35. Blower, M.D., B.A. Sullivan, and G.H. Karpen, *Conserved Organization of Centromeric Chromatin in Flies and Humans.* Developmental cell, 2002. **2**(3): p. 319-330.
36. Nishino, T., et al., *CENP-T-W-S-X forms a unique centromeric chromatin structure with a histone-like fold.* Cell, 2012. **148**(3): p. 487-501.
37. Yu, H.-G., et al., *The plant kinetochore.* Trends in Plant Science, 2000. **5**(12): p. 543-547.
38. Braselton, J.P. and C.C. Bowen, *The Ultrastructure of the Kinetochores of Lilium Longiflorum During the first Meiotic Division.* Caryologia, 1971. **24**(1): p. 49-58.
39. Baskin, T.I. and W.Z. Cande, *The Structure and Function of the Mitotic Spindle in Flowering Plants.* Annual Review of Plant Physiology and Plant Molecular Biology, 1990. **41**(1): p. 277-315.
40. Dawe, R.K., et al., *A Maize Homolog of Mammalian CENPC Is a Constitutive Component of the Inner Kinetochore.* The Plant Cell, 1999. **11**(7): p. 1227.
41. Roudier, F., et al., *Integrative epigenomic mapping defines four main chromatin states in Arabidopsis.* The EMBO journal, 2011. **30**(10): p. 1928-1938.
42. Fransz, P., R. ten Hoopen, and F. Tessadori, *Composition and formation of heterochromatin in Arabidopsis thaliana.* Chromosome Research, 2006. **14**(1): p. 71-82.
43. Chodavarapu, R.K., et al., *Relationship between nucleosome positioning and DNA methylation.* Nature, 2010. **466**(7304): p. 388-392.
44. Bauwens, S., et al., *Distribution of the rDNA and three classes of highly repetitive DNA in the chromatin of interphase nuclei of Arabidopsis thaliana.* Chromosoma, 1991. **101**(1): p. 41-48.

45. Nasmyth, K. and C.H. Haering, *Cohesin: Its Roles and Mechanisms.* Annual Review of Genetics, 2009. **43**(1): p. 525-558.
46. Gent, J.I., et al., *Strong epigenetic similarity between maize centromeric and pericentromeric regions at the level of small RNAs, DNA methylation and H3 chromatin modifications.* Nucleic acids research, 2012. **40**(4): p. 1550-1560.
47. Yuan, G.-C., et al., *Genome-Scale Identification of Nucleosome Positions in <em>S. cerevisiae</em&gt.* Science, 2005. **309**(5734): p. 626.
48. Raisner, R.M., et al., *Histone Variant H2A.Z Marks the 5′ Ends of Both Active and Inactive Genes in Euchromatin.* Cell, 2005. **123**(2): p. 233-248.
49. Santisteban, M.S., T. Kalashnikova, and M.M. Smith, *Histone H2A.Z Regulates Transcription and Is Partially Redundant with Nucleosome Remodeling Complexes.* Cell, 2000. **103**(3): p. 411-422.
50. Meneghini, M.D., M. Wu, and H.D. Madhani, *Conserved Histone Variant H2A.Z Protects Euchromatin from the Ectopic Spread of Silent Heterochromatin.* Cell, 2003. **112**(5): p. 725-736.
51. Zilberman, D., et al., *Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks.* Nature, 2008. **456**(7218): p. 125-129.
52. Dion, M.F., et al., *Dynamics of Replication-Independent Histone Turnover in Budding Yeast.* Science, 2007. **315**(5817): p. 1405.
53. Mizuguchi, G., et al., *ATP-Driven Exchange of Histone H2AZ Variant Catalyzed by SWR1 Chromatin Remodeling Complex.* Science, 2004. **303**(5656): p. 343.
54. Zemach, A., et al., *Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation.* Science, 2010. **328**(5980): p. 916.
55. Conerly, M.L., et al., *Changes in H2A.Z occupancy and DNA methylation during B-cell lymphomagenesis.* Genome research, 2010. **20**(10): p. 1383-1390.
56. Murphy, P.J., et al., *Placeholder Nucleosomes Underlie Germline-to-Embryo DNA Methylation Reprogramming.* Cell, 2018. **172**(5): p. 993-1006.e13.
57. Yamada, S., et al., *The histone variant H2A.Z promotes initiation of meiotic recombination in fission yeast.* (1362-4962 (Electronic)).
58. Choi, K., et al., *Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters.* Nature Genetics, 2013. **45**(11): p. 1327-1336.
59. Shi, J., et al., *Widespread gene conversion in centromere cores.* PLoS Biol, 2010. **8**(3): p. e1000327.
60. Bennetzen, J.M.a.J.L., *Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice.* Proc Natl Acad Sci U S A, 2006. **103**(2): p. 383-388.
61. Wolfgruber, T.K., et al., *High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events.* Front Plant Sci, 2016. **7**: p. 308.
62. Tanksley, S.D., et al., *High density molecular linkage maps of the tomato and potato genomes.* Genetics, 1992. **132**(4): p. 1141-1160.
63. Jacobs, P.A., *Mutation rates of structural chromosome rearrangements in man.* American journal of human genetics, 1981. **33**(1): p. 44-54.
64. Murat, F., et al., *Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution.* Genome research, 2010. **20**(11): p. 1545-1557.
65. Bowers, J.E., et al., *Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(37): p. 13206-13211.
66. Marshall, O.J., et al., *Neocentromeres: New Insights into Centromere Structure, Disease Development, and Karyotype Evolution.* American Journal of Human Genetics, 2008. **82**(2): p. 261-282.

67. Buckler, E.S., et al., *Meiotic drive of chromosomal knobs reshaped the maize genome.* Genetics, 1999. **153**(1): p. 415-426.
68. Alonso, A., et al., *A paucity of heterochromatin at functional human neocentromeres.* Epigenetics & Chromatin, 2010. **3**: p. 6-6.
69. Zhang, B., et al., *Formation of a functional maize centromere after loss of centromeric sequences and gain of ectopic sequences.* Plant Cell, 2013. **25**(6): p. 1979-89.
70. Nasuda, S., et al., *Stable barley chromosomes without centromeric repeats.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(28): p. 9842-9847.
71. Yan, H., et al., *Intergenic Locations of Rice Centromeric Chromatin.* PLOS Biology, 2008. **6**(11): p. e286.
72. Nagaki, K., et al., *Sequencing of a rice centromere uncovers active genes.* Nat Genet, 2004. **36**(2): p. 138-145.
73. Yasuhara, J.C. and B.T. Wakimoto, *Oxymoron no more: the expanding world of heterochromatic genes.* Trends in Genetics, 2006. **22**(6): p. 330-338.
74. Miller, J.T., et al., *Retrotransposon-related DNA sequences in the centromeres of grass chromosomes.* Genetics, 1998. **150**(4): p. 1615-1623.
75. Benjamin Gorinsek, F.G., and Dusan Kordis, *Evolutionary Genomics of Chromoviruses in Eukaryotes.* Molecular Biology and Evolution, 2004. **21**(5): p. 781-798.
76. Pavel Neumann, A.N., Andrea Koblížková, Eduard Kejnovský, Eva Hřibová3, Roman Hobza, Alex Widmer, Jaroslav Doležel, Jiří Macas, *Plant centromeric retrotransposons: a structural and cytogenetic perspective.* Mobile DNA, 2011. **2**(4): p. 1-16.
77. Gernot G. Presting, L.M., Jorg Fuchs and Ingo Schubert, *A Ty3/Gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes.* The Plant Journal, 1998. **16**(6): p. 721 - 728.
78. Gao, X., et al., *Chromodomains direct integration of retrotransposons to heterochromatin.* Genome research, 2008. **18**(3): p. 359-369.
79. Sharma, A. and G.G. Presting, *Evolution of centromeric retrotransposons in grasses.* Genome Biol Evol, 2014. **6**(6): p. 1335-52.
80. Phillip SanMiguel, Brandon S. Gaut, Alexander Tikhonov, Yuko Nakajima & Jeffrey L. Bennetzen, *The paleontology of intergene retrotransposons of maize.* Nature Genetics, 1998. **20**: p. 43-45.
81. Cédric Feschotte, N.J.a.S.R.W., *Plant transposable elements: Where genetics meets genomics.* Nature Reviews Genetics, 2002. **3**: p. 329-341.
82. Sharma, A., K.L. Schneider, and G.G. Presting, *Sustained retrotransposition is mediated by nucleotide deletions and interelement recombinations.* Proc Natl Acad Sci U S A, 2008. **105**(40): p. 15470-4.
83. Hu, W.-S. and H.M. Temin, *Retroviral Recombination and Reverse Transcription.* Science, 1990. **250**(4985): p. 1227-1233.
84. Anupma Sharma, T.K.W.a.G.G.P., *Tandem repeats derived from centromeric retrotransposons.* Biomed Central Genomics, 2013. **14**(142): p. 1 - 10.
85. Bennetzen, J.L., J. Ma, and K.M. Devos, *Mechanisms of Recent Genome Size Variation in Flowering Plants.* Annals of Botany, 2005. **95**(1): p. 127-132.
86. Zhang, Q.-J. and L.-Z. Gao, *Rapid and Recent Evolution of LTR Retrotransposons Drives Rice Genome Evolution During the Speciation of AA-Genome Oryza Species.* G3: Genes|Genomes|Genetics, 2017. **7**(6): p. 1875-1885.
87. BELYAYEV, A., *Bursts of transposable elements as an evolutionary driving force.* Evolutionary Biology, 2014. **27**: p. 2573-2584.
88. Ma, J., K.M. Devos, and J.L. Bennetzen, *Analyses of LTR-Retrotransposon Structures Reveal Recent and Rapid Genomic DNA Loss in Rice.* Genome Research, 2004. **14**(5): p. 860-869.

89. Vitte, C., O. Panaud, and H. Quesneville, *LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss.* BMC Genomics, 2007. **8**: p. 218-218.
90. Tian, Z., et al., *Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons?* Genome Research, 2009. **19**(12): p. 2221-2230.
91. Devos, K.M., J.K.M. Brown, and J.L. Bennetzen, *Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis.* Genome Research, 2002. **12**(7): p. 1075-1079.
92. McClintock, B., *Mutable loci in maize.* Carnegie Institution of Washington Yearbook, 1951. **50**: p. 174-181.
93. McClintock, B., *Controlling elements and the gene.* Cold Spring Harbor symposia on quantitative biology, 1956. **21**: p. 197.
94. Walsh, A.M., et al., *Widespread horizontal transfer of retrotransposons.* Proceedings of the National Academy of Sciences, 2013. **110**(3): p. 1012.
95. Jordan, I.K., L.V. Matyunina, and J.F. McDonald, *Evidence for the recent horizontal transfer of long terminal repeat retrotransposon.* Proceedings of the National Academy of Sciences, 1999. **96**(22): p. 12621.
96. Schubert, R.t.H.R.M.J.D.L.M.I., *Evolutionary conservation of kinetochore protein sequences in plants.* Chromosoma, 2000. **109**: p. 482-489.
97. Stephen Henikoff, K.A., Harmit S. Malik, *The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA.* Science, 2001. **293**: p. 1098-1102.
98. Gershenson, S., *A New Sex-Ratio Abnormality in DROSOPHILA OBSCURA.* Genetics, 1928. **13**(6): p. 488-507.
99. Chmátal, L., et al., *Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice.* (1879-0445 (Electronic)).
100. Rhoades, M.M., *PREFERENTIAL SEGREGATION IN MAIZE.* Genetics, 1942. **27**(4): p. 395.
101. Dawe, R.K. and W.Z. Cande, *Induction of centromeric activity in maize by suppressor of meiotic drive 1.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(16): p. 8512-8517.
102. Edward S. Buckler IV, T.L.P.-D., Carlyn S. Keith Buckler, R. Kelly Dawe, John F. Doebley and Timothy P. Holtsford, *Meiotic Drive of Chromosomal Knobs Reshaped the Maize Genome.* Genetics, 1999. **153**: p. 415-426.
103. Amor, D.J. and K.H.A. Choo, *Neocentromeres: role in human disease, evolution, and centromere study.* American journal of human genetics, 2002. **71**(4): p. 695-714.
104. Schnable, P.S., et al., *The B73 Maize Genome: Complexity, Diversity, and Dynamics.* Science, 2009. **326**(5956): p. 1112.
105. Jiao, Y., et al., *Improved maize reference genome with single-molecule technologies.* Nature, 2017. **546**(7659): p. 524-527.
106. Hufford, M.B., et al., *De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes.* Science, 2021. **373**(6555): p. 655.
107. Kurtz, S., et al., *Versatile and open software for comparing large genomes.* Genome biology, 2004. **5**(2): p. R12-R12.
108. Wolfgruber, T.K. and G.G. Presting, *JunctionViewer: customizable annotation software for repeat-rich genomic regions.* BMC bioinformatics, 2010. **11**: p. 23-23.
109. Ananiev, E.V., R.L. Phillips, and H.W. Rines, *Chromosome-specific molecular organization of maize (Zea mays L.) centromeric regions.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(22): p. 13073-13078.

110. Sharma, A. and G.G. Presting, *Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity.* Molecular Genetics and Genomics, 2008. **279**(2): p. 133-147.
111. Sharma, A. and G.G. Presting, *Evolution of centromeric retrotransposons in grasses.* Genome biology and evolution, 2014. **6**(6): p. 1335-1352.
112. Page, B.T., M.K. Wanous, and J.A. Birchler, *Characterization of a maize chromosome 4 centromeric sequence: evidence for an evolutionary relationship with the B chromosome centromere.* Genetics, 2001. **159**(1): p. 291-302.
113. Kearse, M., et al., *Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data.* Bioinformatics, 2012. **28**(12): p. 1647-1649.
114. Albert, P.S., et al., *Diversity of Chromosomal Karyotypes in Maize and Its Relatives.* Cytogenetic and Genome Research, 2010. **129**(1-3): p. 6-16.
115. Ogiyama, Y., et al., *Epigenetically induced paucity of histone H2A.Z stabilizes fission-yeast ectopic centromeres.* Nature Structural & Molecular Biology, 2013. **20**(12): p. 1397-1406.
116. Shang, W.-H., et al., *Chromosome engineering allows the efficient isolation of vertebrate neocentromeres.* Developmental cell, 2013. **24**(6): p. 635-648.
117. Maggert, K.A. and G.H. Karpen, *The activation of a neocentromere in Drosophila requires proximity to an endogenous centromere.* Genetics, 2001. **158**(4): p. 1615-1628.
118. Olszak, A.M., et al., *Heterochromatin boundaries are hotspots for de novo kinetochore formation.* Nature Cell Biology, 2011. **13**(7): p. 799-808.
119. Nishimura, K., et al., *3D genomic architecture reveals that neocentromeres associate with heterochromatin regions.* The Journal of cell biology, 2019. **218**(1): p. 134-149.
120. Sura, W., et al., *Dual Role of the Histone Variant H2A.Z in Transcriptional Regulation of Stress-Response Genes.* The Plant Cell, 2017. **29**(4): p. 791-807.
121. Presting, G.G., *Centromeric retrotransposons and centromere function.* Current Opinion in Genetics & Development, 2018. **49**: p. 79-84.
122. Bauwens, S., et al., *Distribution of the rDNA and three classes of highly repetitive DNA in the chromatin of interphase nuclei of Arabidopsis thaliana.* Chromosoma, 2004. **101**: p. 41-48.
123. Zlatanova, J. and A. Thakar, *H2A.Z: View from the Top.* Structure, 2008. **16**(2): p. 166-179.
124. Neumann, P., H. Yan, and J. Jiang, *The centromeric retrotransposons of rice are transcribed and differentially processed by RNA interference.* Genetics, 2007. **176**(2): p. 749-761.
125. Lisch, D., *Epigenetic Regulation of Transposable Elements in Plants.* Annual Review of Plant Biology, 2009. **60**(1): p. 43-66.
126. Lermontova, I., et al., *Loading of Arabidopsis centromeric histone CENH3 occurs mainly during G2 and requires the presence of the histone fold domain.* The Plant cell, 2006. **18**(10): p. 2443-2451.
127. Potok, M.E., et al., *Arabidopsis SWR1-associated protein methyl-CpG-binding domain 9 is required for histone H2A.Z deposition.* Nature Communications, 2019. **10**(1): p. 3352.
128. Stellfox, M.E., A.O. Bailey, and D.R. Foltz, *Putting CENP-A in its place.* Cellular and molecular life sciences : CMLS, 2013. **70**(3): p. 387-406.
129. Talbert, P.B. and S. Henikoff, *Environmental responses mediated by histone variants.* Trends in Cell Biology, 2014. **24**(11): p. 642-650.
130. Kumar, S.V. and P.A. Wigge, *H2A.Z-Containing Nucleosomes Mediate the Thermosensory Response in Arabidopsis.* Cell, 2010. **140**(1): p. 136-147.
131. Berriri, S., S.N. Gangappa, and S.V. Kumar, *SWR1 Chromatin-Remodeling Complex Subunits and H2A.Z Have Non-overlapping Functions in Immunity and Gene Regulation in Arabidopsis.* Molecular plant, 2016. **9**(7): p. 1051-1065.

132. Smith, A.P., et al., *Histone H2A.Z regulates the expression of several classes of phosphate starvation response genes but not as a transcriptional activator.* Plant physiology, 2010. **152**(1): p. 217-225.
133. Laspisa, D.J., K.L. Schneider, and G.G. Presting, *Improved Centromere Assemblies for RefGen_v4.* bioRxiv, 2022: p. 2022.03.30.486462.
134. Burrack, L.S. and J. Berman, *Neocentromeres and epigenetically inherited features of centromeres.* Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 2012. **20**(5): p. 607-619.
135. Ishii, K., et al., *Heterochromatin Integrity Affects Chromosome Reorganization After Centromere Dysfunction.* Science, 2008. **321**(5892): p. 1088-1091.
136. Topp, C.N. and R.K. Dawe, *Reinterpreting pericentromeric heterochromatin.* Current Opinion in Plant Biology, 2006. **9**(6): p. 647-653.
137. Glémin, S. and T. Bataillon, *A comparative view of the evolution of grasses under domestication.* (1469-8137 (Electronic)).
138. Haudry, A., et al., *Grinding up wheat: a massive loss of nucleotide diversity since domestication.* (0737-4038 (Print)).
139. Sanou, H., et al., *Loss of Genetic Diversity of Jatropha curcas L. through Domestication: Implications for Its Genetic Improvement.* Crop Science, 2015. **55**(2): p. 749-759.
140. Dixon, T.J., et al., *Shifts in genetic diversity during domestication of Black Tiger shrimp, Penaeus monodon, monitored using two multiplexed microsatellite systems.* Aquaculture, 2008. **283**(1-4): p. 1-6.
141. Choi, K., *Advances towards Controlling Meiotic Recombination for Plant Breeding.* (0219-1032 (Electronic)).
142. Lyzenga, W.J., C.J. Pozniak, and S. Kagale, *Advanced domestication: harnessing the precision of gene editing in crop breeding.* Plant Biotechnology Journal, 2021. **19**(4): p. 660-670.
143. May, K.M., et al., *The parental origin of the extra X chromosome in 47,XXX females.* (0002-9297 (Print)).
144. Sears, D.D., et al., *Cis-acting determinants affecting centromere function, sister-chromatid cohesion and reciprocal recombination during meiosis in Saccharomyces cerevisiae.* (0016-6731 (Print)).
145. Nambiar, M. and G.R. Smith, *Repression of harmful meiotic recombination in centromeric regions.* (1096-3634 (Electronic)).
146. Underwood, C.J., et al., *Epigenetic activation of meiotic recombination near Arabidopsis thaliana centromeres via loss of H3K9me2 and non-CG DNA methylation.* (1549-5469 (Electronic)).
147. Darling, A.E., B. Mau, and N.T. Perna, *progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement.* PLOS ONE, 2010. **5**(6): p. e11147.
148. Kumar, S., et al., *MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis.* Bioinformatics (Oxford, England), 2012. **28**(20): p. 2685-2686.
149. Clark, R.M., S. Tavare, and J. Doebley, *Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus.* Mol Biol Evol, 2005. **22**(11): p. 2304-12.
150. Tian, F., N.M. Stevens, and E.S. Buckler, *Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10.* Proceedings of the National Academy of Sciences, 2009. **106**(Supplement 1): p. 9979.
151. Andorf, C.M., et al., *MaizeGDB update: new tools, data and interface for the maize model organism database.* Nucleic Acids Research, 2016. **44**(D1): p. D1195-D1201.
152. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.* Molecular biology and evolution, 2018. **35**(6): p. 1547-1549.

153. Bukowski, R., et al., *Construction of the third-generation Zea mays haplotype map.* GigaScience, 2018. **7**(4): p. 1-12.
154. Fang, Z., et al., *Megabase-scale inversion polymorphism in the wild ancestor of maize.* (1943-2631 (Electronic)).
155. Fuentes, R.R., et al., *Domestication Shapes Recombination Patterns in Tomato. LID - 10.1093/molbev/msab287 [doi] LID - msab287.* (1537-1719 (Electronic)).
156. Zapata, L., et al., *Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms.* Proceedings of the National Academy of Sciences, 2016. **113**(28): p. E4052-E4060.
157. Bansal, V., A. Bashir, and V. Bafna, *Evidence for large inversion polymorphisms in the human genome from HapMap data.* Genome research, 2007. **17**(2): p. 219-230.
158. Munté, A., et al., *Chromosomal inversion polymorphism leads to extensive genetic structure: a multilocus survey in Drosophila subobscura.* Genetics, 2005. **169**(3): p. 1573-1581.
159. Sun, S., et al., *Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes.* Nature Genetics, 2018. **50**(9): p. 1289-1295.
160. Lemmon, Z.H., et al., *The Role of cis Regulatory Evolution in Maize Domestication.* PLOS Genetics, 2014. **10**(11): p. e1004745.
161. Danilevskaya, O.N., X. Meng, and E.V. Ananiev, *Concerted modification of flowering time and inflorescence architecture by ectopic expression of TFL1-like genes in maize.* Plant physiology, 2010. **153**(1): p. 238-251.
162. Beissinger, T.M., et al., *Recent demography drives changes in linked selection across the maize genome.* Nature Plants, 2016. **2**(7): p. 16084.
163. Petit, M., et al., *Mobilization of retrotransposons in synthetic allotetraploid tobacco.* New Phytologist, 2010. **186**(1): p. 135-147.
164. O'Neill, R.J.W., M.J. O'Neill, and J.A.M. Graves, *Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid.* Nature, 1998. **393**(6680): p. 68-72.
165. Cavrak, V.V., et al., *How a retrotransposon exploits the plant's heat stress response for its activation.* PLoS genetics, 2014. **10**(1): p. e1004115-e1004115.
166. McClintock, B., *The origin and behavior of mutable loci in maize.* Proceedings of the National Academy of Sciences, 1950. **36**(6): p. 344.
167. Griffith, J.L., et al., *Functional genomics reveals relationships between the retrovirus-like Ty1 element and its host Saccharomyces cerevisiae.* Genetics, 2003. **164**(3): p. 867-879.
168. Yang, N. and H.H. Kazazian, *L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells.* Nature Structural & Molecular Biology, 2006. **13**(9): p. 763-771.
169. Xie, Z., et al., *Genetic and functional diversification of small RNA pathways in plants.* PLoS biology, 2004. **2**(5): p. E104-E104.
170. Xie, M. and B. Yu, *siRNA-directed DNA Methylation in Plants.* Current genomics, 2015. **16**(1): p. 23-31.
171. Hufford, M.B., et al., *De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes.* bioRxiv, 2021: p. 2021.01.14.426684.
172. Katoh, K. and D.M. Standley, *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.* Molecular Biology and Evolution, 2013. **30**(4): p. 772-780.
173. Tenaillon, M.I., et al., *Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians.* Genome biology and evolution, 2011. **3**: p. 219-229.

174. Chia, J.-M., et al., *Maize HapMap2 identifies extant variation from a genome in flux.* Nature Genetics, 2012. **44**(7): p. 803-807.
175. Bukowski, R., et al., *Construction of the third generation Zea mays haplotype map.* GigaScience, 2017: p. gix134-gix134.
176. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Meth, 2012. **9**(4): p. 357-359.
177. Feng, Y.X., et al., *The genomic RNA in Ty1 virus-like particles is dimeric.* Journal of virology, 2000. **74**(22): p. 10819-10821.
178. Ross-Ibarra, J., M. Tenaillon, and B.S. Gaut, *Historical divergence and gene flow in the genus Zea.* Genetics, 2009. **181**(4): p. 1399-1413.
179. El Baidouri, M., et al., *Widespread and frequent horizontal transfers of transposable elements in plants.* Genome research, 2014. **24**(5): p. 831-838.
180. Koo, D.-H., et al., *Distinct DNA methylation patterns associated with active and inactive centromeres of the maize B chromosome.* Genome research, 2011. **21**(6): p. 908-914.
181. Du, C., Z. Swigoňová, and J. Messing, *Retrotranspositions in orthologous regions of closely related grass species.* BMC Evolutionary Biology, 2006. **6**(1): p. 62.
182. Estep, M.C., J.D. DeBarry, and J.L. Bennetzen, *The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution.* Heredity, 2013. **110**(2): p. 194-204.
183. Sharma, A., T.K. Wolfgruber, and G.G. Presting, *Tandem repeats derived from centromeric retrotransposons.* BMC Genomics, 2013. **14**(1): p. 142.
184. Schotanus, K., V. Yadav, and J. Heitman, *Epigenetic dynamics of centromeres and neocentromeres in Cryptococcus deuterogattii.* PLOS Genetics, 2021. **17**(8): p. e1009743.
185. Barra, V. and D. Fachinetti, *The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA.* Nature Communications, 2018. **9**(1): p. 4340.
186. van Heerwaarden, J., et al., *Genetic signals of origin, spread, and introgression in a large sample of maize landraces.* Proceedings of the National Academy of Sciences, 2011. **108**(3): p. 1088.

APPENDIX

Table of Contents

Supplemental Background 113
SB1 Tandem repeats and tandem repeat evolution 113
SB2 DNA methylation in the centromere 114
SB3 A primer on transposable elements 115
SB3.1 LTR transposable elements 115
SB3.2 An overview of Class 1 TE transposition 116
SB3.3 Non-LTR transposable elements 117
SB3.4 Class 2 Transposable elements 117
SB3.5 MITES, Mutators and Helitrons 118
SB3.6 Transposable element regulation 119
Supplemental Text 120
SX1. Helitron related gene graveyards 120
SX2. Gene damage can result in persistent H2A.Z loci 123
CEN4 123
CEN9 124
CEN10 124
SX3. New substitution rates calculated from six genomes 124
SX4. Comparison of sequence divergence between physical maps and HM3 SNP data 125
SX5. Deletions in NC358/Oh43 result in HapMap3 artifacts 126
SX6. The CEN2 HAP2B inversion breakpoints are correlated with deactivated centromeres 126
SX7. The progenitor of the CEN2 domestication region 127
SX8. A distantly related pericentromeric introgression in CEN9 128
SX9. CEN9 proximal introgression contains PHS1 129
SX10. Independent CTR1 formation in CEN9 and CEN2 130
SX11. Discrepancies between HapMap3 and the NAM Physical maps 132
SX12. CentC loss during inbreeding 133
SX13. H2A.Z in CR2s 134
SX14. TE Insertions of Centromeric and Pericentromeric genes 136
SX15. The Landscape of ZmCEN3, ZmCEN8 and Orthologous OsCEN1 (*Oryza sativa*) and SiCEN5 (*Setaria italica*) 137

SX16. Gene Fractionation and Loss 138
SX17. Gene Movement and Acquisition 139
SX18. Shared Genes and Retained Centromeric Genes 140
Supplemental Figures 142
Figure S1. CEN5 tree generated from HapMap3 SNP data. 142
Figure S2. Nucleotide level coverage levels for cenH3 (blue), H2AZ (green) and RNAseq (Orange) at positions covered by the maximal 101mers. 143
Figure S3. RNAseq coverage for 73 datasets mapped to B73 143
Figure S4. Variation in MACS2 H2A.Z Peak calling among Orange lines using varied fold change thresholds. 144
Figure S5. Single peaks/deletions are not responsible for the differential ratios observed between CEN5R and CEN5L lines. 145
Figure S6. Overview of H2A.Z and RNAseq peaks in the active B73_v4CEN centromeres. 146
Figure S7. Transcription over the CEN5 gene graveyard. 147
Figure S8. Schematic of Gene graveyard alignment. 147
Figure S9. Isolated H2A.Z loci lacking RNA-seq can form due to gene damage. 149
Figure S10. CR2 Associated H2A.Z peaks in B73_v5. 150
Figure S11. Defined and delocalized nucleosome positioning influences the observed inverse H2A.Z/cenH3 relationship. 151
Figure S12. CEN2 HapMap3 SNP tree for REGON13_SYN 152
Figure S13. CEN2 Physical Map tree for REGION13_SYN 153
Figure S14. Comparisons of Physical map and HapMap3 K2p over equivalent regions for CENs 1 (a), 4 (b), 6 (c) and 2 (d). 157
Figure S15. Distribution of divergence dates among alignments of CEN1, CEN2, CEN4, and CEN6. 158
Figure S16. SNP density relative to B73 v4 per 100 covered HapMap3 sites across CEN2. 158
Figure S17. MUMmer (word length 200) of NC358 and Oh43. 159
Figure S18. Syntenic relationships between Zea mays chromosome 2 (B73_v5) and Sorghum bicolor (NCBI_v3, GCF_000003195.3). 159
Figure S19. Syntenic relationships between Zea mays (B73_v5) and Sorghum bicolor (NCBI_v3, GCF_000003195.3) 160
Figure S20. TIL progenitors and recombination breakpoints for 56 domesticated maize lines. 161
Figure S21. HapMap3 trees of the CEN2 domestication region 162
Figure S22. CEN2 gene models and orthologs among the NAM lines. 163
Figure S23. Schematic diagram of the introgressed region in the upstream pericentromere of B73 v4CEN. 164

Figure S24. Phylogenetic tree of the introgressed PHS1 gene. 165
Figure S25. Schematic phylogeny of the CEN9 PHS1 gene. 166
Figure S26. Schematic diagram of B73 v4CEN position and divergence dates of full-length monomers. 166
Figure S27. Phylogenetic tree of CEN2 CTR1 monomers to CR1 consensus sequences. 167
Figure S28. Junction viewer [70] annotation of the CEN2 CTR1 loci in CML277 (top) and P39 (bottom). 168
Figure S29. Correlation between sequence that is not present in B73 and the R2 value for Mauve/HapMap3 comparisons. 168
Figure S30. Comparison of phylogenetic trees for matching regions for HapMap3 and Mauve in the CEN1 (a), CEN4 (b.) and CEN6 (c) alignments. 170
Figure S31. Comparison of trees corresponding to the 500kb upstream (US) and downstream (DS) of the CEN6 Mauve alignment region. 171
Figure S32. Phylogenetic tree of unimputed HapMap3 data over the CEN6 alignment region. 172
Figure S33. CEN2 HapMap3 Subtrees 173
Figure S34. CEN2 Bootstrap consensus trees 174
Figure S35. B73 v5 CR1 insertion date violin plot by subtype. 175
Figure S36. Distribution of UMRs across the B73 v5 genome. 176
Figure S37. Distribution of UMRs across the CEN5 lineages. 177
Figure S38. UMR, H2A.Z and cenH3 coverage for CR2 parental subtypes. 178
Figure S39. UMR, H2A.Z and cenH3 coverage for CR1 subtypes. 180
Figure S40. CR2 sub-recombinant phylogenetic trees. 185
Figure S41. Enrichment of SNPs in NAM compared to teosinte. 186
Figure S42. SNP density of NAM, TIL and landraces compared to Zea luxurians in CENs 1 (e), 5 (a), 6 (b), 8 (c), and 9 (d). 191
Figure S43. Percentage of nucleotides corresponding to centC in six cultivars across six generations of inbreeding. 192
Figure S44. Schematic phylogeny of CEN5 AUG6 and PHS1 Genes. 193
Figure S45. Syntenic regions of Os1, Zm3 and Zm8. 194
Figure S46. CEN2 CR2 and CR1 gene models and orthologs among the NAM lines. 195
Figure S47. CEN5 Schematic overview, gene models, and orthologs among the NAM lines. 196
Figure S48. CEN5 CR2 and CR1 gene models and orthologs among the NAM lines. 197
Supplemental Tables 199
Table ST1. Details of expressed genes flanking CEN5M. B73 gene models corresponding to CEN5L markers A-D and CEN5R markers E-H 199

Table ST2. Coordinates and H2A.Z coverage data for the 26 NAM lines in the A-D region and the E-H region. 201
Table ST3. Ratios for CEN5L and CEN5R among the NAM lines 202
Table ST4. CEN5L deletion details. 202
Table ST5. H2A.Z Read count and mapping percentage. 204
Table ST6. Maximum H2A.Z coverage over CR1 subtypes and a Consensus CR2 element compared to the CEN5 H2A.Z control peak. 204
Table ST7. Coordinates of Region13_SYN in B73 v4, v5 and HapMap3 SNP coordinates. 204
Table ST8. Coordinates and anchors for the regions compared between HapMap3 and the NAM physical maps. 205
Table ST9. Nucleotide substitution rate for dating maize centromere sequence. 205
Table ST10. B73 v5 coordinates of the deletions in NC358/Oh43 and the estimated deletion size. 206
Table ST11. HapMap3 SNP counts for different comparisons across the CEN2 domestication region. 206
Table ST12. Coordinates of the CML277/BKN009 introgression over Zm00001eb087250. 207
Table ST13. B73 v5 gene models between the recombination locus R3 and the B73 active centromere. 207
Table ST14. Physical map coordinates for the CEN2 HAP2B inversion breakpoints in each of the four physical maps containing the inversion. 207
Table ST15. SNP analysis of the PHS1 gene. 208
Table ST16. Deletion junctions for P39 Monomers in the CEN2 CTR1 locus. 209
Table ST17. Coordinates of CEN1 alignment subsections used to search for potential local divergence maxima 209
Table ST18. Comparisons of divergence date (ka) for single haplotype centromeres CEN1 CEN4 and CEN6. 209
Table ST19. CR2 deletion variant details 210
Table ST20. Divergence (ka) between CR1 and CR2 elements 210
Table ST21. Z. luxurians putative shared CR2s 211
Table ST22. CR counts in the active centromere by subtype 212
Table ST23. CR1s in post-domestication neocentromeres 213
Table ST24. Datasets unavailable for analysis of centC loss. 213
Table ST25. Syntenic coordinates and genes of Os1, Zm3 and Zm8. 214
SUPPLEMENTAL REFERENCES 215

Supplemental Background

SB1 Tandem repeats and tandem repeat evolution

Tandem repeats on the most basic level may be defined as two or more contiguous, approximate copies of a pattern of nucleotides and may be either local or global [187, 188]. Local tandem repeats being chromosome specific, and global repeats being distributed genome wide. In *Zea mays* the primary centromeric tandem repeat is CentC, which has been shown to interact with cenH3 [25, 189]. The size of the tandem arrays can vary greatly from just a few thousand bases (kb) to megabases (Mb) in length and consist primarily of short (156nt) monomers [188]. Tandem repeat monomers are often species specific and are usually between 100 and 200bp, however repeats up to 1.4kb have been described [189-191]. Monomers are not strictly identical and often contain polymorphisms and form higher order repeats (HOR) which can be defining features of repeat islands [188]. Although referred to as "centromeric tandem repeats", islands of these monomers may also be found in pericentromeric sequences, which is the case with multiple *Zea mays* centromeres (CEN1, CEN3, CEN5, CEN7, CEN10) [24]. This movement of CentC is likely due to hemicentric inversion or double strand break (DSB) and repair [61]. It is important to note that these ectopic centC islands do not retain centromere activity but can still expand or contract. It has been shown in yeast that DSB and repair are a major source of amplification or deletion of tandem repeat monomers though variation in a tandem repeat island can also occur via segmental duplication, replication slippage, recombination leading to unequal crossing over, or gene conversion [84, 192] [193]. These mechanisms can result in considerable variation in tandem repeat structure even between closely related maize lines. Interestingly, investigation of the human □-satellite repeat has demonstrated faster SNP acquisition rates in centromeric and pericentromeric regions than in the chromosome arms [191, 194]. This is an important observation as it suggests the evolutionary rate within centromeres and pericentromeres is faster than on the chromosome arms. As such, a centromere specific conversion factor is required when calculating divergence time.

The underlying functional significance of centromeric tandem repeats is unknown, however there has been significant speculation regarding their putative function. A more classical hypothesis suggests that centromeric tandem repeats may aid in chromosome disjunction during cell division [195]. A more recent proposed role is that tandem repeats allow for kinetochore plasticity [34]. However, one could speculate that due to the mechanical nature of the centromere, and the frequency of DSB events, repeats act as a defined region of structural failure. They simultaneously protect centromere proximal genes, and allow for easy chromosome repair much like the “crumple zone” in modern automobiles [121]. Such a function could also explain the large variety of centromeric tandem repeats as a repeat need only be repetitive to function in this manner.

SB2 DNA methylation in the centromere

Histones and their variants affect chromatin composition and transcription in the regional level but at the nucleotide level DNA methylation acts as a key regulator of transcription. DNA methylation in plants can occur in three contexts CG, CHG and CHH where H is any nucleotide but G. Analysis of centromeres using 5-methyl-cytosine (5mC) immunostaining have reported that centC and CR elements are hypomethylated in the active centromere but centromeric repeats outside of the active centromere, were hypermethylated [180]. This contrasts with more recent bisulfite sequencing analysis of methylation in the centromeres of maize which reported a gradual increase in methylation as one approaches the centromere. Though notably, this analysis did detect lower CHG methylation levels in CR2 (and centC) compared to the unique centromere mean methylation [46]. They suggest that variability in centromeric repeat methylation levels among the three methylation contexts, is partially to blame for the previously observed hypomethylation in active maize centromeres (i.e. poor 5mC immunostaining). The increase in observed methylation near centromeres may be due to the increased concentration of retrotransposons and tandem repeats in the active centromere and pericentromeric

regions which can be regulated by siRNA directed DNA methylation [169, 170]. These tandem repeats and transposable elements represent most of the sequence found in centromeric regions.

SB3 A primer on transposable elements

SB3.1 LTR transposable elements

In addition to tandem repeats centromeres are rich in transposable elements (TEs), mobile elements capable of changing position in the genome [92]. Originally discovered by Barbara McClintock through her detailed cytological studies of the *Ds/Ac* loci in maize and their effects on the pericarp of maize [166], these elements are highly diverse and vary greatly in length, sequence composition and transposition mechanism. They are also abundant in grasses and make up an estimated ~80% of the maize genome [196]. TEs are subdivided into two major groups based on their transposition intermediate, Class 1 (RNA elements) and Class 2 (DNA elements).

Class 1 TEs (RNA elements) may be further subdivided into LTR (Long Terminal Repeat) and Non-LTR retrotransposons. Studies in the genus *Oryza* have shown LTR retrotransposons to be the most frequent class of TEs [197]. LTR retrotransposons have five primary functional domains including a group specific antigen (GAG), protease (PRT), reverse transcriptase/RNAse-H (RT/RNAse-H) and an endonuclease/integrase (INT) [196, 198]. LTR retrotransposons can be subdivided into two major families Ty3-*Gypsy* and Ty1-*Copia*. The most prominent difference between the two groups is the position of the integrase. Ty1-*Copia* elements have the integrase domain upstream of the RT domain [199]. The Ty3-*Gypsy* elements contain the integrase downstream of the RT domain. Studies in *Oryza* have shown the Ty3-*Gypsy* type retrotransposons generally have greater representation in the genome, and Ty1-*Copia* tend to be more evolutionarily diverse and smaller in size [89, 197].

The transcribed region of the retrotransposon is often called the polyprotein region (PPT) which is flanked on the 5' side by a repetitive untranslated region (UTR). The UTR and PPT is nested between two terminal LTRs which are themselves nested between target site duplications (TSD). TSDs represent the insertion site of the retrotransposon and the TSD/LTR junctions act as a unique signature for each element [24, 200]. The LTRs are identical upon initial insertion into the genome and accumulate SNPs over time. As such insertion date can be inferred from genetic distance between the LTRs with a suitable conversion factor. Furthermore, each insertion forms two unique markers in the form of TSD/LTR junctions which are useful for determining shared lineage.

SB3.2 An overview of Class 1 TE transposition

Transposition of TEs is an intricate biochemical process by which a piece of integrated TE genomic DNA mobilizes in a host genome. Studies of the Class 1 barley retroelement BARE1 suggests transcription is driven by a Pol II promoter in the 5' LTR. Transcription of BARE1 terminates inside of the 3' LTR, before the terminus of the element [201]. A transcribed RNA intermediate then travels to the cytoplasm where it is translated. The GAG proteins encapsulate a copy of the RNA, and the translated PRT, RT, RNAse-H and INT into a virus like particle (VLP) [202]. Within the VLP the polyprotein is cleaved into functional retrotransposon proteins by the PRT [203]. Within the VLP the RNA intermediate self-primes, and is reverse transcribed into cDNA, the RNA intermediate is simultaneously degraded by the RNAse-H [204]. The VLP must then return to and enter the nucleus. The precise mechanism by which this is accomplished is poorly understood, though likely there is a nuclear localization signal (NLS) that mediates the process, as is the case in the maize Grande retrotransposon [205]. The highly conserved INT then cuts and joins the retrotransposon cDNA to the genomic DNA forming the TSDs and completing the cycle [206]. Interestingly, studies of HIV suggest it is common for multiple RNA molecules to be

packaged in a single VLP and one might infer similar processes may occur in retroelement VLPs [207]. This would potentially allow transpositional recombination between elements to take place.

SB3.3 Non-LTR transposable elements

Non-LTR Class 1 TEs are structurally characterized by a 5' and 3' UTRs with two ORFs including a *gag* and *pol* flanked nested between TSDs additionally non-LTR retros often have a poly(A) stretch between the 3' UTR and terminal TSD [198, 200]. And include such families as the autonomous, long interspersed nuclear elements (LINEs) and non-autonomous, short interspersed nuclear elements (SINEs). These families comprise the majority of the non-LTR retrotransposons [208]. These elements are ubiquitous among plant genomes and are proposed as the phylogenetic progenitors of the LTR, Ty3-*Gypsy* and Ty1-*Copia* retrotransposons [209, 210]. Interestingly LINE-like sequences do not randomly distribute thought the genome but were found to instead form clusters interspersed across the genome observable with FISH [211, 212]. While not centromere targeting, these elements can be found in active centromeres.

SB3.4 Class 2 Transposable elements

Class 2 transposons (DNA elements) are characterized by a single open reading frame (ORF) containing a transposase flanked by terminal inverted repeats (TIR), which are nested between two short direct repeats (DR) [198, 200]. The TIR sequences are usually quite short ~10-40 bp but can be up to ~200 bp [213]. Class 2 elements transpose with a cut-and-paste mechanism, wherein the transposase binds with sequence specificity to the parent element, and any non-autonomous family members which effectively allows for transposition of both autonomous and non-autonomous elements. The elements are then excised from the genome and reinsert in a different location [213]. This mechanism is non-replicative, therefore a Class 2 TE must rely on host machinery to indirectly increase copy number [214]. Class 2

transposons are not known to be enriched in maize centromeres but were the first TEs to be discovered by Barbara McClintock (Ac/Ds system) [92, 93, 166].

SB3.5 MITES, Mutators and Helitrons

Miniature inverted-repeat transposable elements (MITEs) are a type of short (100-600bp) non-autonomous Class 2 transposon that are the result of internal deletions in autonomous DNA elements like the Ds element [214, 215]. As such MITEs generally have both TIRs and TSDs but lack the transposase ORF found in autonomous elements. Generally speaking, MITEs are randomly distributed on chromosome arms and those that are associated with genes and are frequently transcribed with genes [216]. Additionally, MITEs associated with genes have been reported to significantly lower gene expression [215]. MITE density tends to decrease at the centromere and pericentromeric regions possibly due to the heterochromatic nature of the region [215].

Mutator (*Mu*) elements are Class 2 TEs that all share conserved TIR sequences approximately 220nt in length, but have differing internal sequences [217]. *Mu*-like elements (MULEs) are common among flowering plants and have multiple functional variants. The *Mu* element is highly active and mutagenic preferentially inserting into low copy number DNA [218]. This property makes it particularly likely to alter a genes behavior through insertion in introns or promoter sequences. Indeed, it's mutagenic properties has led to the creation of the engineered *Mu* element *RescueMu* for the purpose of studying *Mu* element behavior, and for high throughput mutation and cloning [219].

Helitrons are Class 2 elements characterized by a transposition mechanism related to rolling circle replication [220]. The majority of helitrons are non-autonomous and differ in structure from the majority of Class 2 TEs as they lack TIRs and instead possess conserved terminal motifs [220]. The helitron transposition mechanism targets 5'A and 3'T sites and results in the insertion of helitrons without TSDs [220]. Helitrons are the only known TEs that do not produce an insertion scar on the host

DNA (TSD or other insertion mark). Helitrons have a unique structure among TEs having a Rolling Circle (RC), Replication initiator (Rep) and DNA Helicase (Hel) [221]. Remarkably, studies have shown that helitrons can capture genic fragments such as introns or exons [220].

SB3.6 Transposable element regulation

TEs are often considered "parasitic" elements. Presumably, increases in copy number decreases the fitness of the host. As TEs have the capability of altering expression, it is expected that certain mechanisms exist to suppress the mobility or amplification of TEs. Analysis of yeast deletion mutants showed that >100 host genes may influence activity of the Ty1 retrotransposon, indicating close interplay between the host and retrotransposon transposition [167]. In humans it has been shown that siRNAs are generated from TEs which subsequently target those same TEs [167, 168]. Analysis of siRNAs in *Arabidopsis thaliana* indicated ~27% of the siRNAs investigated were derived from TEs, with many siRNAs originating from the centromere or pericentromere [169]. The siRNAs generated from the elements also have the potential to modify chromatin to further regulate TE transcription through siRNA triggered DNA methylation [170]. There is currently no information regarding the regulation of centromeric retrotransposons in the literature, though they are likely regulated by these same mechanisms. Deregulation of TEs can potentially lead to a rapid increase in TE copy number called a transposon burst.

Supplemental Text

SX1. Helitron related gene graveyards

Analysis of the alignment between B73, Mo17, NKH8431 and LH145 uncovered a 37,873nt region unique to B73 (v4 coordinates 105,091,738 – 105,129,611). The position of the region in CEN5L of B73 was confirmed by the independent BAC by BAC assembly RefGen_v2. Junction viewer images revealed the region to be comprised of genic sequence that contained no LTR retrotransposon insertions. Blast of the region to eight physical maps (CML247, F7, EP1, W22, PH207, Mo17, PE0075, DK105) indicated the region was present in Mo17, F7 and PE0075 on chromosomes 3, 3 and 7 respectively. It is remarkable that Mo17 differs from B73 at the CEN5L locus given the divergence of the two lines is less than 10Ka. These data indicate that this region is mobile and likely inserted into B73. The fact that this region was not found as a single fragment in the other five genomes suggests that either the region can be lost without consequence, or it was heavily disrupted by retrotransposon insertions in the other genomes.

An alignment of the region was constructed from the four genomes and the region was found to be highly colinear across much of the region. However, there were two regions in the alignment where B73 differs from the other three genomes. Blast of the first region of poor homology (333bp) indicated the B73 region was homologous to a fragment of a *Zea mays* transcriptional activator gene (L19495.2). Blast of the second region (520bp) NCBI indicated homology to an Ankyrin repeat. The corresponding regions from Mo17, F7 and PE0075 show homology to an alpha/beta-Hydrolase (XM_008653148). Additionally, a 2095 bp sequence was found in B73 that was duplicated in Mo17 and triplicated in F7 which has homology to argonaute 18a and clathrin heavy chain 1 (XM_020541814.1 and XM_020542866.2).

The region in B73 contains seven gene models Zm00001d015671- Zm00001d015677 (v4 annotation) four of which were truncated. These gene models were analyzed using the Gramene database for paralogs and orthologs to determine the origin of the region. Five of the gene models had a paralog and

three of the paralogs had syntenic orthologs. However, the syntenic orthologs in *Sorghum bicolor* were from different chromosomes indicating this region is not derived from a group of linked genes that travel together, but rather an amalgam of genic fragments.

Zm00001d015673 is highly repetitive and had 391 paralogs in the maize genome. Conserved domain analysis indicated that it contained a PIF1 helicase and a helitron like domain (N-terminus). Gene trees from the Gramene database also indicate this repeat is present in *Setaria Italica*. Zm00001d015674 has 29 maize homologs, no orthologs and no domains.

Gene models in the gene graveyard do not have any H2A.Z associated with them. And only gene models Zm00001d015671, Zm00001d015675 and Zm00001d015676 have B73 RNAseq associated with them though all 27 lines display some expression with defined splicing (*SI Appendix Figure S7*). There are differences in the expression patterns for both Zm00001d015671 and Zm00001d015676. The expression differences in Zm00001d015671 are likely due to the Tophat mapping parameters which required uniquely mapped reads. SNPs allowed reads to map uniquely increasing read coverage. This is especially evident at 105,105,400 in B73 (T-C). CML333 and CML69 have expression covering a larger region of Zm00001d015676. One could speculate this is due to loss or damage to a paralog allowing reads to map uniquely to the graveyard region.

To investigate the recent history of the gene graveyard an alignment was generated from sites of the gene graveyards in B73, F7, Mo17, PE0075 and the sites of insertion or excision in Mo17 CHR05, B73 CHR07 and B73 CHR03 (*SI Appendix Figure S8*). This alignment revealed the locus at chromosome 3 is flanked by AT microsatellite repeat. Investigation of the equivalent region on B73 revealed heavy truncation of the repeat. Recombination over the microsatellite likely led to the excision of the gene graveyard from chromosome 3 in B73 which was likely incorporated into the CEN5L locus due to double strand break and subsequent repair.

We identified a second gene graveyard in CEN9 in the PHS1 introgression upstream of CEN9. The graveyard contains five gene models (Zm00001d045996 - Zm00001d046000), which are either truncated or have >5 paralogs. Blast of this region to B73_v4CEN uncovered a complete copy of the CEN9 graveyard on CHR01 and partial copies on CHR03, CHR05 and CHR06. This was unlike the CEN5 gene graveyard which did not have any full-length duplicates. A Mauve alignment was generated from the copies in CEN9 (B73, DK105), CHR01 (B73, CML247, F7), CHR03 (B73, EP1, PH207), CHR05 (B73) and CHR06 (Mo17, PE0075). The B73 CHR03 copy appears to be duplicated, but the duplicated region is split by a gap of 100 N's indicating a possible chimera. The CEN9 graveyard is between 53,101,789 and 53,118,468 and is 16,680nt. The graveyard contains three gene models (Zm00001d045997 - Zm00001d046000), which are either truncated or have >5 paralogs none of which are associated with H2A.Z.

The CEN9 graveyard was searched for conserved domains, and it was found that the region contained DEAD-like helicase and Helitron like N domains (RF+1 and +3). Conserved domain searches of the copies indicate the DEAD-like helicase (PIF1-like helicase) and Helitron like N domains are present in all copies. Additionally, there appears to be two possible palindromic regions. The larger of the two is 404nt and is found in CHR05, CHR01 and CEN9. The shorter one (152nt), is found in CHR03 and CHR06 but is also present in CHR01 and CEN9. The CEN5 graveyard also contains DEAD-like helicase (PIF1-like helicase) and Helitron like N domains (both are RF-2). However, the CEN5 sequence does not contain the hairpin palindrome. The region corresponding to the CEN5 gene graveyard on Mo17 CHR03 with an additional 5kb flanking either side was extracted and blasted to itself to look for a palindromic region in the ancestral location. This blast did not uncover any palindromic sequence in the ancestral CHR03 locus. We blasted the flanking 1kb regions of the B73 CEN9 copy to PH207 to find the exact insertion site but since the CEN9 graveyard inserted into the distantly related PHS1 introgression, the exact insertion site was unable to be located.

Most Helitrons start with a 5' T/C and end with CTRR (Most often CTAG) additionally they preferentially insert into AT dinucleotides (for additional details regarding helitrons and other TES see supplemental background SB3). These motifs are found at 12,680 and 29,342 (original base coords, AC dinucleotide), both within 15nt of the estimated start and stop of the CEN9 gene graveyard. In CEN5 the first TC is found at 40,244 and CTGA is found at 400. Though admittedly these are very common motifs and occur quite frequently throughout the sequence [2]. Additionally, the sizes of both CEN9 and CEN5 regions are near the reported size estimates for helitrons of 2-30kb [1].

This information suggests the gene graveyards in CEN9 and CEN5 are helitron related. The CEN5 gene graveyard contains helitron related domains but lacks the palindromic hairpin which is a key identifier of a both autonomous and non-autonomous helitrons (except for a rolling circle transposon lacking a hairpin found in Aspergillus [3]). The fact that there are no additional copies of the CEN5 gene graveyard in B73, coupled with the lack of the hairpin structure (Presumed to be involved in determining the termination site in rolling circle transposition [2,4]), suggest that the CEN5 graveyard likely moved via recombination over the AT microsatellite rather than rolling circle transposition.

SX2. Gene damage can result in persistent H2A.Z loci

CEN4

There is a syntenic region in CEN4 that contains an H2A.Z peak that is also associated with RNAseq. This peak (CEN4-h) has no gene model in *Zea mays,* but the peak has 2 HSPs matching a Sorghum gene model. Alignments of the Sorghum gene model to CEN4-h showed differences in the CDS that may lead to the Zm gene no longer being functional. To verify that the Zm peak is not a functional gene we examined both the Sb gene model and the CEN4-h peak using the NCBI conserved domain search [222] and found a frame shift mutation between Zinc finger domains that would render this gene non-functional in *Zea mays*. This is a different situation than CEN9. In the CEN9 region we saw separation of

H2A.Z from the parent gene. In CEN4 we find persistent H2A.Z and RNAseq after gene damage, but there is no separation of H2A.Z from the gene corpse (*SI Appendix Figure S9a*).

CEN9

In CEN9 we find a syntenic group of 4 HSPs in *Zea mays* matching v4 gene models Zm00001d046014 and Zm00001d046016 and their orthologs in *Sorghum bicolor* SORBI_3010G146000 and SORBI_3010G145900 respectively. In Sorghum HSP1 corresponds to the 5' promoter region of the gene, where one would expect H2A.Z to localize. However, in *Zea mays* HSP1 has moved ~84kb away from the parent gene. Investigation of the sequence between HSP1 and the parent gene in *Zea mays* uncovered four retrotransposon insertions, the oldest of which inserted ~216ka. These insertions contributed to the separation of the solo H2A.Z locus (*SI Appendix Figure S9b*).

CEN10

CEN10 contains evidence of a persistent 3' H2A.Z locus being retained in *Zea mays*. We also find evidence of an H2A.Z locus in *Zea mays* that has no hit to any gene model in Sorghum bicolor and the Sorghum bicolor gene model was lost from *Zea mays*. The situation at CEN10-n is a clear example of an H2A.Z locus that survived the destruction of the parent gene (*SI Appendix Figure S9c*).

SX3. New substitution rates calculated from six genomes

Homologous regions from six sequenced genomes (B73, CML247, EP1, F7, Ph207 and W22) were identified using MUMmer, extracted and aligned using Mauve to identify regions present in all six inbreds [147, 223]. The pairwise distances calculated between B73, and the other five lines helped to precisely identify the conserved regions in the new reference assembly. The alignments also indicate that previously reported deletions represent highly divergent regions [24].

The artificially constructed sequences using the portions of the mauve alignment from the six sequenced genomes of the single haplotype centromeres of chromosomes 1, 4 and 6 had maximum K2P distance values that ranged from 0.00143605 - 0.00149082 (*SI Appendix Table ST9*) [24]. Assuming, these regions diverged near domestication ~10 thousand years ago (ka), the substitution rates of these three centromeres ranged from 7.2 – 7.5 $*10^{-8}$. The weighted average nucleotide substitution rate of $7.3*10^{-8}$ was used for centromere divergence dating. This is a faster rate than that calculated for the region upstream of *tb1* ($5.5 * 10^{-8}$) and the weighted average rate calculated for the domestication locus on the long arm of chromosome 10 ($6.4 * 10^{-8}$), which is in the location of a previously active centromere marked by conserved single copy sequences in the pericentromere (CSCP) of chromosome 6 [224]. The centromere substitution rate derived here also places previously characterized post-domestication haplotypes from the other centromeres with divergence dates <10 ka [24].

SX4. Comparison of sequence divergence between physical maps and HM3 SNP data

To examine the divergence of maize lines for which we have no physical map we compared the divergence between physical maps and HapMap3 SNP data over equivalent regions (*SI Appendix Table ST8*). Single haplotype centromeres among the NAM lines (CEN1, 4 and 6) were used to calibrate the HapMap3 conversion factor as they are all estimated to have diverged at domestication (~10ka). Each of the CENs 1, 4, and 6 contained a total of 325 pairwise divergence comparisons between HM3 and the physical map (*SI Appendix Figure S14 a-d*). The data are roughly linear, so a linear trend line was added along with the R^2 values and estimated slope. The best correlation was found in CEN4 (R^2=0.94), followed by CEN6 (R^2=0.83) and then CEN1 (R^2=0.51). Occasional outliers are observed and represent misalignments in the mauve or artifacts in HapMap3. Analysis of Region13_SYN of CEN2 within HAP2 and HAP1 (estimated to have diverged ~10.3 ka and 5.8ka respectively) display variation in the linear trends. While the slope of the HAP1 lines is like what we observe with the single haplotype CENs, the

slope of the HAP2 lines is considerably lower. This difference is likely caused by the fact that HapMap3 data uses a HAP1 line as a reference and thus is missing variation shared among the HAP2 lines alone.

SX5. Deletions in NC358/Oh43 result in HapMap3 artifacts

The CEN2 domestication region is defined on the upstream side by a recombination breakpoint shared by NC358 and Oh43. Initially this breakpoint was suggested by SNP density charts generated from HM3 data. However, it manifested as narrow peaks of high SNP density instead of plateaus of increased SNP density which are observed with most recombination breakpoints (*SI Appendix Figure S16*). The release of the NAM line physical maps allowed us to confirm the presence of the NC358/Oh43 breakpoint and revealed two indels (0.18Mb and 1.02Mb) relative to the other 24 NAM lines upstream of CEN2 domestication (*SI Appendix Figure S17, Table ST10*). Analysis of the region in B73 to the NC358 and Oh43 physical maps using MUMmer (word length 200) indicate both are deletions in NC358/Oh43. These deletions suggest that the SNPs in HM3 over these regions are mis-mapped and NC358/Oh43 are not accurately represented in these regions.

SX6. The CEN2 HAP2B inversion breakpoints are correlated with deactivated centromeres

Investigation of genes across CEN2 indicated that HAP2B inversion breakpoints occurred in regions of relatively low gene density. Analysis of the syntenic relationships between B73 v5 gene models and the Sorghum bicolor physical map (NCBI_v3, GCF_000003195.3) suggest the regions of low gene density are proximal to deactivated Sorghum centromeres (*SI Appendix Figure S18*). The upstream inversion breakpoint is located within a cluster of pericentromeric markers for *Sorghum bicolor* chromosome 6 which is also proximal to the currently active B73 CEN2. the downstream breakpoint is found near a

group of *Sorghum bicolor* chromosome 5 pericentromeric markers and is also proximal to the junction between Sb6 and Sb5 syntenic regions on Zm2.

SX7. The progenitor of the CEN2 domestication region

To determine if CEN2 domestication is derived from TIL10 or TIL14 we initially mapped TIL10 and 14 GSS data (Bowtie -v1 and -v2, -m1, Trimmed and untrimmed) to a representative of HAP1 (B73), HAP2A (Ky21), HAP2B (P39) and HAP2_BASAL (NC358). Visual inspection of mapped reads in IGV did not indicate whether TIL14 or TIL10 was the progenitor. We then turned to HM3 data. The CEN2 domestication region was defined on the upstream side by the NC358/Oh43 recombination and on the downstream side by the HAP1/HAP2 recombination and is 164,525nt (B73 v4 chr2:94064285-94228810). The coordinates were verified by alignments between B73 RefGen_v4 and NC358 physical maps. We extracted HM3 SNPs from the closest HM3 positions internal to the CEN2 domestication region (B73 v4 chr2:94064331-94228806). To ensure that ambiguities in the SNP data did not affect the tree structure, trees were generated from the HM3 data in the CEN2 domestication region both with and without ambiguous sites (*SI Appendix Figure S21*, pairwise deletion, bootstrap 1000). Evaluations of the trees indicate neither TIL10 or TIL14 are basal to the domesticated maize group in the CEN2 domestication region. TIL10 groups with NC358 and Oh43, and TIL14 groups with BKN030 and BKN033. Analysis of the 3,621 HM3 SNPs in the CEN2 domestication region indicate 848 positions are uninformative and 11 positions have >40 ambiguous calls over the position and were excluded from further analysis (*SI Appendix Table ST11*). A total of 4 SNPs were found to be shared between TIL10 and NC358/Oh43. The four SNPs are distributed across the CEN2 domestication region suggesting the SNPs are not a small introgression. A total of 9 SNPs were found to be shared between TIL14, BKN030 and BKN033. However, only 1 of these SNPs was non-ambiguous in all 3 lines. Interestingly 7 of the 9 SNPs shared by TIL14, BKN030 and BKN033 are also shared with the HAP2B group (both inverted and non-

inverted). This suggests that TIL14 was the progenitor the HAP2B group. We cannot be entirely certain that TIL14 was the progenitor of CEN2 domestication. Though two lines of evidence suggest that TIL14 was the progenitor. First, TIL10 shares SNPs with modern lines suggesting the CEN2 domestication region in TIL10 may have escaped from a NC358/Oh43 relative into TIL10. Second TIL14 is basal to HAP2 across the non-recombinant region, and TIL14 is within the domesticated maize clade in both the CEN2 domestication region. Further TIL14 is in the domesticated maize clade upstream of the CEN2 domestication region (Only NC358/Oh43 recombine). TIL10 on the other hand contains a recombination breakpoint upstream of CEN2. These data would suggest that TIL14 was the progenitor of the CEN2 domestication region and CEN2 while TIL10 acquired the CEN2 domestication region from a line related to NC358/Oh43.

SX8. A distantly related pericentromeric introgression in CEN9

Approximately 294 kb upstream of the B73 active centromere region rests a region where five non-B73 genomes share no alignment markers greater than 3 kb in length (Mauve alignment). The pericentromeric introgressed region spans from 52,788,188 to 53,366,330 (RefGen_v4CEN) totaling approximately 578kb [133]. The region is comprised primarily of various retrotransposons and repeats but contains a single expressed gene (PHS1, Zm00001d045993 RefGen_v4 annotation). Phylogenetic analysis over PHS1 revealed three haplotypes (*SI Appendix Figure S24, Text SX9*). A total of 33 retrotransposons were identified and comprise approximately 389kb or 67% of the sequence in the B73 introgressed region. 31 of these retrotransposons are unique to B73 with insertion dates ranging from 13.9 to 505.8 ka. One RLG doke (RLG_doke_AC197224-5479) retrotransposon was found to be shared between the six genomes. The LTRs of the shared retro were incomplete, did not have TSDs and spanned approximately 73 kb. The shared retro was estimated to have inserted ~619ka. From these data, we estimate the divergence of the introgressed region to be between 505.8 and 619.5 ka. The

PHS1 introgression upstream of CEN9 is evidence that recombination can occur from distant relatives near an active centromere. Further, another study of CR elements has shown evidence of numerous putative horizontal transfers of CR elements between distant oryzoid and panicoid lineages [111].

SX9. CEN9 proximal introgression contains PHS1

The introgressed gene (v4 annotation Zm00001d045993) is located from 52,750,697 to 53,365,663 in B73 RefGen_v4CEN. Analysis of combined RNA-seq data from four different tissues mapped to the B73 v4CEN genome and visualized in IGV determined this gene to be expressed. Qteller expression over the gene indicate low expression in almost all tissues [225]. Investigation of the gene data available on the Gramene database revealed it to have no paralogs in *Zea mays* and orthologs in *Oryza sativa*, *Setaria italica*, *Sorghum bicolor* and the *Arabidopsis thaliana* Poor Homologous Synapsis 1 gene (PHS1). PHS1 has previously been described by Ronceret *et. al.* 2009 as being a regulator of meiotic recombination and acts in the coordination of homologous chromosome pairing during cell division [226].

Evaluation of retrotransposon insertions in PHS1 among six physical maps (B73, EP1, PH207, CML247, F7 and W22) indicate three distinct lineage specific retrotransposon insertions (*SI Appendix Figure S25*). The oldest of which was a Pute retrotransposon that inserted into B73 (414.9ka). The EP1 and CML247/F7/W22 lineages contain a Jare (20ka) and a Bosohe retrotransposon (0ka) respectively.

To determine the phylogeny of the PHS1 homolog the coding regions for this gene were extracted from RNA-seq data for 26 maize lines and 9 TIL lines. Alignments between these sequences and the CDS from the 5 non-B73 sequenced lines reveal there are three haplotypes for this gene, the B73 haplotype, The W22 haplotype and the CML69 haplotype (*SI Appendix Figure S24*). To establish a divergence date for the PHS1 homolog between B73 and the 5 sequenced lines, K2p distances were calculated from synonymous substitutions over the CDS alignment (*SI Appendix Figures S23 and S24*). Dates were calculated using the genic substitution rate of 7.9*10^-9 proposed by Gaut *et. al.* 1991 [227]. The B73

haplotype was determined to have diverged from the other five non-B73 genomes approximately 914.8ka though this date is considerably higher than the estimated divergence of the surrounding sequence and may be due to the use of different conversion factors.

To further analyze the gene, SNP analysis of the coding regions was conducted uncovering 15 SNPs. Alignments of the maize sequences with the ancestral *Sorghum bicolor* ortholog revealed 8 of the SNPs occurred in the B73 haplotype, 4 SNPs occurred in the W22 haplotype, 2 occurred in the CML69 haplotype and one non-ancestral SNP was acquired by both the B73 and CML69 haplotype (*SI Appendix Table ST15*). Eight SNPs were found to be non-synonymous with two coding for amino acids of different chemical properties (from S to F and P to H, respectively, in the B73 haplotype).

SX10. Independent CTR1 formation in CEN9 and CEN2

Upstream of the CentC in CEN9 are two loci (LI and LII) containing CR1 derived tandem repeats (CTR1) previously described in Sharma *et. al.* 2013 [84]. B73 v4CEN contains the completely assembled CTR1 loci. This permits the study of the organization of each locus, the relationships between loci and allows comparison the loci among the 6 sequenced genomes. LI and LII are found from 55,377,887 to 55,562,535 and 56,998,082 to 57,124,172, respectively in the B73 v4CEN genome. LI was found to contain 114, 115, 116, 124 and 71 total monomers for B73, W22, CML247, F7, EP1 and PH207 respectively. LI contained nine retrotransposons with insertion dates ranging from 32kya to 339kya (*SI Appendix Figure S26*). LII contained 103, 102, 110, 97, 98 and 115 total monomers for B73, W22, CML247, F7, EP1 and PH207 respectively. Four retrotransposons were dated in B73 with insertion dates ranging from 0 to 187kya. All monomers in both loci are in the reverse orientation in B73 v4CEN.

Line specific local duplications and deletions of monomers are common occurrences in both CTR1 loci. Five, line specific local duplications were found in LI and 4 in LII. LI was found to have five, line specific deletions and LII contained 12. Additionally, there is evidence of two instances of deletion with

recombinant element formation in LII. First, over CLII_50 and CLII_51 to form the recombinant type of BLII_44 and second over CLII_22 and CLII_23 to form the recombinant type BLII_18. These evolutionary mechanisms are like those described by Wolfgruber *et. al*. 2016 [61] for the evolution of CentC. However, it is important to note that the CTR1 loci in the 5 non-B73 genomes contain gaps and the possibility monomer homogenization during assembly may cloud analysis.

In CEN2 CTR1 loci can be found in 3 physical maps (Oh43, CML277 and P39) though it is present in the scaffolds of four additional lines (CML322, Il14H, NC358 and Oh7B). The CEN2 CTR1 loci was investigated in P39 and CML277 which contain 4 and 7 monomers respectively. Monomers are 1,577-1,582nt long and are comprised of CR1 LTR/UTR sequence. All monomers except CML277 Monomer 7 and P39 Monomer 1 are derived from a CR1_R5. CML277 Monomer 7 and P39 Monomer 1 are 128.3-177.8ka diverged from the other monomers and are separated in the tree (*SI Appendix Figure S27*). Furthermore, CML277 Monomer 7 and P39 Monomer 1 have CR1_R4-like UTRs and many SNPs that differentiate them from the other monomers suggesting this they are derived from a different CR1. If CML277 Monomer 7 and P39 Monomer 1 are derived from a different CR1 and are excluded, the tree suggests that this locus is derived from P39 Monomer 4 and CML277 Monomer 1. All monomers excluding CML277 Monomer 7 and P39 Monomer 1 are diverged by ~56.8ka. This is far younger than the maximum monomer divergence of the CTR1 in CEN9 estimated to have formed 558.2-573.7ka for LII and LI respectively.

Alignments of the P39 and CML277 region suggest P39 lost the CML277 equivalent monomers 4 and 3 via two deletions over "T" and "A" rich regions respectively (*SI Appendix Figure S28*). This suggests an instance of deletion and recombinant monomer formation, but we were unable to identify SNPs in the alignment to positively confirm this. The deletion coordinate based on a muscle alignment in P39 and deletion junctions are found in *SI Appendix Table ST16*.

Analysis of a tree containing monomers from both the CEN9 and CEN2 loci indicate that CEN9 CTR1 monomers are derived primarily from a CR1_B element while the CEN2 locus is derived primarily from a CR1_R5. From these data it is likely that the CEN2 CTR1 formation was independent of the CEN9 CTR1 formation, however both tandem repeats were derived from CR1 LTR/UTR sequence.

SX11. Discrepancies between HapMap3 and the NAM Physical maps

Analysis of K2P divergence for regions in HapMap3 vs matching regions in the NAM physical maps indicate there are discrepancies between the datasets. Generally, the correlation between HapMap3 and Mauve alignments was acceptable (R2 = 0.94 in CEN4 and R2 = 0.82 in CEN6), though the R2 was particularly low in CEN1 (R2 = 0.52). Investigation of the outlier points indicate that Tx303 had an elevated K2p in HapMap3 relative to the Mauve K2P divergence in CEN1. To ensure that the poor correlation was not due to some local misalignment in the mauve alignment we extracted five, 500kb subalignments from the CEN1 alignment and generated a K2p comparison for each window (*SI Appendix Table ST17*). The correlations across each of these windows did not improve suggesting that the poor correlation was not a local issue in the alignment. We then ran a MUMmer analysis of the aligned CEN1 region to B73 v5 and B73_v4 physical maps (word length 1000) to look for any repeats or duplications, as well as any potential inversions that may have been missed. The MUMmer analysis indicated no major repetitive regions or substantial duplications that would have affected HM3 mapping. Further we found no evidence of any additional inversions.

Interestingly, visual inspection of the Mauve alignments in Geneious for CENs 1,4 and 6 revealed that there were far more unique insertions in the CEN1 alignment than in CEN6, and CEN4. We calculated how much sequence in the alignments was not present in B73 for CENs 1, 4 and 6 and plotted them against the R^2 and found that there is a linear correlation between the amount of sequence missing from B73 and the R^2 (*SI Appendix Figure S29*). It is possible that the correlation is so low in CEN1 because

a large amount of unique sequence from the other lines is unable to be mapped with the B73 reference leading to poor correlation. Essentially the more unique sequence that is absent in the reference for a given region, the poorer the correlation between physical map and HM3 divergence.

Given the maximum maize-maize distances in the HM3 data are roughly identical to those in the Mauve data, we suspected that the poor correlation could be the result of variable ordering at the clades level. To investigate this, phylogenetic trees were generated for HapMap3 and Mauve data over the regions from CENs 1, 4, and 6 and then compared (*SI Appendix Figure S30*). In general, the trees matched though we observed discrepancies in the position of CML228 in CEN6 and, Tx303 and B97 in CEN1 between HapMap3 and the physical maps. Investigation of the position of CML228 in 500kb (physical map region size) HapMap3 trees flanking the CEN6 region indicate CML228 does not change position in the flanking trees suggesting it is not a local HapMap3 issue (*SI Appendix Figure S31*). Furthermore, trees generated from the same HapMap3 regions using unimputed data over the CEN6 alignment region show CML228 still groups with Ki11 in the non-imputed data suggesting this issue is not due to imputation (*SI Appendix Figure S32*). While the root cause of the discrepancies and poor correlation in CEN1 for HapMap3 and Mauve data are not known, taken together these data suggest that they may be the result of the use of B73 as a reference in HapMap3 or due to differences in the germplasm used for HapMap3 and physical map assembly.

SX12. CentC loss during inbreeding

Reads from six maize cultivars that underwent six successive generations of selfing (PRJNA483084) were aligned to a centC dimer (blastn -word_size 7 -evalue 0.001 -max_hsps 1 -perc_identity 85 -dust no) to determine the read and sequence fraction derived from centC in each generation. There was significant missing data as not all generations, cultivars and replicates were sequenced (*SI Appendix Table ST24*). Using the available data, we calculated the centC percentage for each of the 6 cultivars across

inbreeding generations and found no significant reduction in overall centC levels (SI Appendix *Figure S43*). This may be due to a variety of factors including centromere position (it was assumed the centromeres were in the centC locus, but this may not be the case), level of inbreeding (perhaps further successive generations were required to observe a centC loss), or general lack of data. The authors of the study did however observe an overall decrease in genome size because of purging during inbreeding which the authors note may be due to the loss of genes, variation in repetitive DNA copy number, fluctuations in repeat monomer counts (including centC), and the loss of supernumerary B chromosomes [228]. A more targeted experiment that evaluates centromere position, size and repeat content would be necessary to further evaluate centC loss specifically during inbreeding.

SX13. H2A.Z in CR2s

To evaluate the relationship between H2A.Z and CR2 I remapped the B73 H2A.Z data to the B73_v5 physical map allowing multimapping to account for CR2 LTRs (Bowtie, X 300, m 2). I then ran MACS2 peak calling on the resulting BAM file (default parameters). Using the MACS2 peak output file I compared the peak coordinates to the coordinates of CR2 elements I had identified in B73_v5 looking for any coordinate overlap. I identified 42 peaks that overlapped with CR2 elements. I then checked this CR2 subset to see if the elements were in the active centromere, inserted post domestication, were uninterrupted and whether the elements were associated with the previously identified unmethylated regions.

Initially we focused on CR2 elements overlapping an H2A.Z peak, within the active centromere that inserted post-domestication, and were uninterrupted. Only 3 elements in B73 fit these criteria, one in CEN1 and two in CEN8. Investigation of H2A.Z and cenH3 over these elements with 2kb flanking sequence indicated that the H2A.Z peaks were much lower than the H2A.Z control peak located upstream of CEN5. Furthermore, the H2A.Z peaks appear to only overlap the LTR junction and do not

extend into the LTR. We expanded our investigations to include elements that were associated with H2A.Z peaks and inserted post domestication but were not located in the active centromeres. Two peaks fit these criteria including one on chromosome 1 and another on chromosome 4. The peak on chromosome 1 was ~35Mb downstream of the active centromere but had substantial coverage, though this peak was overlapping the junction of the 3' CR2 LTR. The peak on chromosome 4 was in the 5' LTR but was at a coverage level like those in the active centromere. We also investigated 5 additional peaks that were present in the active centromere and uninterrupted regardless of insertion date and found that H2A.Z peaks were in the LTRs but again H2A.Z coverage was very low.

In general, we don't see significant H2A.Z peaks in the 5'LTR of post domestication uninterrupted CR2s within or outside the active CENs. However, we see low level H2A.Z coverage over uninterrupted, pre-domestication CR2 5' LTRs in the active centromere. This suggests that H2A.Z does indeed localize to CR2 LTRs (both 5' and 3') but it isn't observable in post domestication elements with our current multimapping parameters. This is further supported by H2A.Z data mapped to consensus sequences for CR2 subtypes which show H2A.Z primarily associates with LTRs in the U3 region. The lack of observable peaks in post domestication CR2s is likely due to the limited variation between post-domestication LTRs. I allowed up to a maximum of two positions to be mapped per read and as post domestication CR2 subtypes can have tens or even >100 copies I am unable to map H2A.Z specifically to post domestication elements. It is also notable that investigation of the maximum H2A.Z coverage over CR2 subtype consensus sequences to the maximum coverage of an H2A.Z control region upstream of CEN5 suggest that only a subset of CR2s are likely to contain H2A.Z.

It is notable that the maximum UMR coverage found in CR2s falls between two H2A.Z peaks when mapped to consensus CR2 subtypes. This suggests that the CR2 UMR peaks may be associated with the H2A.Z. Previous comparisons of H2A.Z and DNA methylation have shown and anticorrelation between H2A.Z and DNA methylation in various systems (reviewed, The histone variant H2A.Z in gene regulation).

Though after comparison of elements with H2A.Z to elements with UMRs show no significant correlation. However, the disconnect may again be a result of the inability to map H2A.Z to specific post domestication CR2s as the majority of CR2s with UMRs were found to be post domestication while most elements with called H2A.Z peaks are >20ka and are mappable due to higher SNP densities. The disconnect may also be caused by differences in tissue or sampling between the H2A.Z and methylation datasets. One might speculate that those CR2 elements with UMRs also have H2A.Z and are the subset of CR2s open to transcription (and therefore expansion) in the B73 genome.

SX14. TE Insertions of Centromeric and Pericentromeric genes

The centromeric AUG6 gene spans from 108270728 to 108321871 in the v4_CEN genome and contains 13 exons. Evaluation of indels revealed several deletions and local duplications as well as an EP1 specific insertion of an RLC ebel retrotransposon with an insertion date of 0 ka. Introns 3, 8 and 12 were found to be large indicating the likelihood of shared retrotransposons between the six lines. Intron 3 was found to have a shared RLC raider retrotransposon (TSD CAATT) with an estimated insertion date of 150 ka in B73. Intron 12 was found to share a Xilon diguus retrotransposon (TSD TAGGG/TAGAG) with an estimated insertion date of 588 ka. Intron 8 was found to have no detectable retrotransposon insertions. Additionally, further analysis of the Xilon uncovered two additional elements, a Zm00370 Mutator element in the 5' LTR and a partial RIL nugimo in the 3' LTR (*SI Appendix Figure S44b*).

The pericentromeric PHS1 was found to have 10 exons and 9 introns in the upstream pericentromere of CEN9. PHS1 (Poor Homologous Synapse 1) has previously been described by Ronceret *et. al.* 2009 as being a regulator of meiotic recombination and acts in the coordination of homologous chromosome pairing during cell division [226]. B73 was found to have a lineage specific partial RLC pute retrotransposon inserted into intron 9. An 80nt alignment was created from the LTR fragments of this retro and insertion was estimated to be 414.9 ka. EP1 also contains a lineage specific insertion of a RIT

jare LINE transposable element. With a 1.5kb flanking sequence determined to be a zein protein fragment. CML247, F7 and W22 all share an insertion of an RLG bosohe retrotransposon into intron 3 with an insertion date of 0 ka. The five non-B73 sequenced lines also share two additional insertions in intron 3, an internal prem 1 retrotransposon fragment and a non-transposable element insertion containing mitochondrial sequence (*SI Appendix Figure S44a*).

SX15. The Landscape of ZmCEN3, ZmCEN8 and Orthologous OsCEN1 (*Oryza sativa*) and SiCEN5 (*Setaria italica*)

The regions of centromeric conservation in ZmCEN3 and ZmCEN8 in the v4CEN genome is found from 75.6 to 90.2 Mb and 43.2 to 52.9 Mb respectively. Functional centromeres are defined by CenH3 ChIP-seq for ZmCEN3 and ZmCEN8 are from 85.7 to 88.0 Mb and 50.9 to 52.3 Mb in Zm3 and Zm8 respectively. OsCEN1 which is syntenic with ZmCEN3 and ZmCEN8, is between 15.7 and 18.1 Mb and was defined as 1Mb upstream and downstream of the Os1 CenH3 CentO tandem repeat (http://rice.uga.edu/annotation_pseudo_centromeres.shtml) which ranged from 16.7 to 17.1 Mb. SiCEN5 was defined by the 1 Mb upstream and downstream of the most prominent tandem repeat cluster and was determined to be 18.6 to 22.3 Mb.

OsCEN1 and SiCEN5 are syntenic along their length. ZmCEN3 is syntenic with OsCEN1 and a 470 kb region in the downstream pericentromere of OsCEN1. A region just downstream of ZmCEN3 was found to be syntenic with OsCEN12 though this region falls outside of the ZmCEN3 conserved region. ZmCEN8 was found to be syntenic with the Os1 region from 12.2 to 15.4 Mb constituting 3.2 Mb in the upstream pericentromeric region of Os1. A region from 52.6 to 58.0 in the downstream pericentromere of Zm8 was found to correspond to OsCEN1 (*SI Appendix Figure S45*).

A total of 599 genes were found from 12.2 to 18.6 Mb in Os1 which contains all regions of synteny to ZmCEN3, ZmCEN8 and SiCEN5. Analysis of these genes using the Gramene database indicated 340 genes contained no domain and no orthologs. Of the remaining 259 genes, 74 were found to be in OsCEN1. ZmCEN3 and ZmCEN8 contained 37 and 48 genes respectively not including those with no orthologs or domains. RNAseq data was analyzed for root, shoot, leaf and tassel tissues among 27 NAM lines indicate 13 genes and 24 genes are expressed in ZmCEN3 and ZmCEN8 respectively (*SI Appendix Table ST25*).

SX16. Gene Fractionation and Loss

Fractionation, or the gradual loss of duplicate genes following a whole genome duplication is a major factor contributing to decreased synteny [229] (along with genomic rearrangements and deletions [230]). It has been reported that Zea mays is an ancient tetraploid containing a dominant, "A" sub-genome and a recessive, "B" sub-genome [229]. Analysis of the genes in the syntenic regions of Zm3, Zm8 and Os1 revealed that sets of genes are lost in Zm3 and Zm8 but very few Os1 genes were found to be lost from Zm3 and Zm8. Genes were determined to be lost from Os1 if they were found to have a domain and no orthologs in Si or Zm. 10 genes were found to be lost in Os1 six of which had >10 paralogs. This suggests that many of the genes lost were those from larger gene families. Three of the genes lost that had <10 paralogs were predicted to be integral membrane proteins (*SI Appendix Table ST25*). Given the divergence of Si and Zm from Os are between ~50-60 Mya and the divergence of Si from Zm is approximately ~26 Mya one can estimate the rate of gene loss to be between 1.9E-7 and 2.0E-7 genes per year [231-233]. Of the 259 Os1 genes with orthologs or predicted domains 131 genes had orthologs in Zm3 or Zm8. However only 23 Os1 genes had orthologs in both Zm3 and Zm8. The pattern of degradation in the collinearity of genes between Os1, Zm3 and Zm8 is consistent with patterns of fractionation observed in ancient tetraploids (*SI Appendix Figure S45*).

Centromere retention in maize was previously analyzed in Wang et al. 2012 [224] and corrected in the centromeric regions using ChIP-seq data published in Schneider et al. 2016 [24]. Many centromeres were corrected in Schneider et al. 2016 leading to three discrepancies in retained ancestral centromeres between the two reports (*SI Appendix Table ST25*). Previous analyses have shown that one maize sub-genome shows dominance via biased fractionation [229]. This bias is also observed in Zm3 and Zm8, 62 genes were found to fractionate to Zm3 and 46 genes fractionated to Zm8. This discrepancy in the number of genes from each sub-genome indicates the Zm3 syntenic region is likely dominant and from the A sub-genome (*SI Appendix Table ST25*).

SX17. Gene Movement and Acquisition

Genes have been reported to change position and show decreased synteny between related species over larger evolutionary distances. A variety of mechanisms have been shown to relocate genes in the genome including translocations and inversions, illegitimate recombination, or retrotransposon mediated movement [230, 234, 235]. To determine the frequency of gene movement orthologs were evaluated between Os1, Zm3, Zm8 and Si5. Forty-Five Os1 genes from the syntenic region were found to be moved in Si5 away from the syntenic region on Os1 (Os1 genes with Si orthologs on non-syntenic chromosomes) and seven genes moved into the syntenic region. A total of 96 Os1 genes from the syntenic region were found to have moved away from the syntenic regions in Zm while 4 and 11 genes from Zm3 and Zm8 moved into the syntenic region with orthologs on non-syntenic Os chromosomes (*SI Appendix Table ST25*). From these data and the reported divergence estimates for Os, Si, and Zm, one can estimate a rate of gene movement in and out of the syntenic regions to be 2.0E-6 and 2.14E-6 genes per year in Zm (for Zm3 and Zm8 respectively) and 1.04E-6 genes per year in Si [231-233].

Many Zm genes were found to have no orthologs in either Os or Si appearing to be acquired by Zm. A total of 12 and 17 genes were found in ZmCEN3 and ZmCEN8 respectively that had evidence of a domain or GO-term and no orthologs in Os or Si (*SI Appendix Table ST25*). Knowing the divergence of the three genomes allows us to estimate the rate of gene acquisition to be between 2.4E-7 and 3.4E-7 genes per year. Analysis of RNAseq data revealed four of these genes are expressed. These acquired genes contained diverse predicted functions based on predicted domain information. It is important to note that the classifications of genes presented here are heavily dependent on the quality and completeness of the physical maps. As there is no ancestral Os/Zm genome available it is difficult to assess these genes.

SX18. Shared Genes and Retained Centromeric Genes

A total of 17 genes were found to be shared between Os1, Si5, Zm3 and Zm8 in the syntenic regions five of which were determined to be expressed (*SI Appendix Table ST25*). Analysis of GO-Domains obtained from Gramene indicate nine of these genes constitute cellular components, three are involved in biological processes and one has molecular function. Two of these shared genes have no predicted domain and one was found to have a domain of unknown function. One gene had no GO-Terms associated with it but conserved domain searches revealed a BCNT-C domain.

No genes were shared by OsCEN1, ZmCEN3 and ZmCEN8 and only one gene in OsCEN1 was found to be retained in ZmCEN3. Blast searches to the NCBI nr database and conserved domain searches revealed it to be a LNK2 type gene shown in Arabidopsis to be a transcriptional coactivator for the expression of genes involved in the circadian clock [236]. Investigation of RNA-seq data from 27 NAM lines indicate maintenance of the gene. Additionally, this gene has a paralog on Zm8 in the region corresponding to OsCEN1 but the Zm8 paralog is non-centromeric. Two additional genes were found in OsCEN1 with non-

centromeric orthologs on both Zm3 and Zm8. Domain information obtained from the Gramene database and subsequent Blast searches to the NCBI nr database indicate a putative nuclear NAC-domain transcription factor or NAM-Domain family protein, and an RJ4 annexin protein possibly involved in salt stress response.

Supplemental Figures

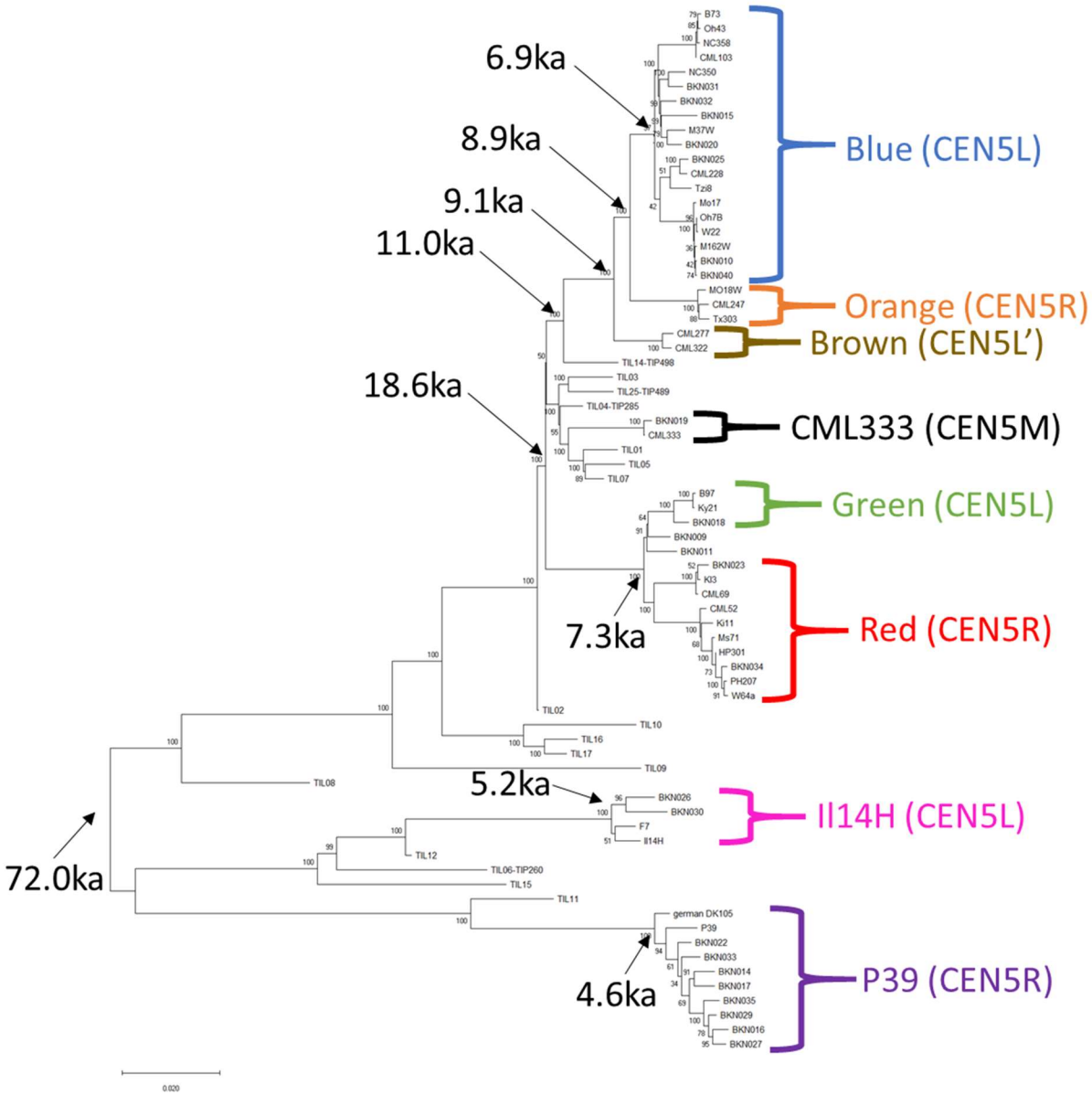


Figure S1. CEN5 tree generated from HapMap3 SNP data.
Tree was generated in MEGAX with SNP data from 105,978,235 to 106,977,241 in HapMap3 (B73_RefGenv4 coordinates) representing a non-recombinant region of CEN5L. Major colored lineages identified in Schneider et al. 2016. Estimated divergence dates for key nodes are shown and were calculated from HapMap3 SNPs. The centromere position is noted for each colored lineage which shows centromere position is not genetically linked.

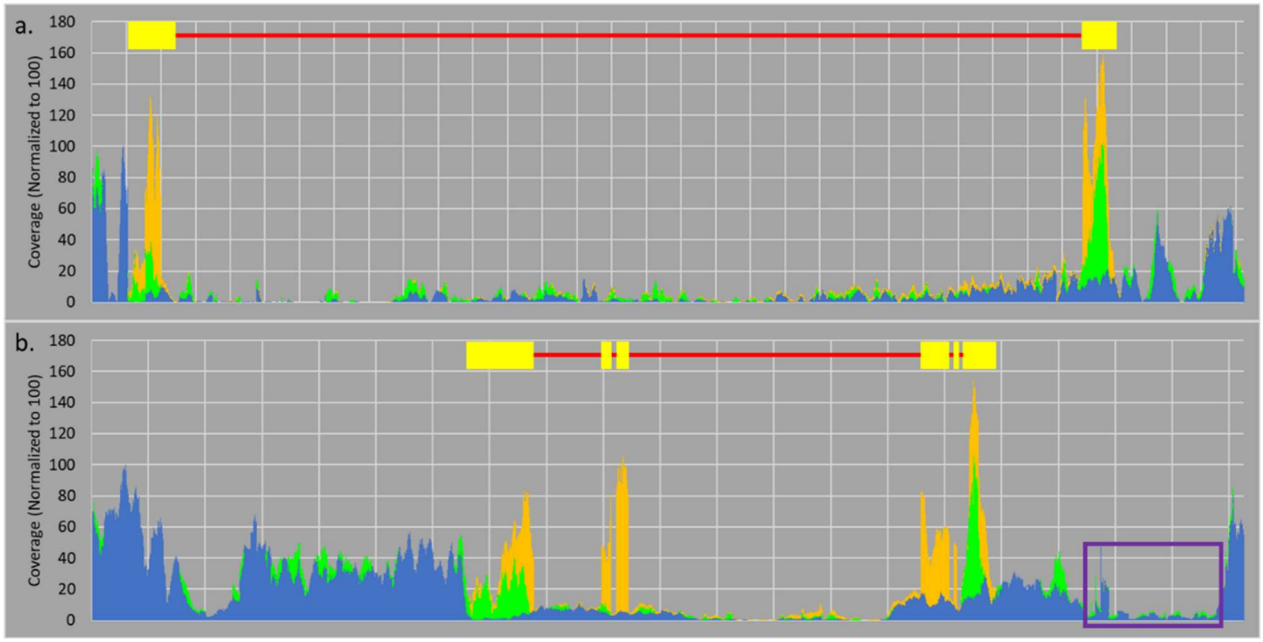


Figure S2. Nucleotide level coverage levels for cenH3 (blue), H2AZ (green) and RNAseq (Orange) at positions covered by the maximal 101mers.

Exons for the unannotated gene in the active CEN3 (**a**) and Zm00001d050645 (v4 annotation), the gene in the active CEN4 (**b**) are marked as yellow boxes, introns are denoted as red lines. The purple box is a region of abnormally low coverage caused by a recent CR insertion. Coverage values are normalized to 100, horizontal gridlines are in 1kb intervals.

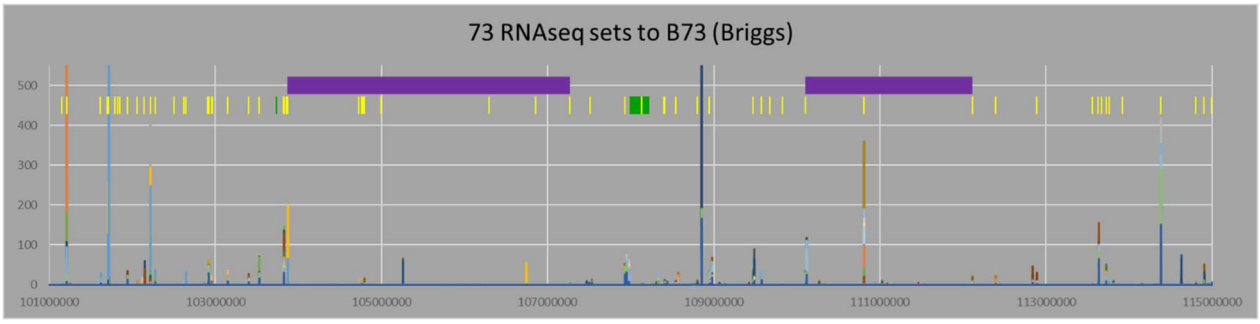


Figure S3. RNAseq coverage for 73 datasets mapped to B73.

Overlapping raw coverage values averaged over 1kb regions across CEN5L and CEN5R for the 73 datasets. CentC tandem repeat is shown as green lines and v5 gene models with yellow lines. The purple bars correspond to regions A-D (left) and E-H (right). There are 3 peaks with coverage levels >50 in CEN5L/R only one of these peaks corresponds to a gene model (CEN5R, Zm00001eb235330, gene marker “F”).

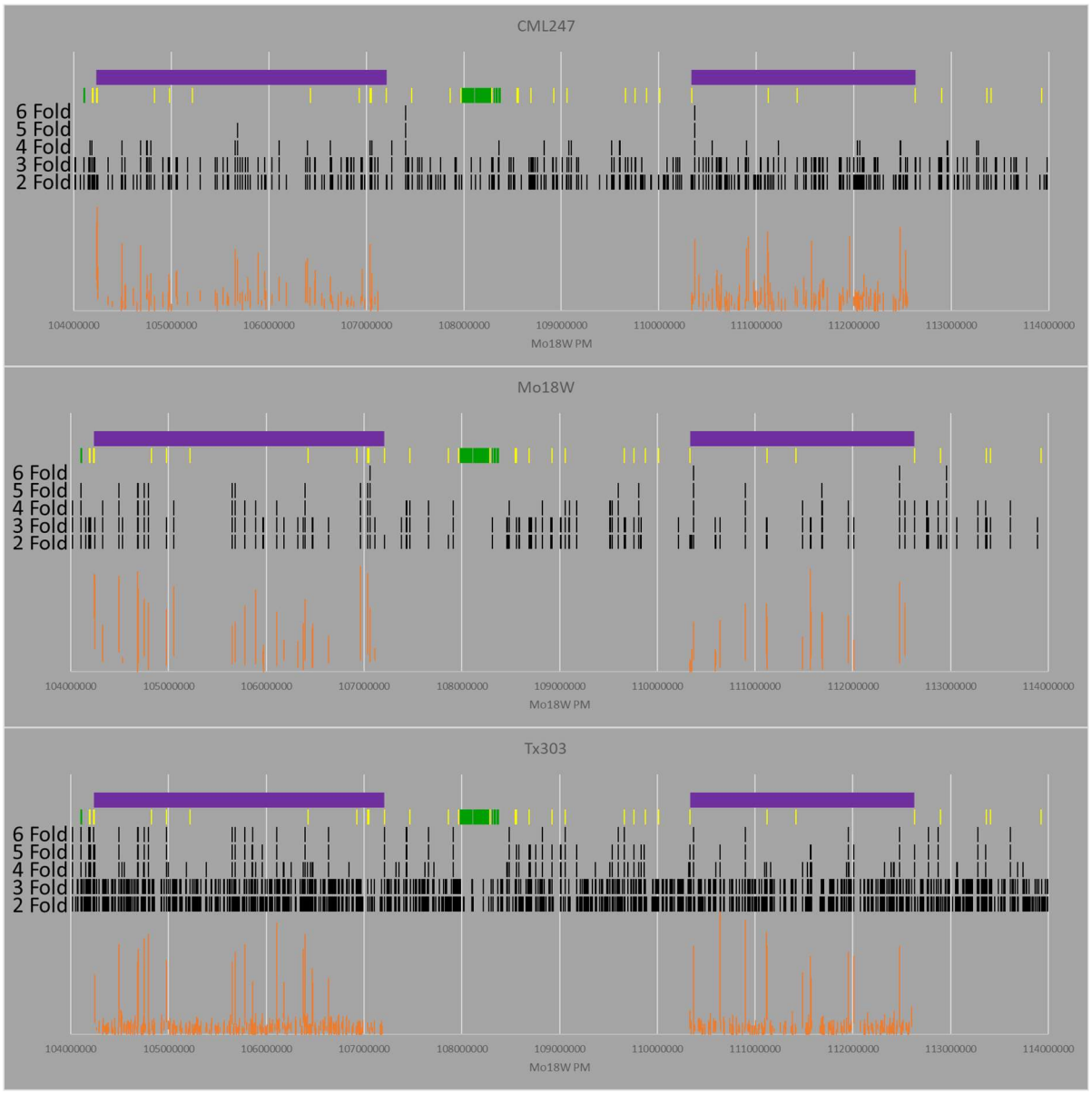


Figure S4. Variation in MACS2 H2A.Z Peak calling among Orange lines using varied fold change thresholds.

Relative coverage levels for peaks called with MACS2 with no fold change threshold are shown in orange. The regions corresponding to peaks called with fold change thresholds 2-6 are shown as black lines. CentC tandem repeat is shown as green lines and v5 gene models with yellow lines. The purple bars correspond to regions A-D (left) and E-H (right). We observe many false positive peak calls in Tx303 and CML247. There are more false positive peak calls in CEN5R than CEN5L in CML247 which led to a skewed H2A.Z ratio. We see a dramatic loss of false positive peaks between fold change thresholds 3 and 4. However at a fold change threshold of 4 we start to lose peaks called in Mo18W which does not display from the false positive calls. A fold change threshold of 3.5 was chosen to remove false positive calls from CML247 and Tx303.

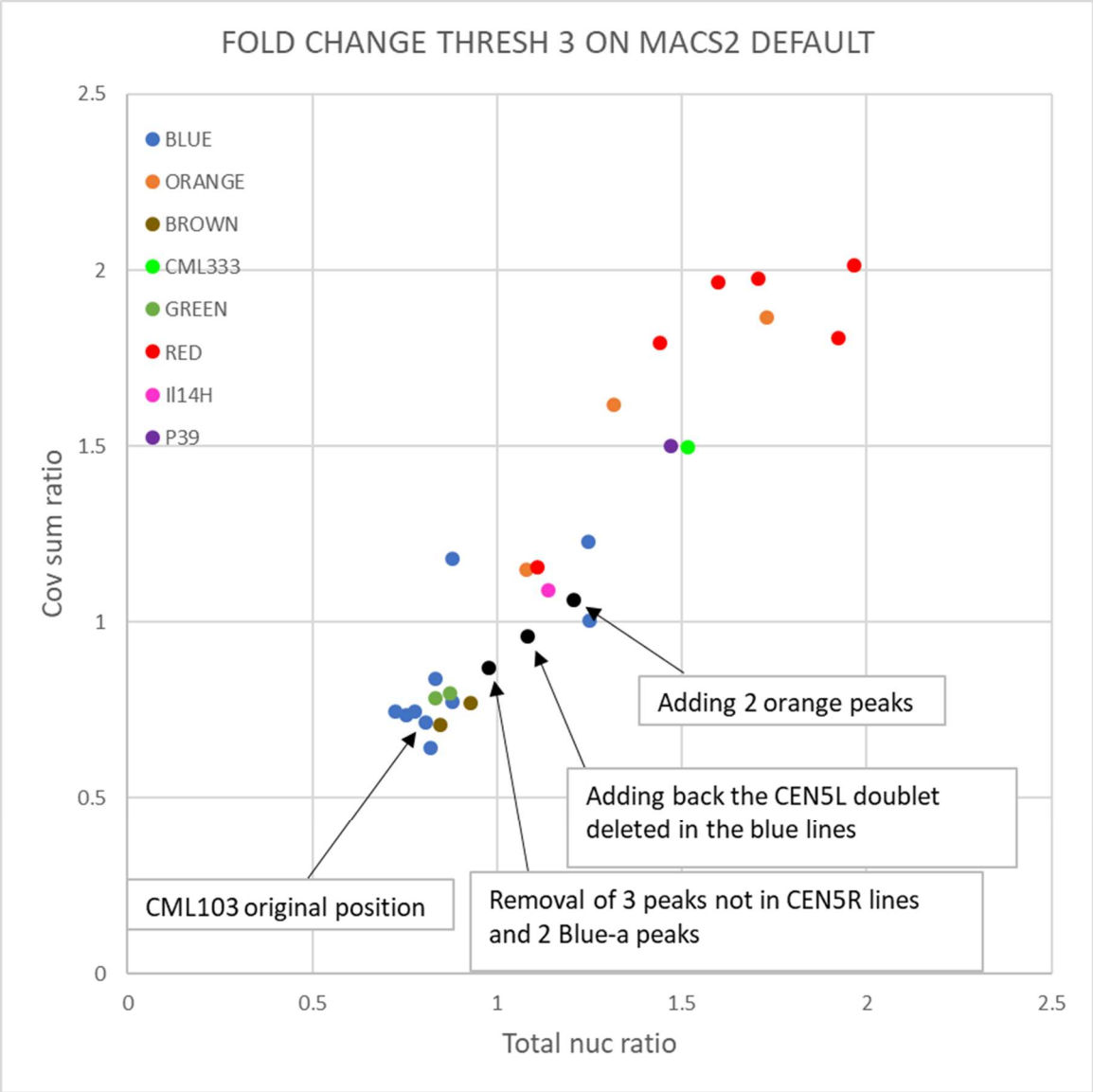


Figure S5. Single peaks/deletions are not responsible for the differential ratios observed between CEN5R and CEN5L lines.

To evaluate the effect of single peaks on the ratios observed between CEN5L and CEN5R lines we manually adjusted the data for the Blue line CML103 to look more like the orange line Mo18W. Initially 5 peaks in CEN5R were removed from CML103. The deleted Blue doublet was then added back to CML103 (using Mo18W coverage data), finally two orange specific peaks were added to CML103 in CEN5L (Mo18W coverage data). With each change the CML103 ratio moves closer to the CEN5R cluster but no single change is sufficient to make the blue CML103 move into the CEN5R ratio cluster.

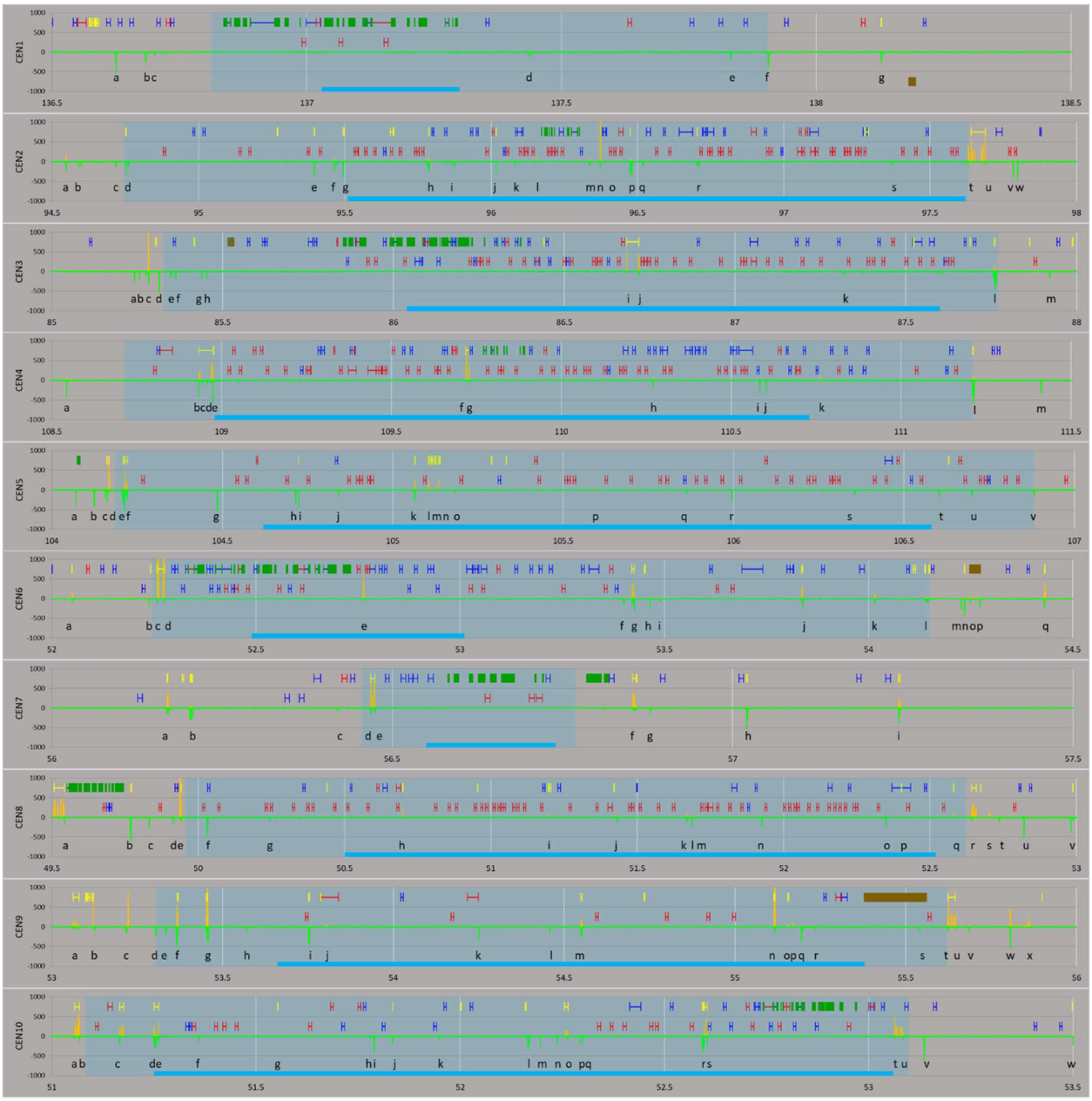


Figure S6. Overview of H2A.Z and RNAseq peaks in the active B73_v4CEN centromeres.

Mean H2A.Z and RNAseq coverage per 100nt window are marked in light green and orange respectively. Active B73 centromeres are marked with a light blue line. Regions of elevated cenH3 (CEN shoulders) are shaded light blue. Post domestication CR1 and CR2 insertions are denoted in red and blue respectively directly above the baseline. RefGen_v4 Gene models (yellow), centC (dark green), non-centC repeats (brown) and post-domestication CR1 and CR2 insertions (blue and red respectively) are marked at the top of each chart. Peaks with a maximum coverage >35 are alphabetically labeled.

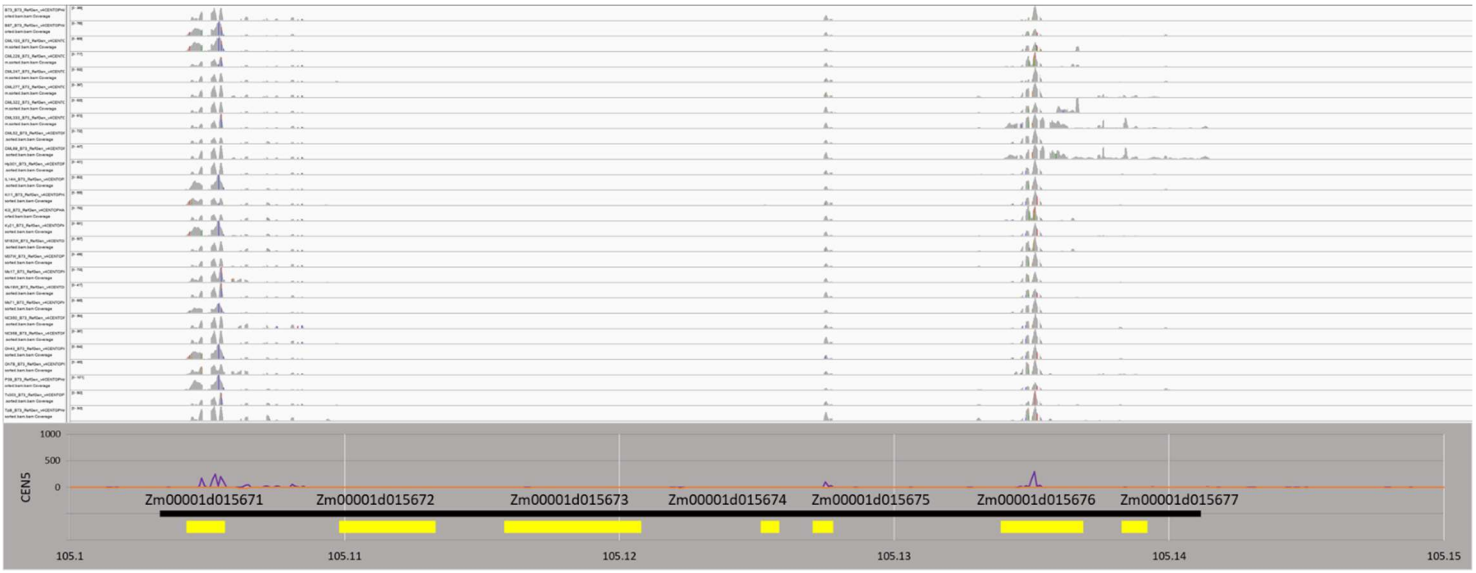


Figure S7. Transcription over the CEN5 gene graveyard.

RNAseq for 27 NAM lines from combined ear, root, shoot, and tassel tissues. The gene graveyard is annotated in black and the each of the seven genes are marked in yellow. The orange line represents H2A.Z coverage and purple represents B73 mean RNAseq coverage per 100nt window.

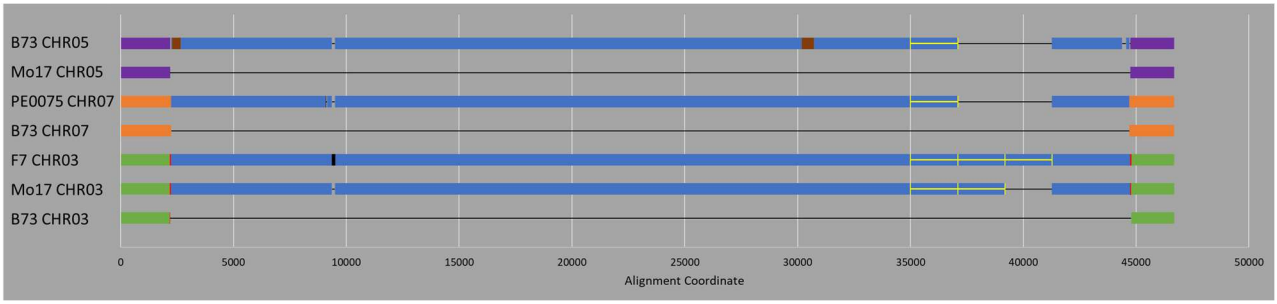
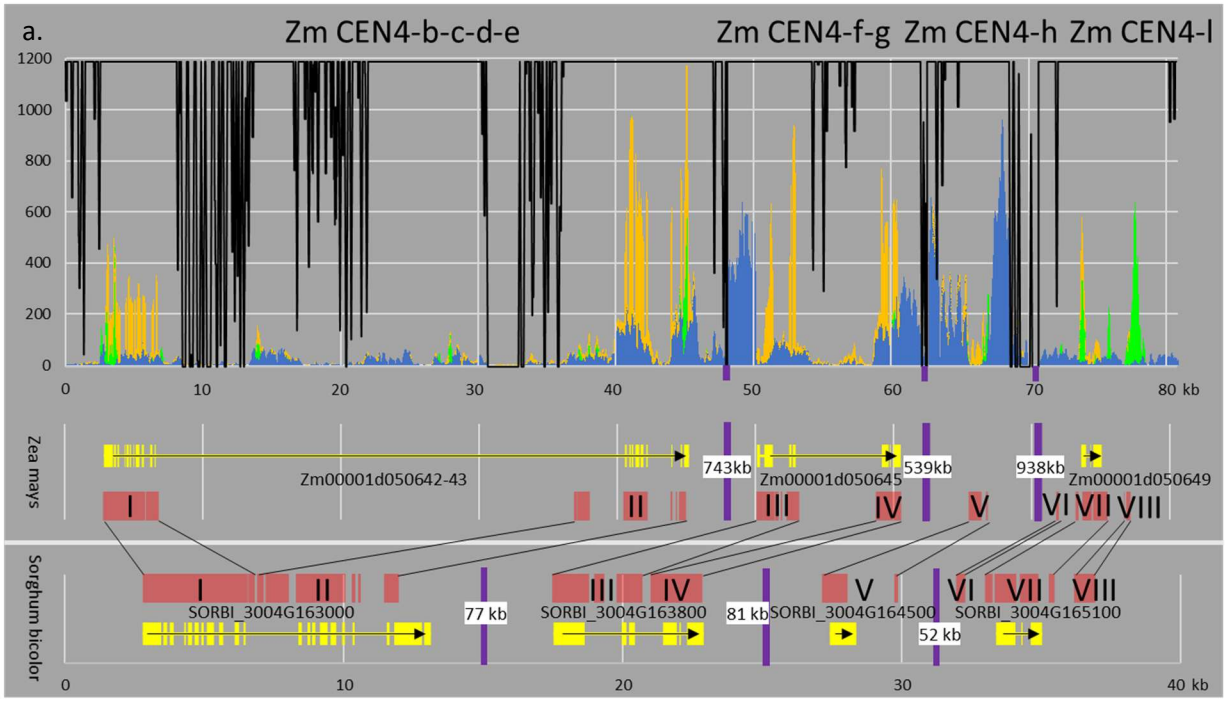
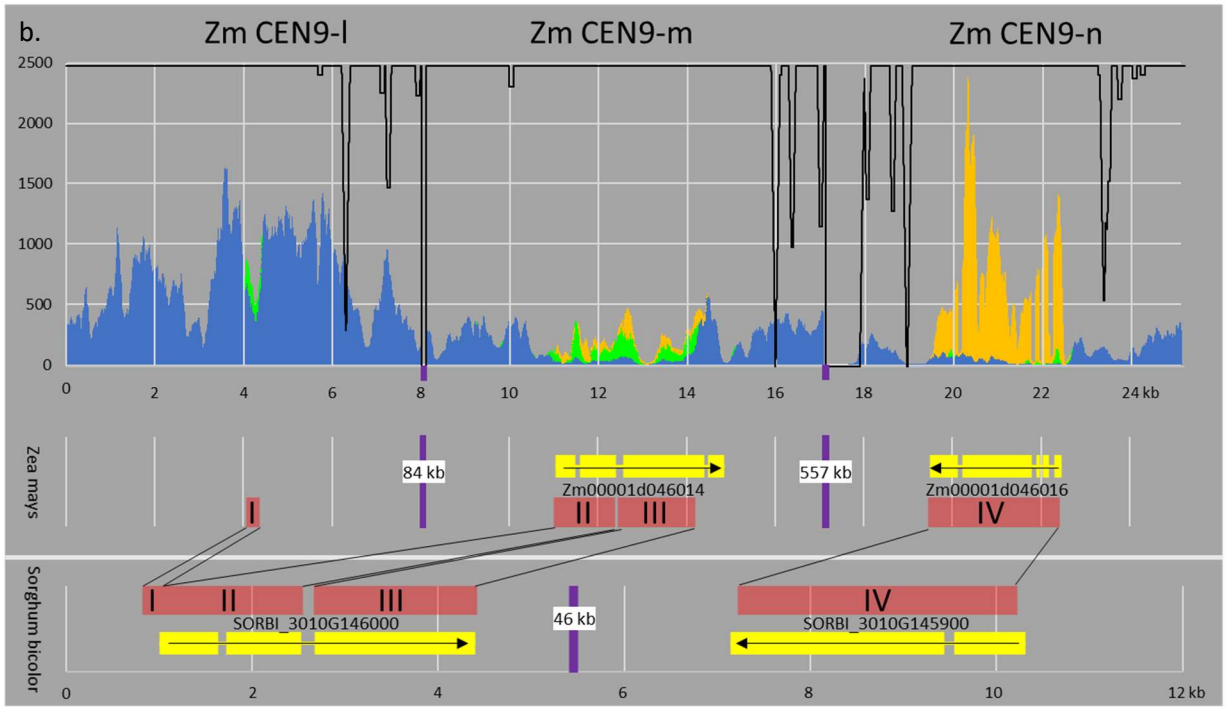


Figure S8. Schematic of Gene graveyard alignment.

Purple Orange and green represent matching flanking sequence between chromosome 5, 7 and 3 respectively. Red indicates the position of the AT satellite repeat. Blue sequence corresponds to aligned gene graveyard sequence. Thin black lines indicate gaps in the alignment and thick black lines indicate spans of N's. Brown are regions that have no homology.


a.
Zm CEN4-b-c-d-e
Zm CEN4-f-g
Zm CEN4-h
Zm CEN4-l
1200
1000
800
600
400
200
0
0
10
20
30
40
50
60
70
80 kb
Zea mays
Zm00001d050642-43
743kb
Zm00001d050645
539kb
938kb
Zm00001d050649
I
II
III
IV
V
VI
VII
VIII
Sorghum bicolor
SORBI_3004G163000
77 kb
SORBI_3004G163800
81 kb
SORBI_3004G164500
52 kb
SORBI_3004G165100
0
10
20
30
40 kb


b.
Zm CEN9-l
Zm CEN9-m
Zm CEN9-n
2500
2000
1500
1000
500
0
0
2
4
6
8
10
12
14
16
18
20
22
24 kb
Zea mays
84 kb
Zm00001d046014
557 kb
Zm00001d046016
I
II
III
IV
Sorghum bicolor
SORBI_3010G146000
46 kb
SORBI_3010G145900
0
2
4
6
8
10
12 kb

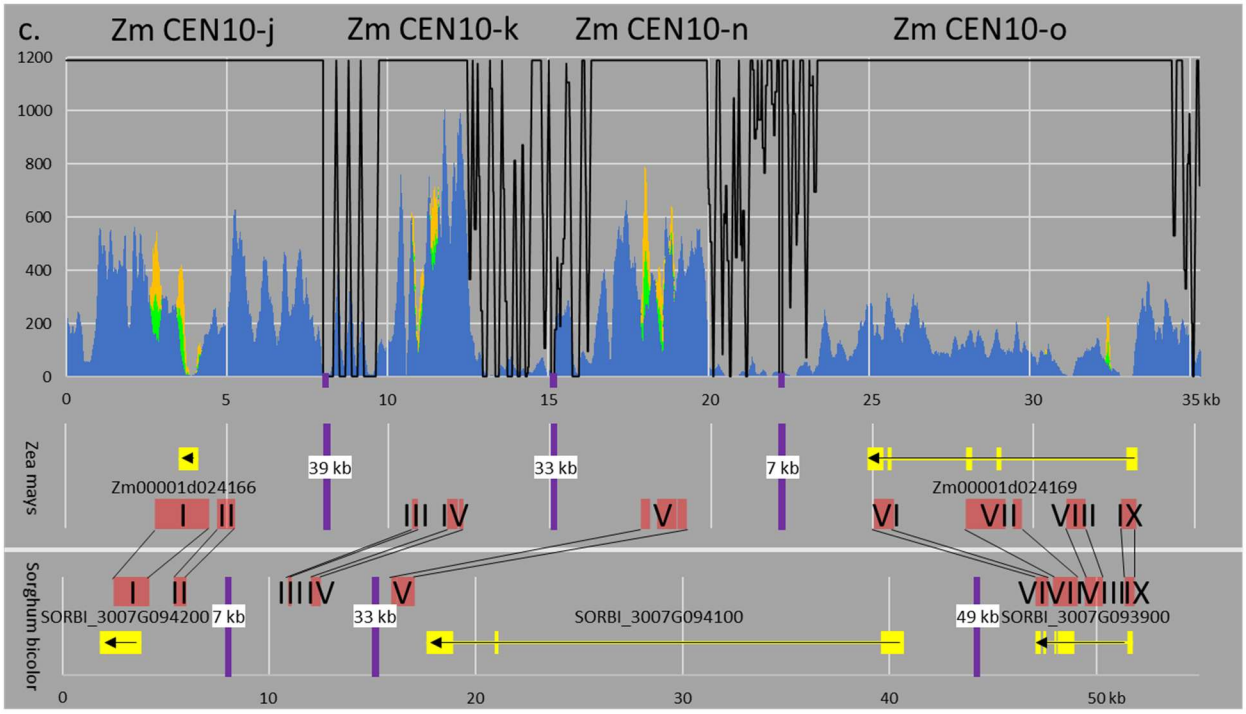


Figure S9. Isolated H2A.Z loci lacking RNA-seq can form due to gene damage.

Panels **a, b,** and **c** correspond to examples of isolated H2A.Z peak formation on chromosomes, 4, 9, and 10, respectively. The top panel contains CenH3 (blue), H2A.Z (green) and RNAseq (orange) coverage in B73 for each window which is separated by a purple bar. Unique 101mer coverage is marked in black. The middle and bottom panel are schematic diagrams of corresponding regions in *Zea mays* and *Sorghum bicolor* respectively. Each contains Gene models (yellow) and blast HSPs (red). Shared blocks of HSPs are connected by black lines. The syntenic regions of *Sorghum bicolor* have been reversed in CEN9 and CEN4 to reflect the orientation of *Zea mays.* See *SI Appendix text SX2* for additional details.

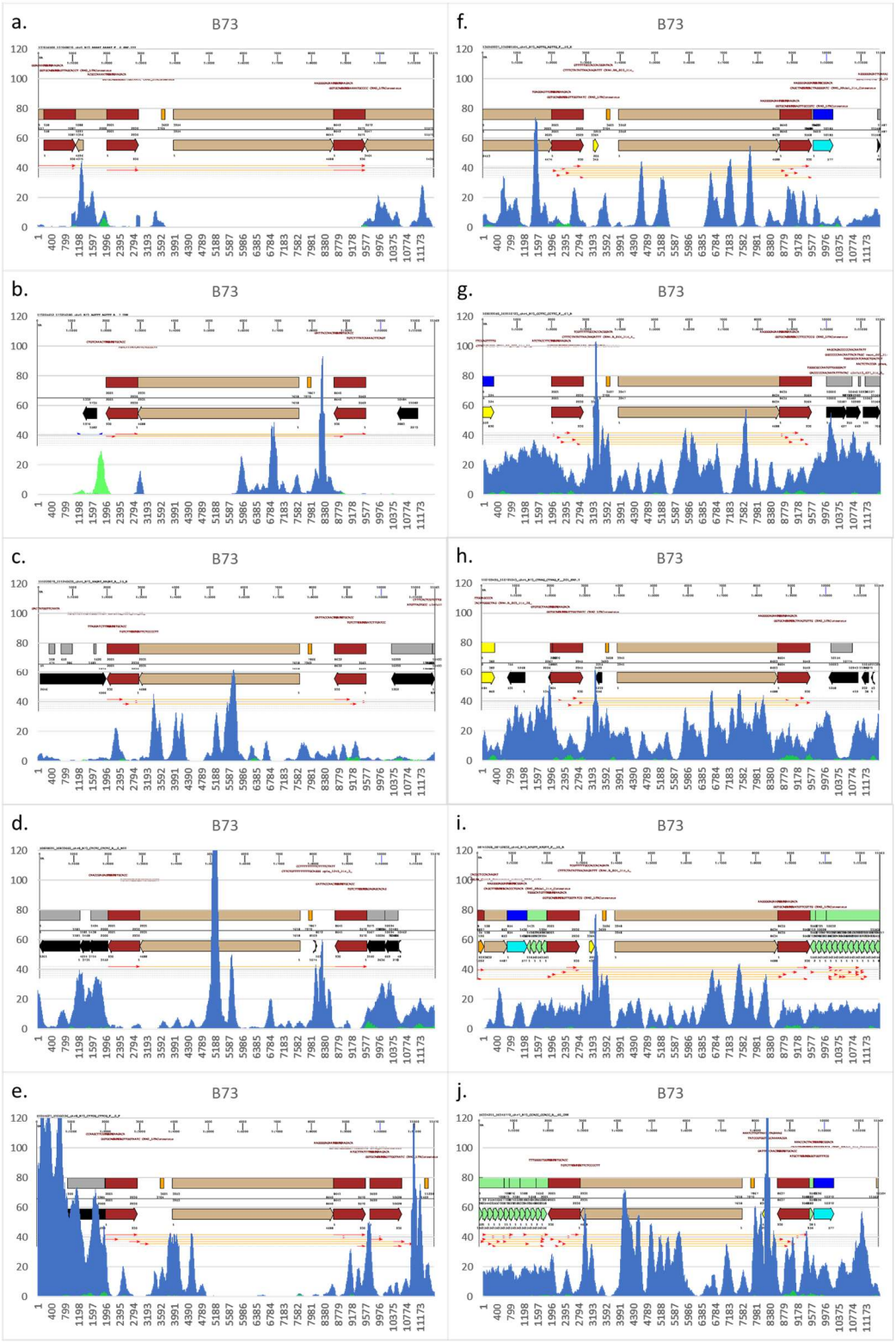


Figure S10. CR2 Associated H2A.Z peaks in B73_v5.

Per nucleotide coverage data for H2A.Z (green) and cenH3 (blue) normalized to the CEN5L control peak overlaid on top of Junction viewer [108] annotations of the region. Each CR2 of interest is shown in the native orientation with 2kb of flanking sequence.

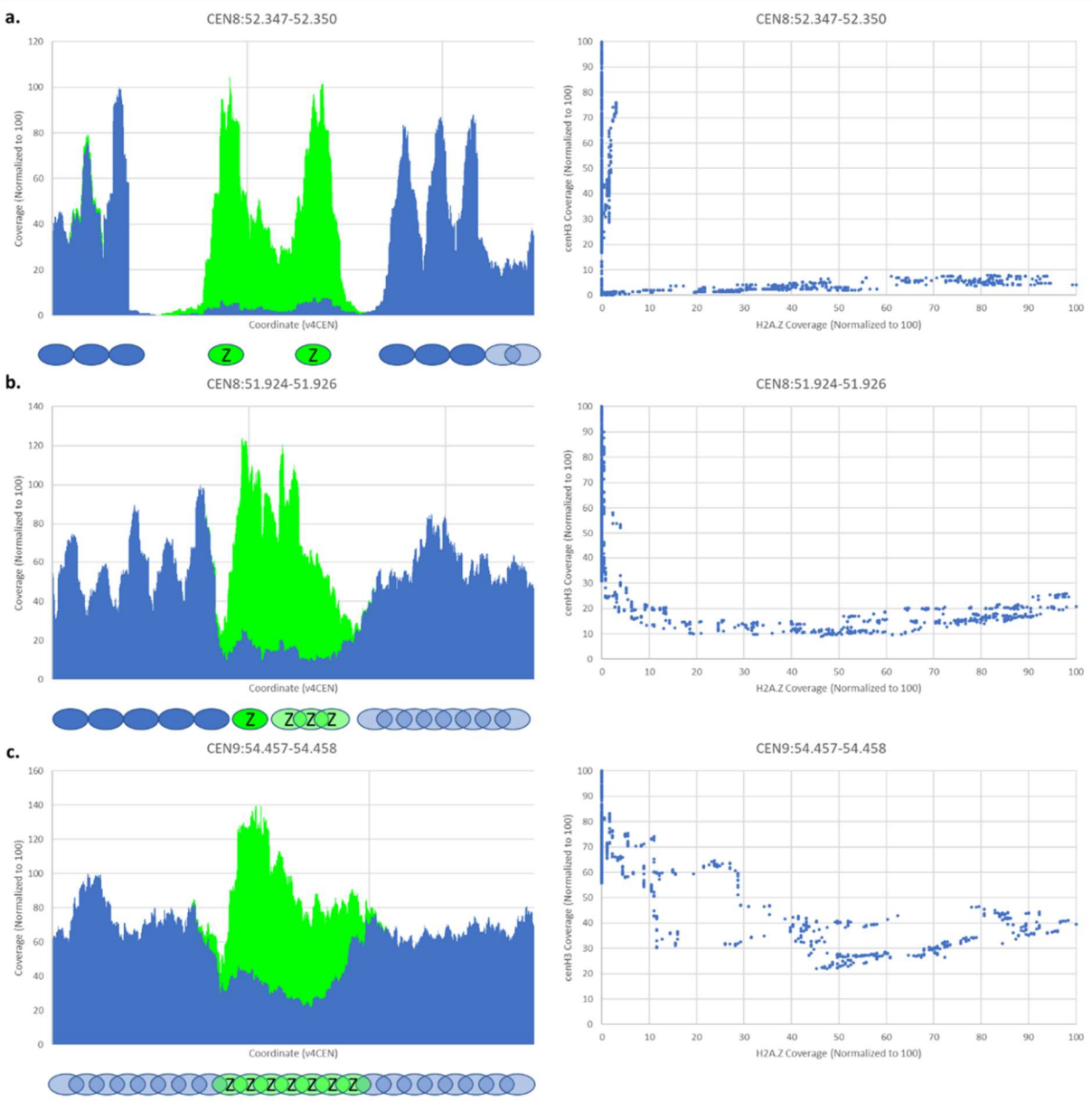


Figure S11. Defined and delocalized nucleosome positioning influences the observed inverse H2A.Z/cenH3 relationship.

Stacked bar graphs of cenH3 (blue) and H2A.Z (green) normalized to 100 for all positions covered entirely by 101mers and the corresponding H2A.Z vs cenH3 comparison normalized to 100. Beneath each stacked bar graph is a schematic representation of approximate nucleosome positions over the window. Solid ovals correspond to defined nucleosome positions and overlapping transparent ovals correspond to undefined regions. **a**. This example displays defined nucleosome positioning on either side of the H2A.Z peaks as indicated by clear and consistent cenH3 peaks. **b**. An example of a partially diffuse H2A.Z locus with defined positioning upstream and diffuse positioning downstream. **c**. An example of a fully diffuse H2A.Z locus having poorly defined positioning throughout the window.

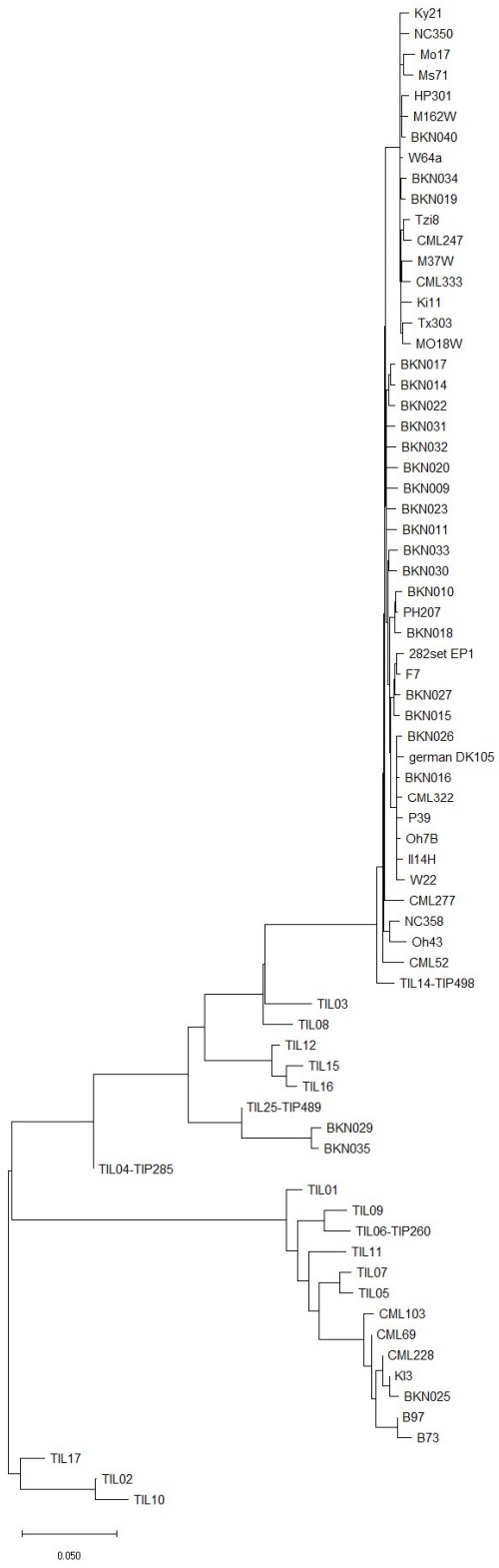


Figure S12. CEN2 HapMap3 SNP tree for REGON13_SYN.
This tree was generated from Hapmap3 region defined in *SI Appendix Table ST7.*

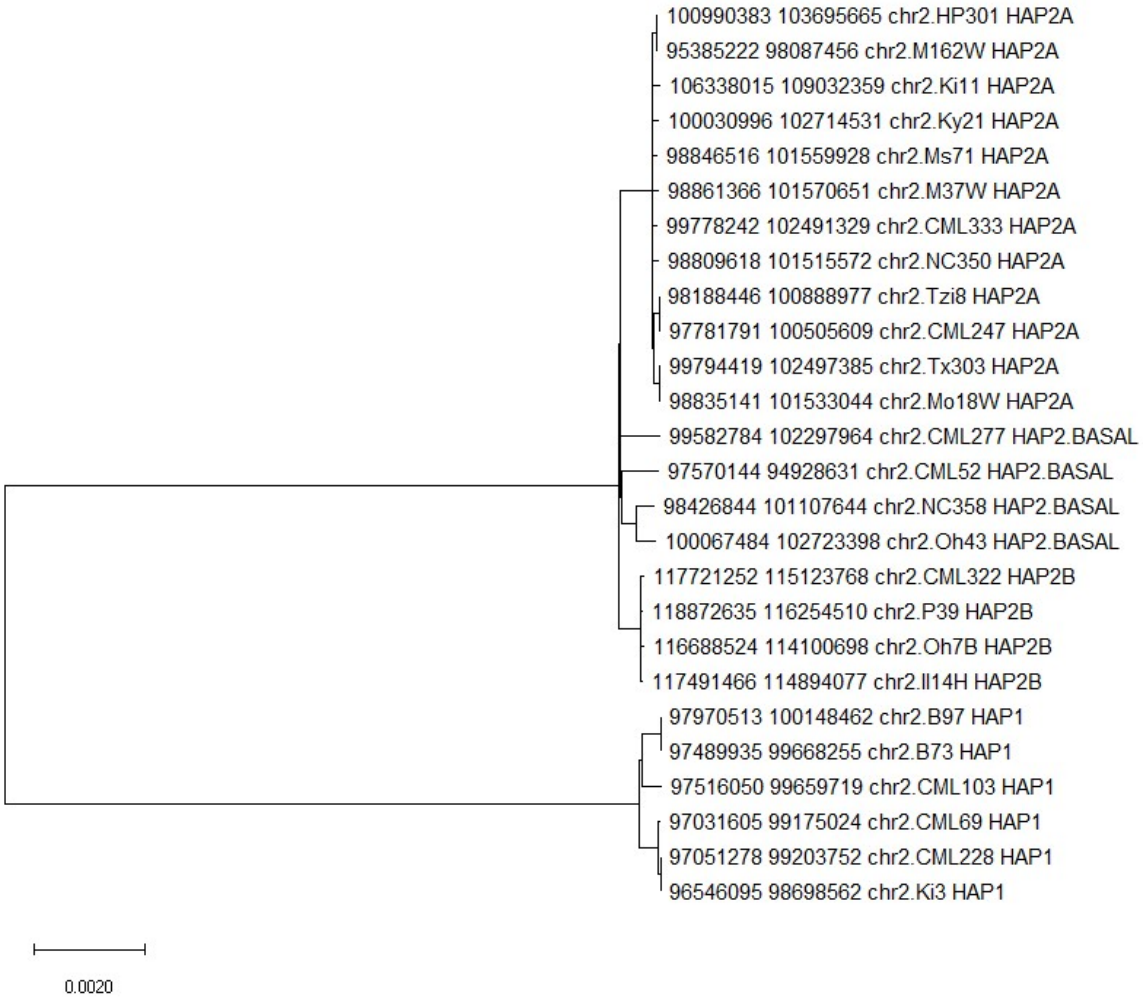
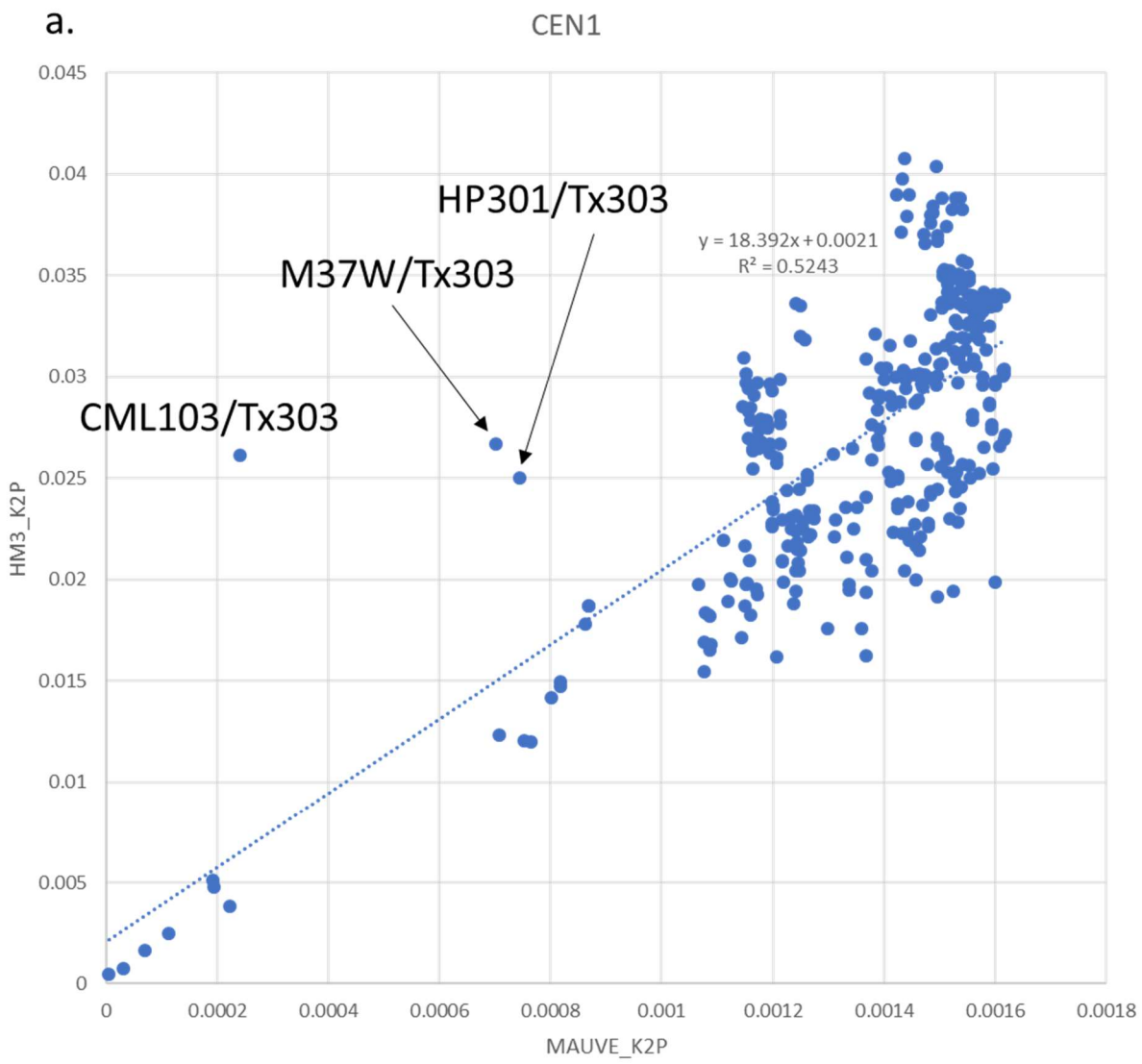
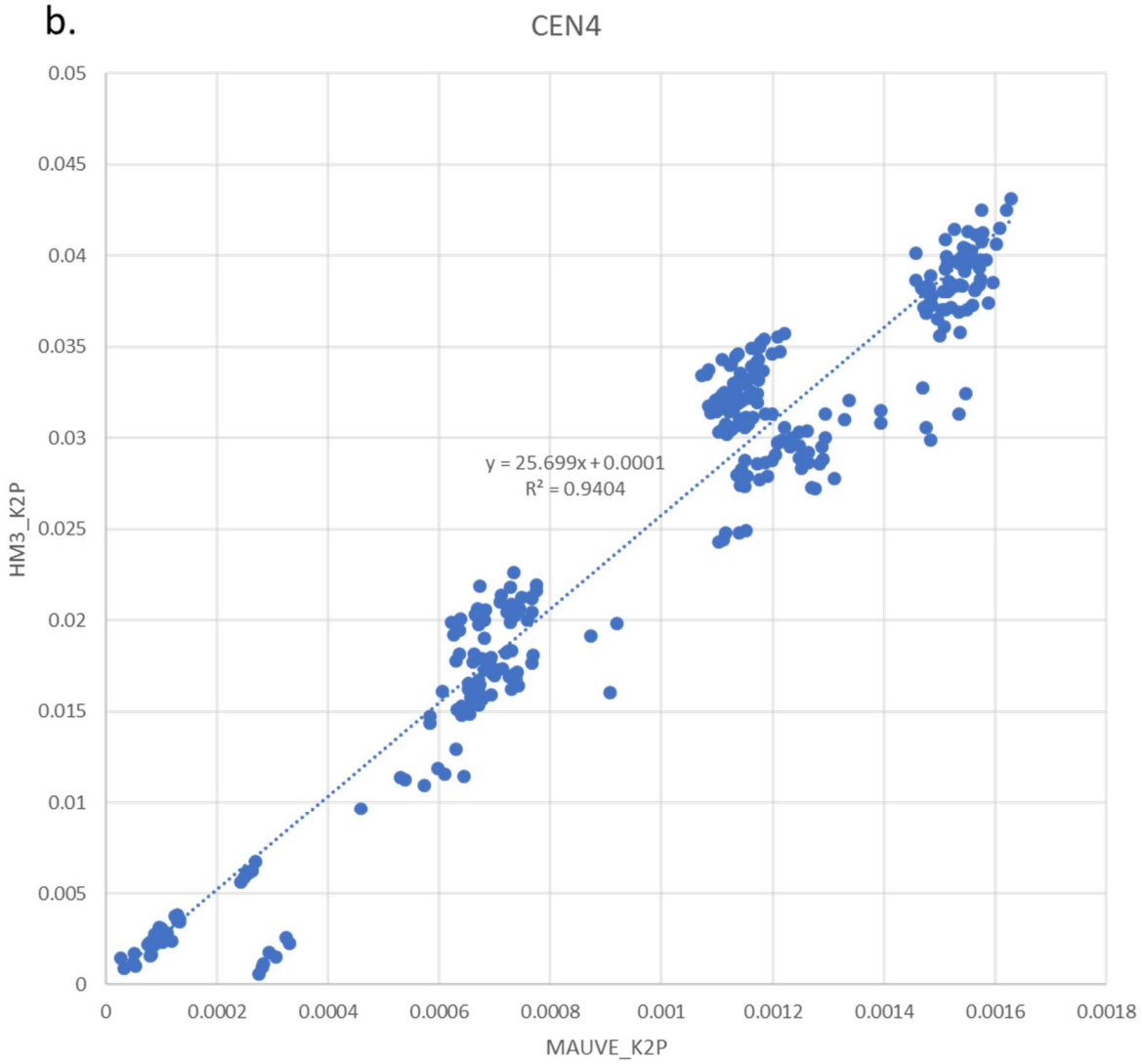


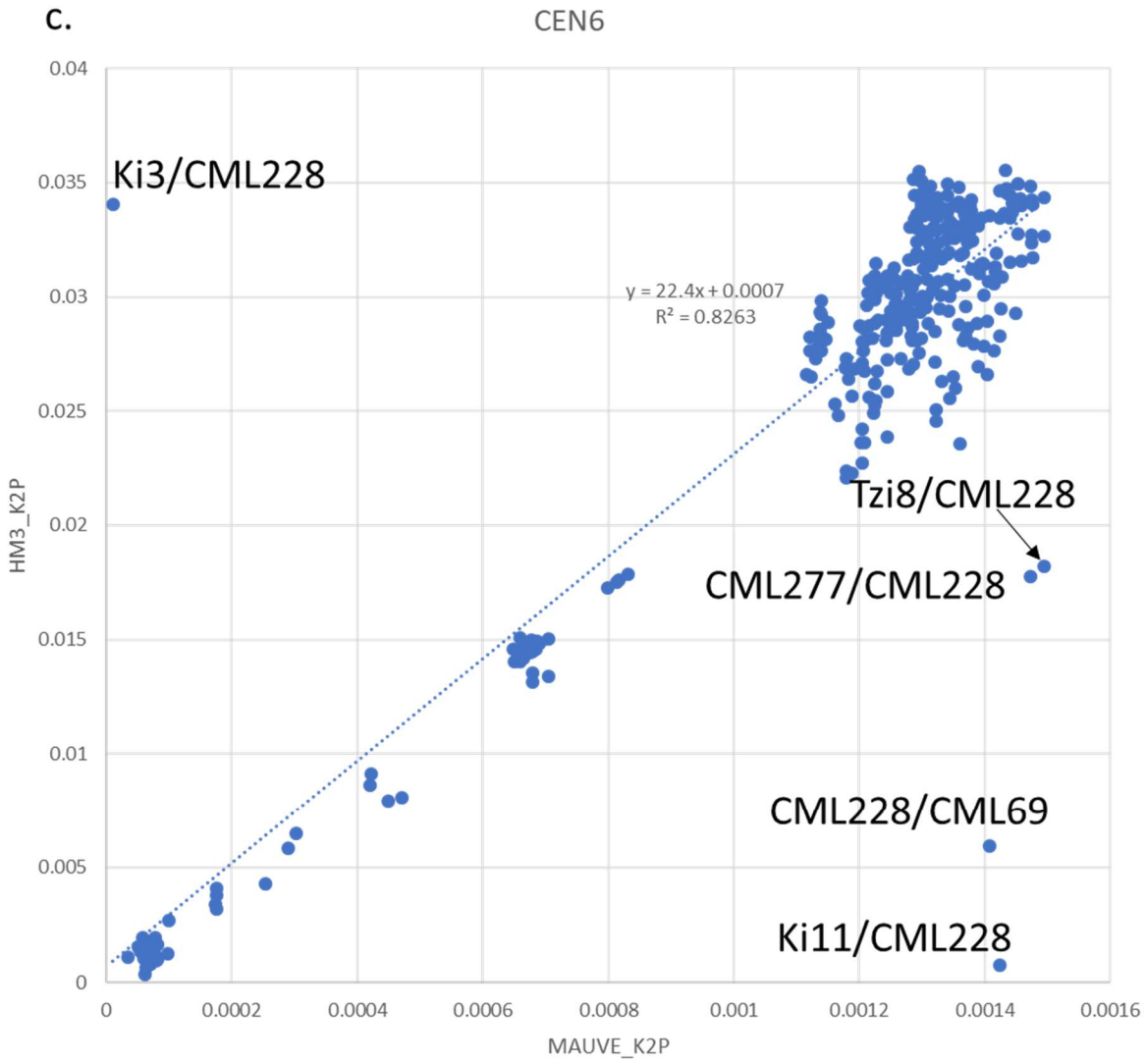
Figure S13. CEN2 Physical Map tree for REGION13_SYN.

This tree was generated from physical map and manually edited for continuity (see *SI Appendix Table ST7* for B73 v4 and HapMap3 coordinates for REGION13_SYN). Headers include the start and stop coordinates for the region in each line, the chromosome, NAM line and CEN2 haplotype.


a.
CEN1
HP301/Tx303
M37W/Tx303
CML103/Tx303
y = 18.392x + 0.0021
R² = 0.5243
HM3_K2P
MAUVE_K2P
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.0002
0.0004
0.0006
0.0008
0.001
0.0012
0.0014
0.0016
0.0018

b.


CEN4
y = 25.699x + 0.0001
R² = 0.9404
HM3_K2P
MAUVE_K2P
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.0002
0.0004
0.0006
0.0008
0.001
0.0012
0.0014
0.0016
0.0018


C.
CEN6
Ki3/CML228
y = 22.4x + 0.0007
R² = 0.8263
Tzi8/CML228
CML277/CML228
CML228/CML69
Ki11/CML228
HM3_K2P
MAUVE_K2P
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.0002
0.0004
0.0006
0.0008
0.001
0.0012
0.0014
0.0016

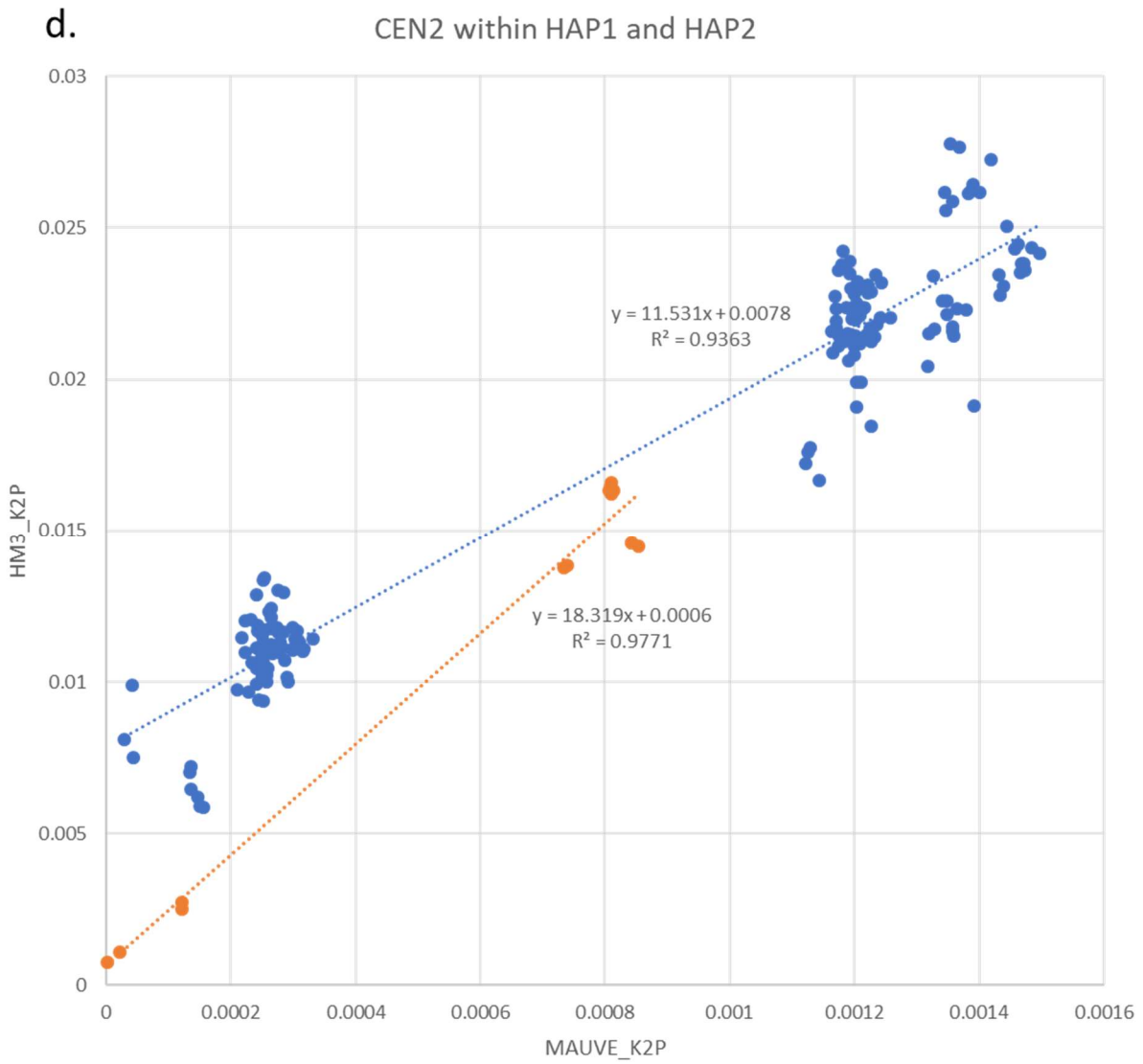


Figure S14. Comparisons of Physical map and HapMap3 K2p over equivalent regions for CENs 1 (a), 4 (b), 6 (c) and 2 (d).

Each point represents a comparison of a pair of NAM lines over identical regions using manually edited physical map alignments (Mauve) and SNP data (HM3). The Mauve K2P value on the X axis and HapMap3 K2P value on the Y axis. Trends are linear, so a linear trend line was added along with the R2 values and estimated slope. The best correlation was found in CEN4 (R2=0.94), followed by CEN6 (R2=0.83) and then CEN1 (R2=0.52). There is a considerable difference between K2P conversion depending on the haplotype as observed in CEN2 this is likely due to the B73 (HAP1 CEN2) reference being unable to detect HAP2 SNPs in CEN2. Coordinates for the regions can be found in *SI Appendix Table ST8.* Maximum divergence dates for CEN1, 4 and 6 can be found in *SI Appendix Table ST18*.

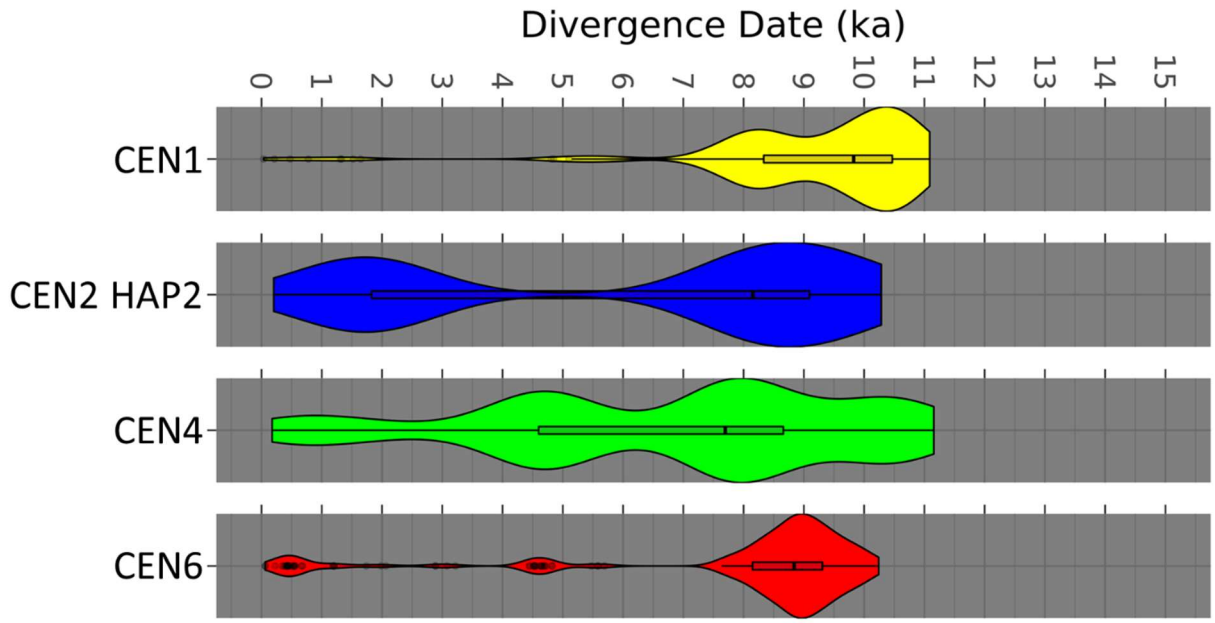


Figure S15. Distribution of divergence dates among alignments of CEN1, CEN2, CEN4, and CEN6.
All dates were generated with the centromeric conversion factor except (*SI Appendix Table ST18*). CEN2 dates were generated within HAP2 only. Coordinates for the regions dated can be found in *SI Appendix Table ST8*. Divergence date distributions among the 3 single haplotype centromeres (CENs 1, 4, and 6) suggest that CEN4 was the first to be domesticated (Max date 11.16ka), followed by CEN1 (11.09ka), then CEN2 HAP2 (10.28ka), and finally CEN6 (10.24ka).

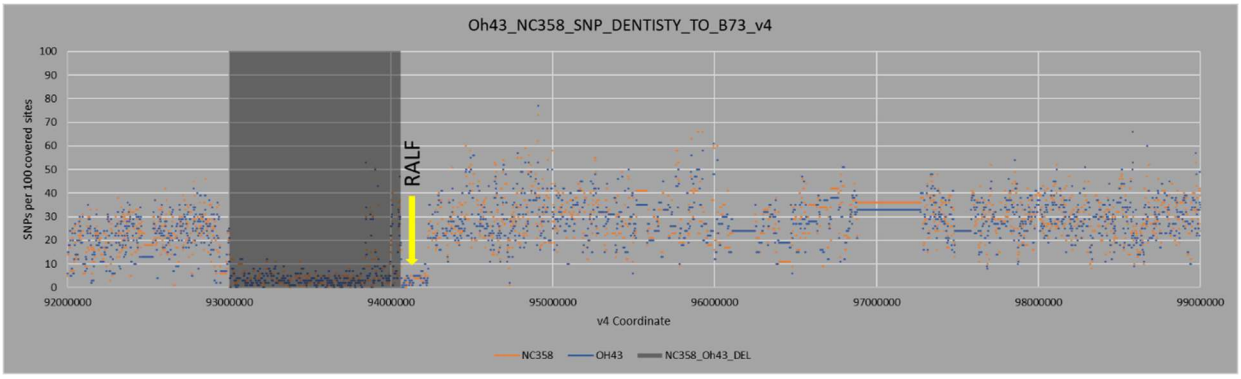


Figure S16. SNP density relative to B73 v4 per 100 covered HapMap3 sites across CEN2.
The larger of the two deletions flanking the CEN2 domestication region is shaded grey and contains artifactual SNP peaks. The CEN2 domestication region is marked with a yellow arrow.

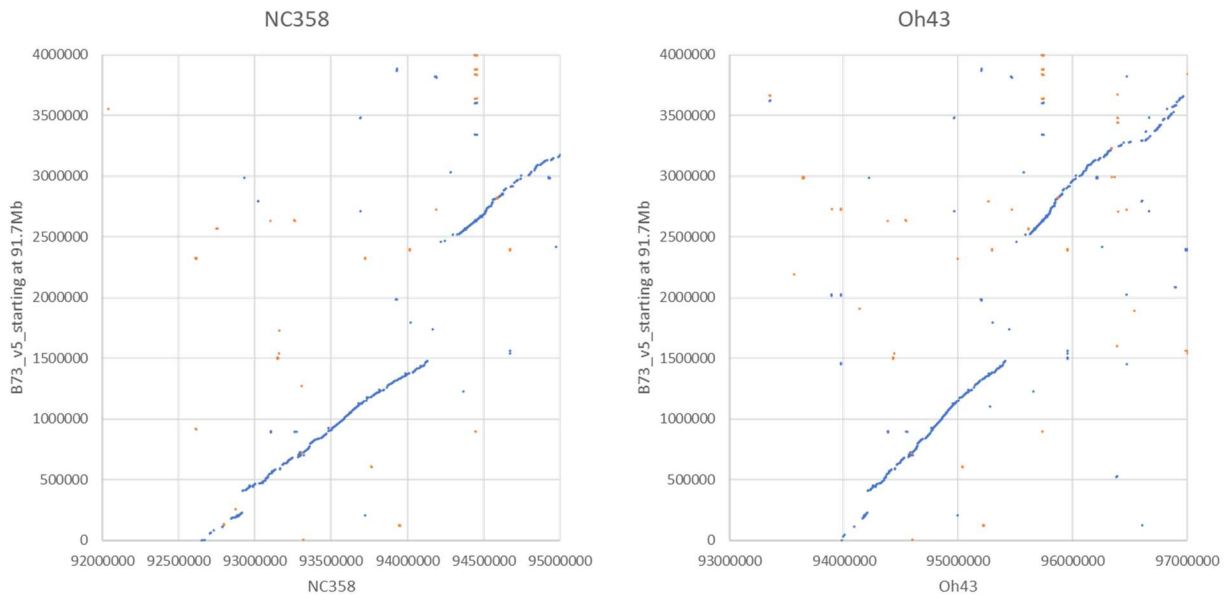


Figure S17. MUMmer (word length 200) of NC358 and Oh43.
MUM200 of a subregion of B73 v5 starting at 91.7Mb showing two deletions which obfuscate SNP density in HapMap3.

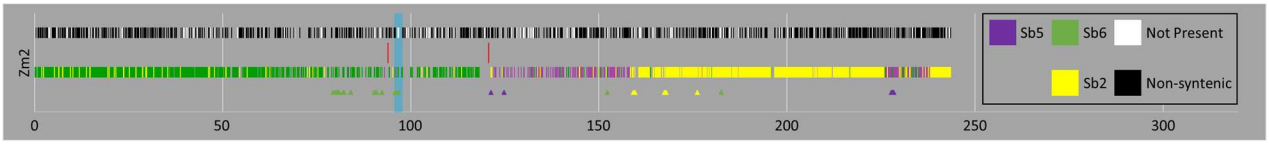


Figure S18. Syntenic relationships between Zea mays chromosome 2 (B73_v5) and Sorghum bicolor (NCBI_v3, GCF_000003195.3).
Lines represent B73 gene models whose highest blast hit was to a syntenic *Sorghum* chromosome and are colored based on *Sorghum* chromosome. Black and white lines represent B73 gene models whose highest blast hit was to a non-syntenic Sorghum chromosome and B73 gene models with no blast hit in the Sorghum genome, respectively. The large red lines mark the HAP2B inversion breakpoints mapped to the B73 genome. The two breakpoints are found within clusters of Sorghum pericentromeric markers for Sb6 and Sb5. The active *Zea mays* centromere 2 is shaded light blue. Triangles represent B73 chr2 gene models whose highest blast hit was within 5Mb of one of the syntenic *Sorghum bicolor* centromere regions. B73 gene models whose highest blast hit was within 5Mb of a non-syntenic centromere were discarded. *Sorghum bicolor* centromere regions were identified by blast of the Sorghum bicolor centromeric tandem repeat (CEN38, AF137608) to the Sorghum genome. The centromeric regions were defined as the largest island of CEN38. See *SI Appendix Figure S19* for an analysis of Sb synteny across the B73 v5 genome.

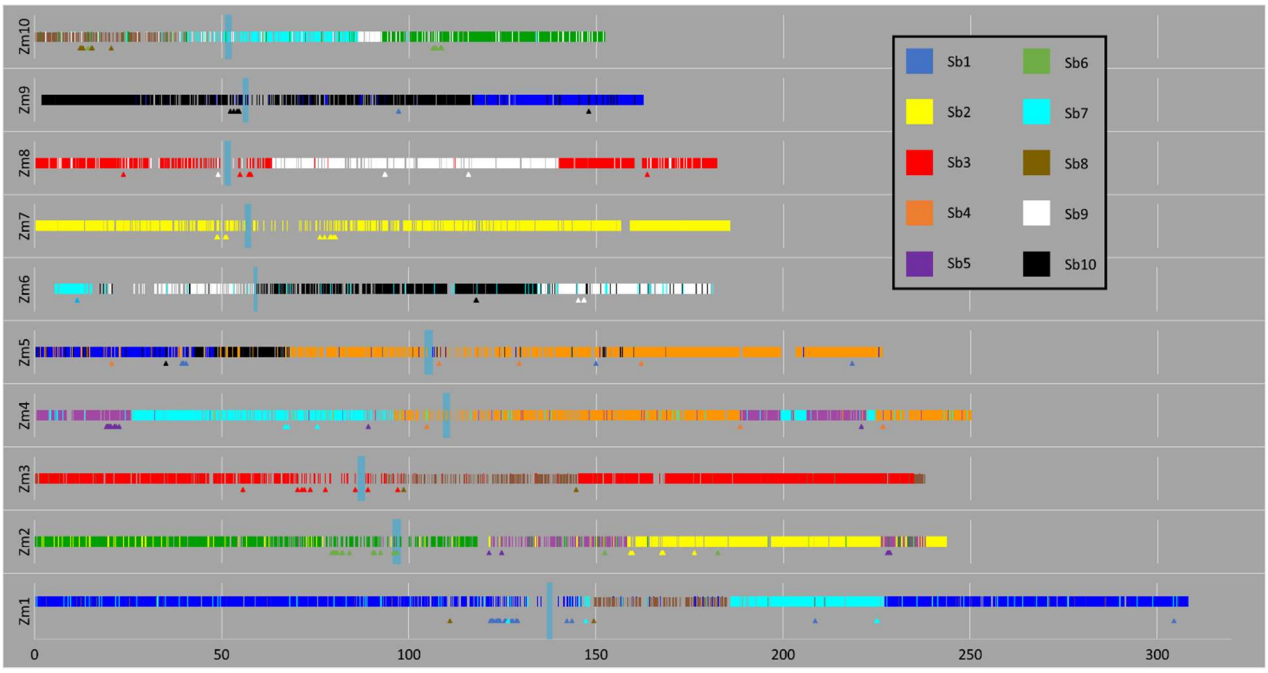


Figure S19. Syntenic relationships between Zea mays (B73_v5) and Sorghum bicolor (NCBI_v3, GCF_000003195.3).

Lines represent B73 gene models whose highest blast hit was to a syntenic *Sorghum* chromosome and are colored based on *Sorghum* chromosome. Active *Zea mays* centromere regions are shaded light blue. All *Zea* chromosomes are drawn to scale. Triangles represent B73 gene models whose highest blast hit was within 5Mb of one of the syntenic *Sorghum bicolor* centromere regions. B73 gene models whose highest blast hit was within 5Mb of a non-syntenic centromere were discarded. Sorghum bicolor centromere regions were identified by blast of the Sorghum bicolor centromeric tandem repeat (CEN38, AF137608) to the Sorghum genome. The centromeric regions were defined as the largest island of CEN38. Sorghum bicolor chromosome 1 lacked any CEN38 sequence. As such the centromere was defined by the largest gap in the chromosome (40,805,245-45,605,245 4.80Mb). Zea mays CHR06 showed a high density of Sorghum chr1 markers at ~25Mb but investigation of these markers indicated that they were repetitive in *Zea mays* (i.e., 101 Zm gene models from a 6.9Mb region hit the same 1404nt region in Sb chr1) and did not reflect a true syntenic relationship and as such are not displayed.

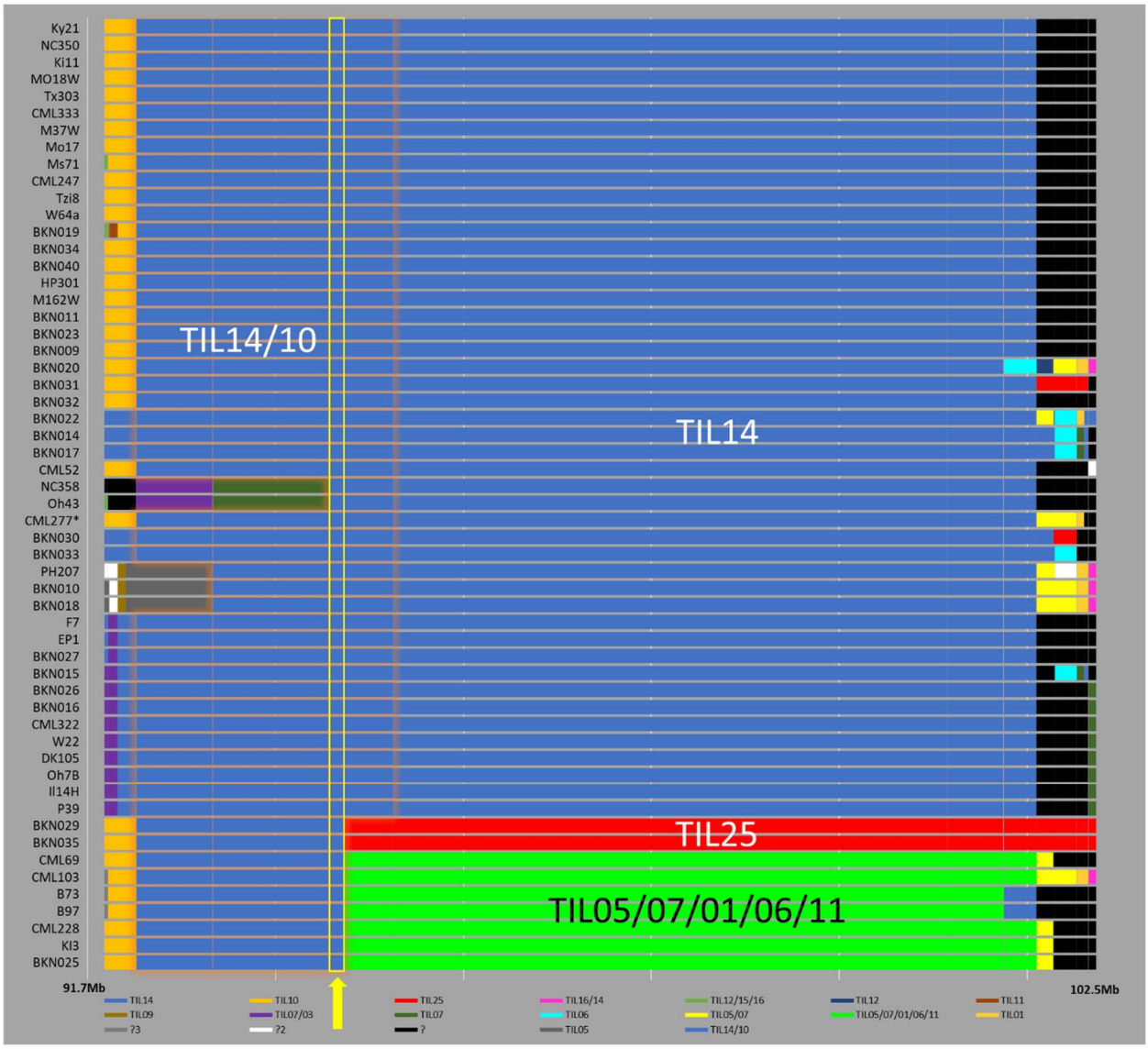


Figure S20. TIL progenitors and recombination breakpoints for 56 domesticated maize lines.
Hapmap3 data was investigated to identify the recombination breakpoints among domesticated maize. Trees were generated from 22 Regions across CEN2 defined by recombination, structural variation, and mauve alignment coverage. The X axis spans from 91,499,128-102,324,959 in B73 v4 (91.7Mb-102.5Mb in B73 v5) CEN2 contains 3 haplotypes among domesticated maize, HAP1, HAP2 and HAP3 are derived from TILs 05/07/06/01/11, TIL14 and TIL25 respectively. Tll10 shares sequence with TIL14 upstream of the CEN2 domestication region in all but NC358/Oh43, in the CEN2 domestication region and downstream of the HAP2/HAP1 and HAP2/HAP3 recombination breakpoints (Blue with orange glow). This breakpoint coupled with phylogenetic investigation of HM3 SNPs over the CEN2 domestication region suggest that TIL14 was the progenitor of both CEN2 HAP2 and CEN2 domestication (*Supplemental Text SX7*). A similar analysis was done for CEN5 and the domestication locus Region-D (See **Section 4.4.5** in main text, and main text **Figure 4.4**).

*CML277 contains a 305nt introgression of HAP1 sequence between the CEN2 domestication region and non-recombinant region (*SI Appendix Table ST12*).

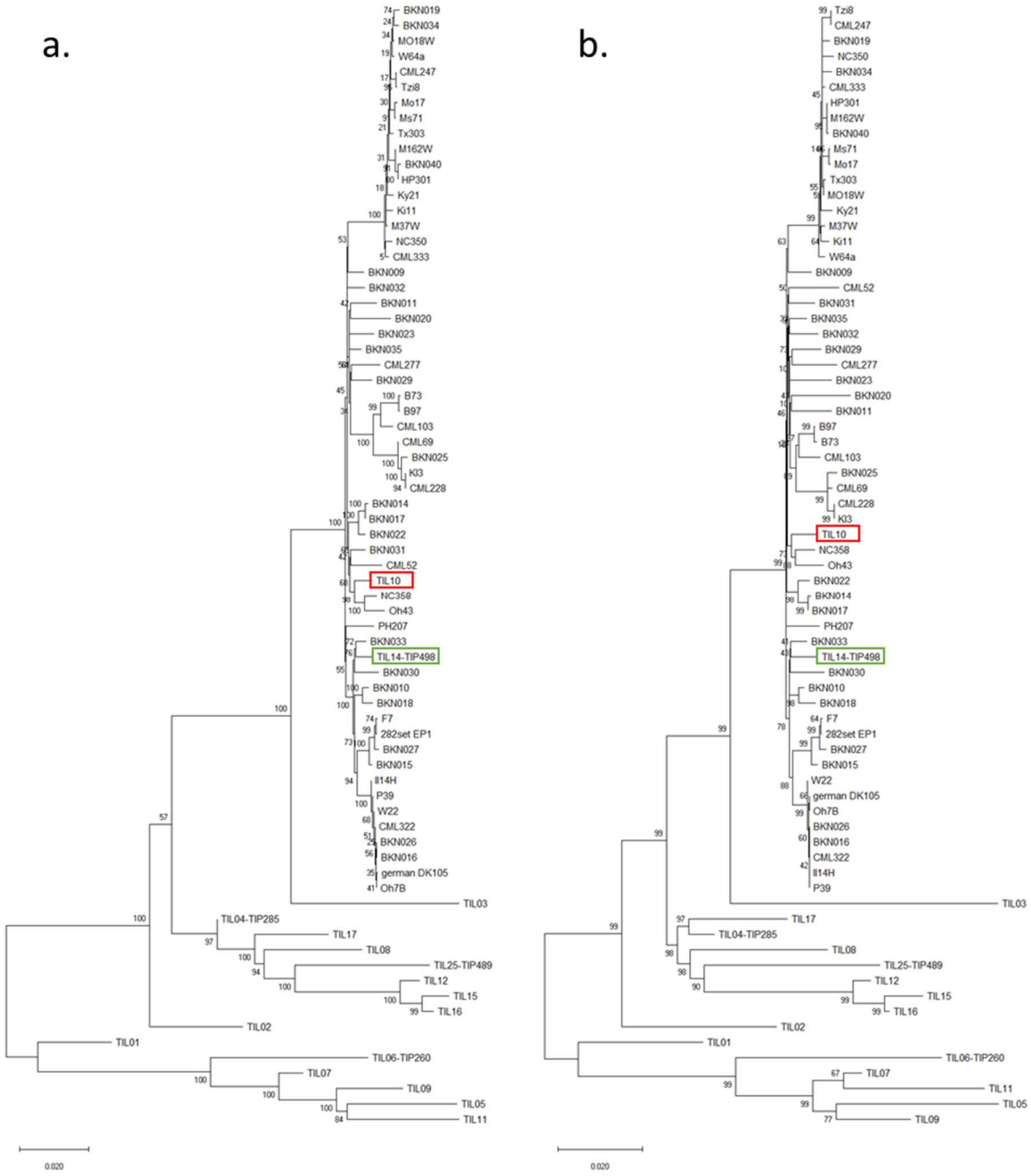


Figure S21. HapMap3 trees of the CEN2 domestication region

With ambiguities (**a**) and without ambiguities (**b**). TIL10 is boxed in red and TIL14 is boxed in green. There is no major difference in structure between the two trees suggesting the position of the TILs in the trees are not cause by an artifact of missing data.

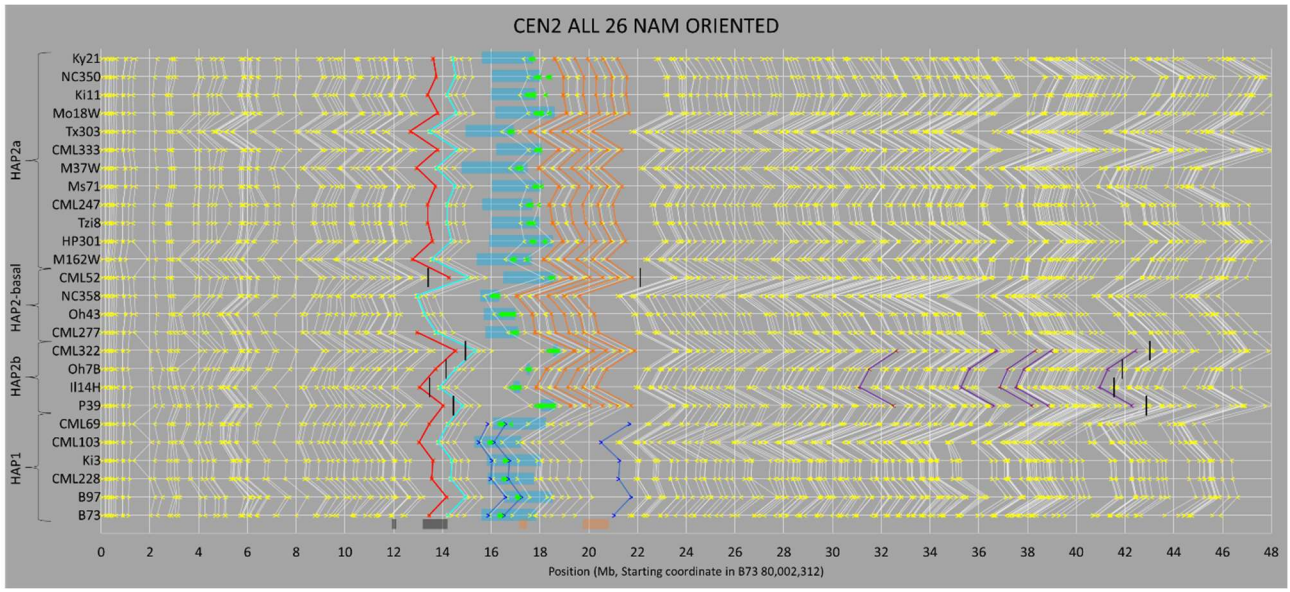


Figure S22. CEN2 gene models and orthologs among the NAM lines.

Orthologous genes among all 26 NAM lines were identified via blastn of B73 v5 gene models to the 26 physical maps and are shown as yellow markers connected by a white line. All NAM lines are zeroed at the start of Zm00001eb086050 (B73 v5 annotation). The arrows denote the orientation of the gene model. Active centromeres were identified via cenH3 ChIP-seq for each NAM line mapped to the respective physical map. Active centromere regions (light blue shade) were manually defined using cenH3 coverage charts. The active CEN was defined as the region where cenH3 coverage is ~3X higher than background. CentC repeats identified with blast are marked with light green arrows. The inversion breakpoints in the HAP2b lines and the unique CML52 pericentric inversion breakpoints are marked with black lines. HAP2 specific gene models and NC358/Oh43 gene deletions were identified via blast of Ky21 (HAP2A) gene models to the 26 NAM lines. HAP2 specific gene models and NC358/Oh43 specific gene deletions are shown as orange arrows connected by orange lines and red arrows connected by red lines respectively. The RALF gene is denoted with cyan arrows connected by a cyan line. The positions of the two HAP1 inversions are shaded orange beneath B73. NC358/Oh43 deletions are shaded black. HAP2B specific gene models were identified using P39 gene models to the 26 NAM lines. Blast results were sorted by line and coordinate and filtered to remove non-syntenic genes.

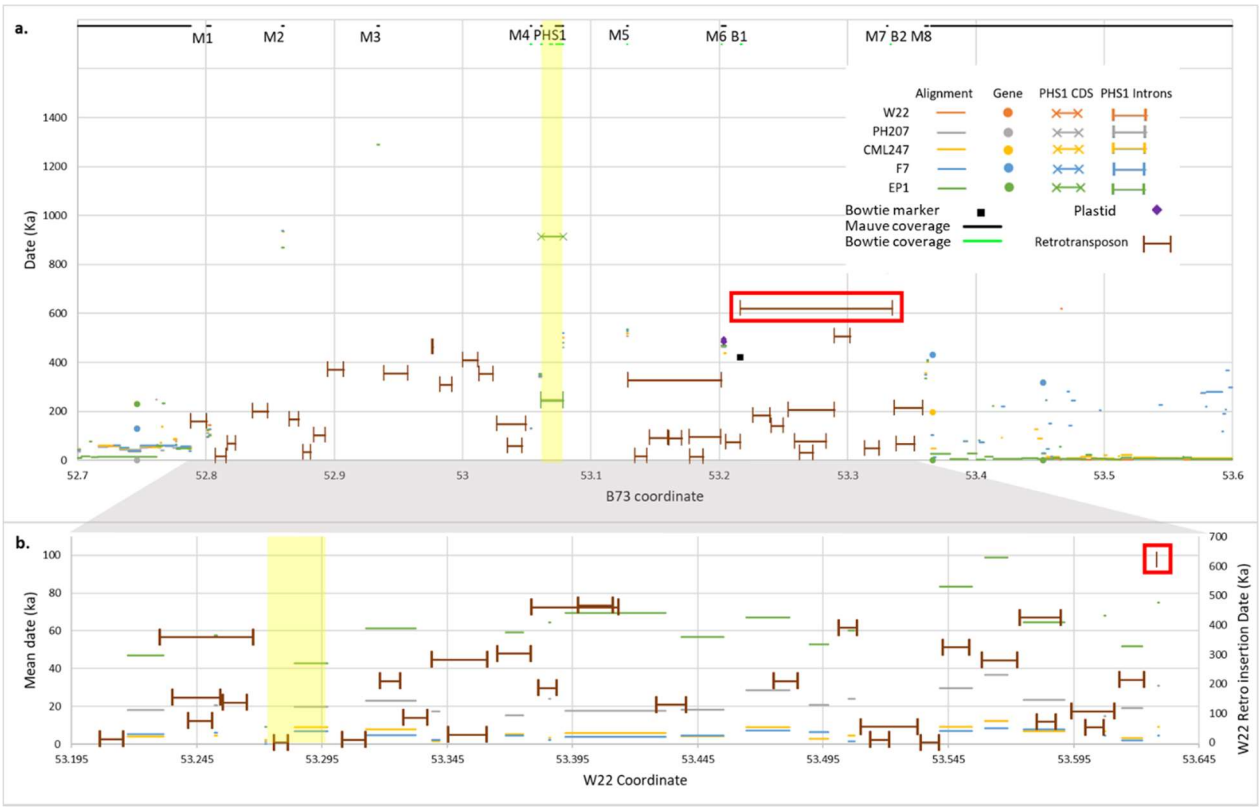


Figure S23. Schematic diagram of the introgressed region in the upstream pericentromere of B73 v4CEN.

(a.) Horizontal bars in orange, grey, yellow, blue and green represent the mean divergence date of 375nt alignments overlapping by 500nt for W22, PH207, CML247, F7 and EP1, respectively. Divergence of genes in W22, PH207, CML247, F7 and EP1 are represented by orange, grey, yellow, blue and green dots, respectively. The dots denote the B73 v4CEN start coordinates (Mb) of the gene. Black horizontal bars represent progressive Mauve alignment coverage. Mauve markers in the introgression are labeled M1 – M8. Bowtie alignments generated from unique 100mers are depicted with light green horizontal bars and are labeled B1 and B2. The yellow region demarcates the coding region of the introgressed PHS1 gene. Divergence dates for the PHS1 CDS are illustrated by bars flanked with X's in orange, grey, yellow, blue and green for W22, PH207, CML247, F7 and EP1, respectively (*SI Appendix Figure S24, Supplemental text SX8*). Divergence dates of concatenated introns for the PHS1 gene are marked as bars flanked by vertical lines using the same color scheme described for the CDS divergence dates. The plastid marker identified with Junction Viewer are marked with purple diamonds. Bowtie marker divergence dates for Bowtie markers 1 and 2 are marked with black boxes. The insertion time and start/end positions (Mb) of retrotransposons are denoted by brown dashes and vertical bars. **b.** Divergence date of non B73 genomes to W22 is <100 ka. Grey, yellow, blue and green bars reflect the divergence date of PH207, CML247, F7 and EP1, respectively on the W22 physical map (Mb) corresponding to the B73 introgressed region. Gaps of in coverage > 5kb are shown as spaces between divergence estimate bars. Red boxes indicate the LTR of the single shared retrotransposon.

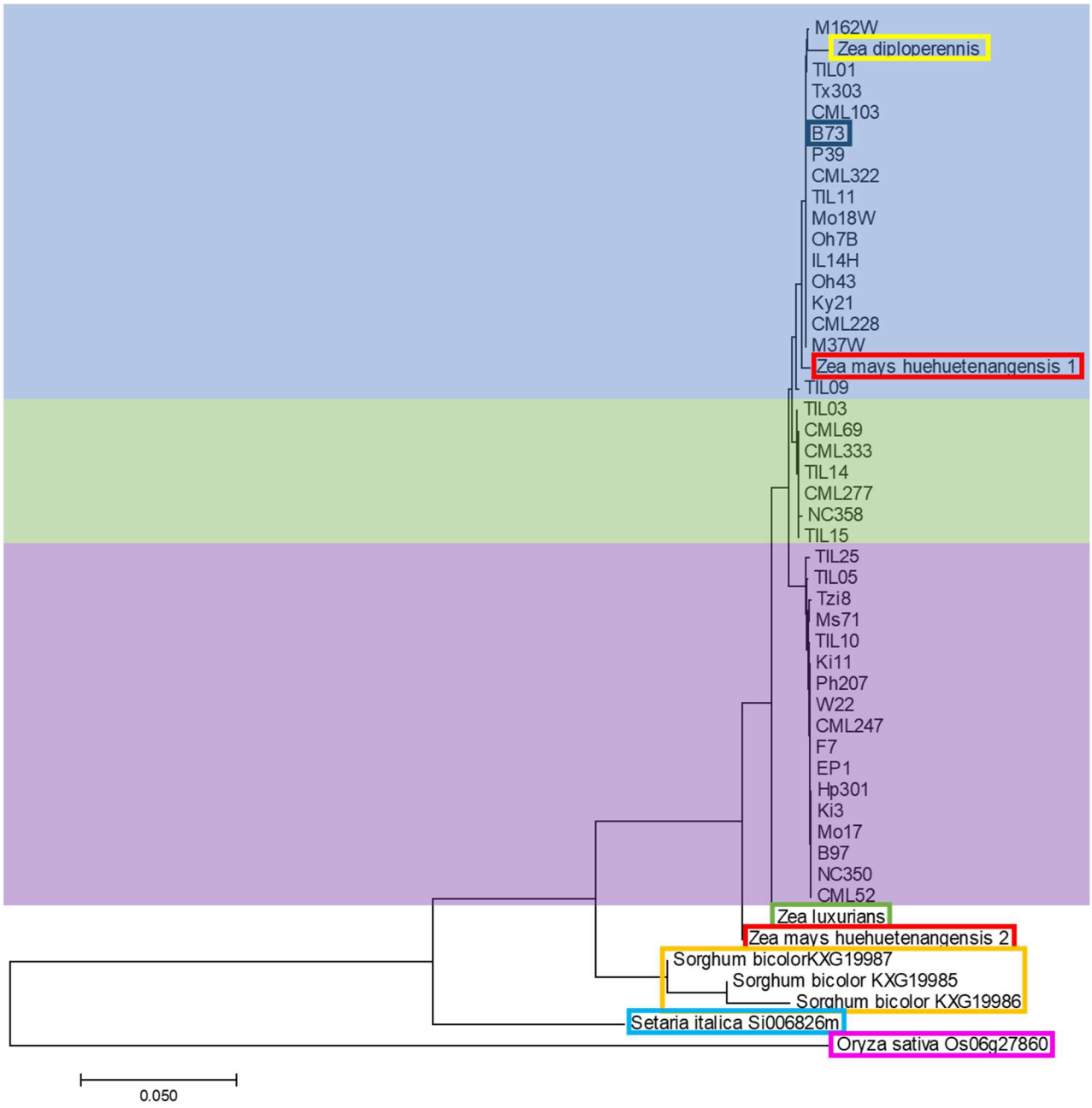


Figure S24. Phylogenetic tree of the introgressed PHS1 gene.

The highlighted clades correspond to the three PHS1 haplotypes. Blue reflects members of the B73 haplotype. Purple represents members of the W22 haplotype and Green represents members of the CML69 haplotype. Transcripts from the B73 v4 CDS file available on the Gramene database are represented as Zm00001d045993_T001-3. CDS extracted from RNA-seq data mapped to B73 v4CEN is represented as B73_Zm00001d04599. *Zea luxurians*, is boxed in green, *Zea diploperennis* is boxed in yellow and two lines of *Zea mays* ssp. *Huehuetenangensis* are boxed in red. *Sorghum bicolor*, *Setaria italica*, and *Oryza sativa japonica* PHS1 orthologs are boxed in orange light blue and pink respectively. For additional information see *SI Appendix Text SX9, Figure S25*.

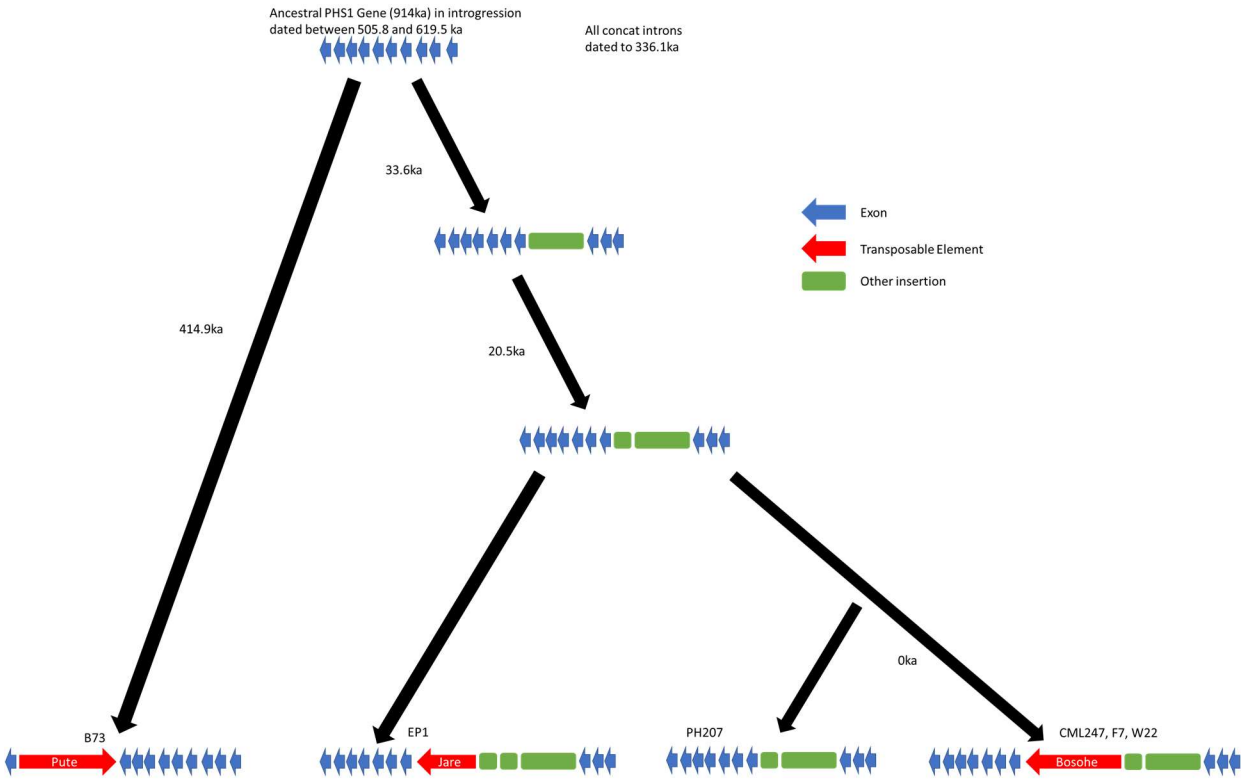


Figure S25. Schematic phylogeny of the CEN9 PHS1 gene.
Schematic representation of the PHS1 gene and the associate retrotransposon and non-retro insertions leading to four variants among the B73, CML247, EP1, F7, PH207 and W22 genomes.

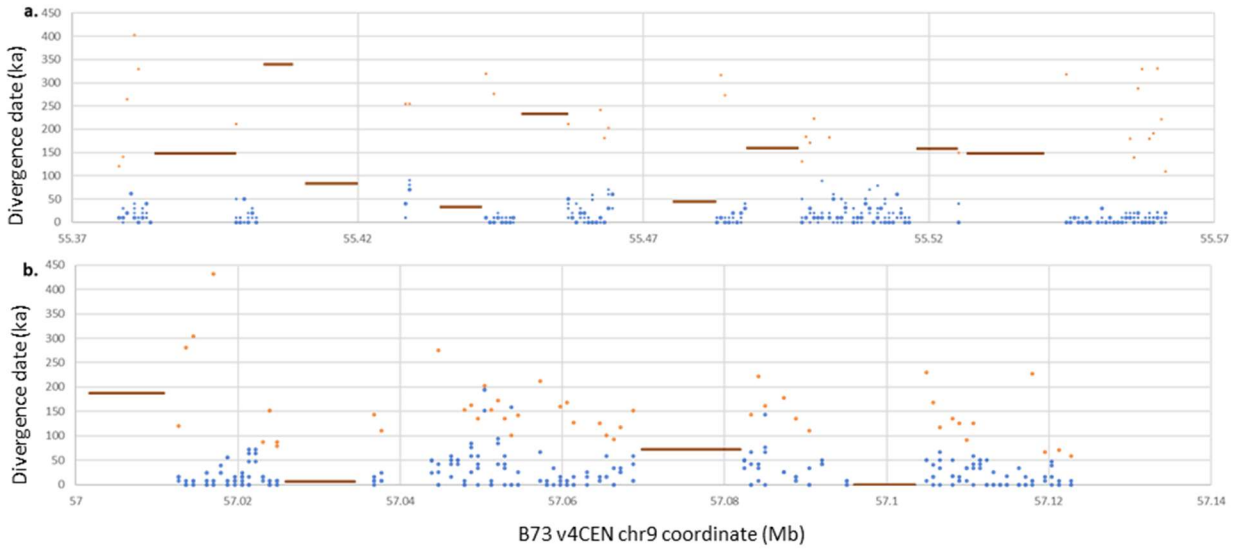


Figure S26. Schematic diagram of B73 v4CEN position and divergence dates of full-length monomers.
Monomers were identified with blastn in CTR1 LI (**a.**) and CTR1 LII (**b.**). Haplotype 2 (PH207) monomers are orange dots, Haplotype 1 monomers (B73, W22, CML247, F7 and EP1) are depicted with blue dots. Retrotransposon B73 v4CEN positions and insertion dates are demarcated with brown bars.

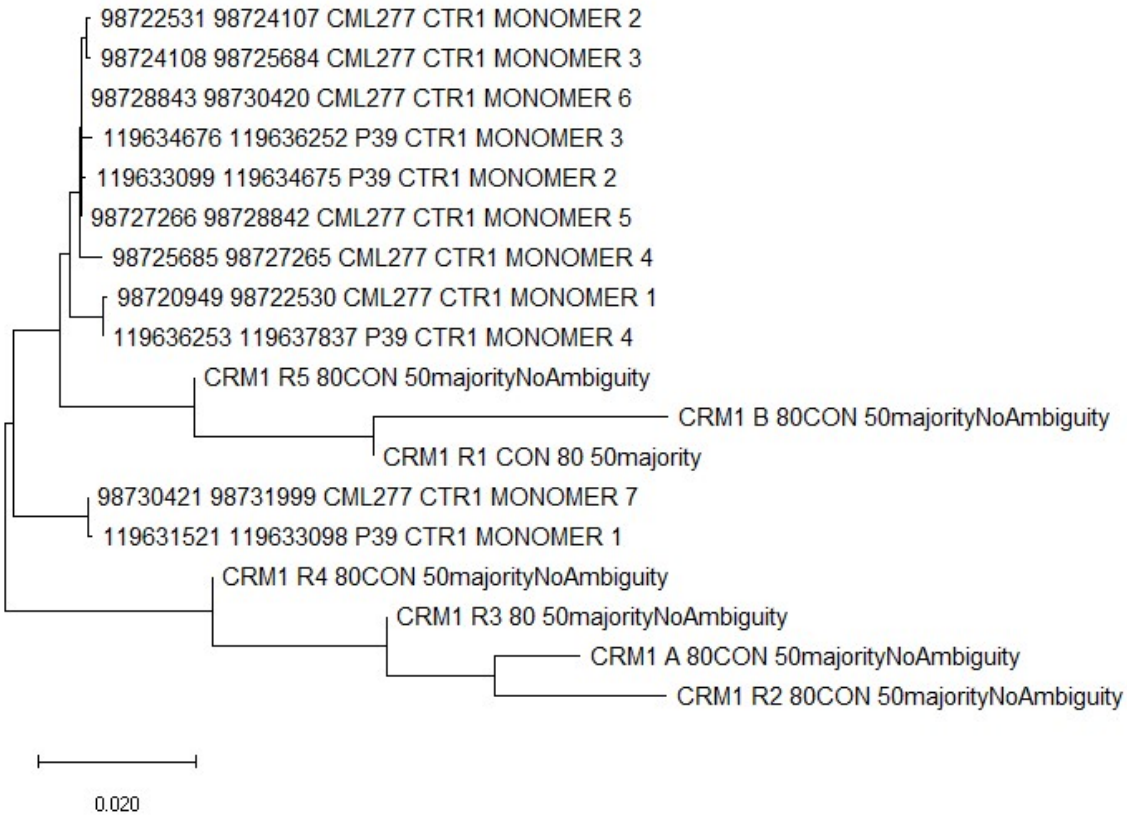


Figure S27. Phylogenetic tree of CEN2 CTR1 monomers to CR1 consensus sequences.

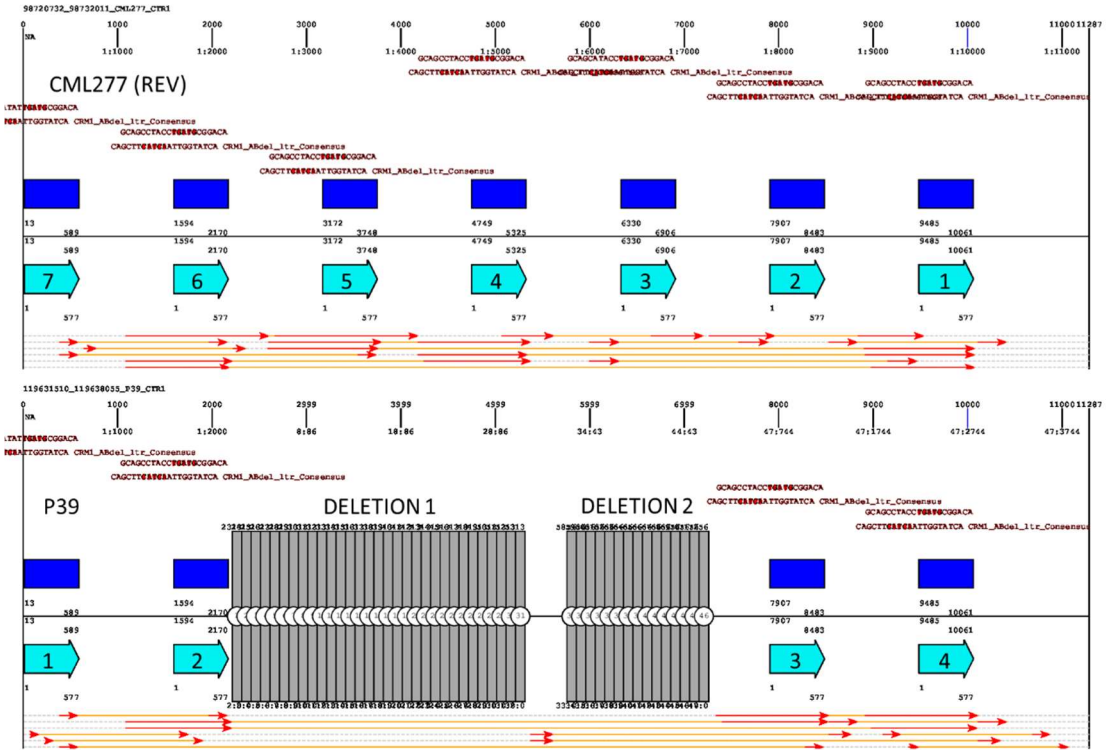


Figure S28. Junction viewer [108] annotation of the CEN2 CTR1 loci in CML277 (top) and P39 (bottom). P39 contains two potential deletions.

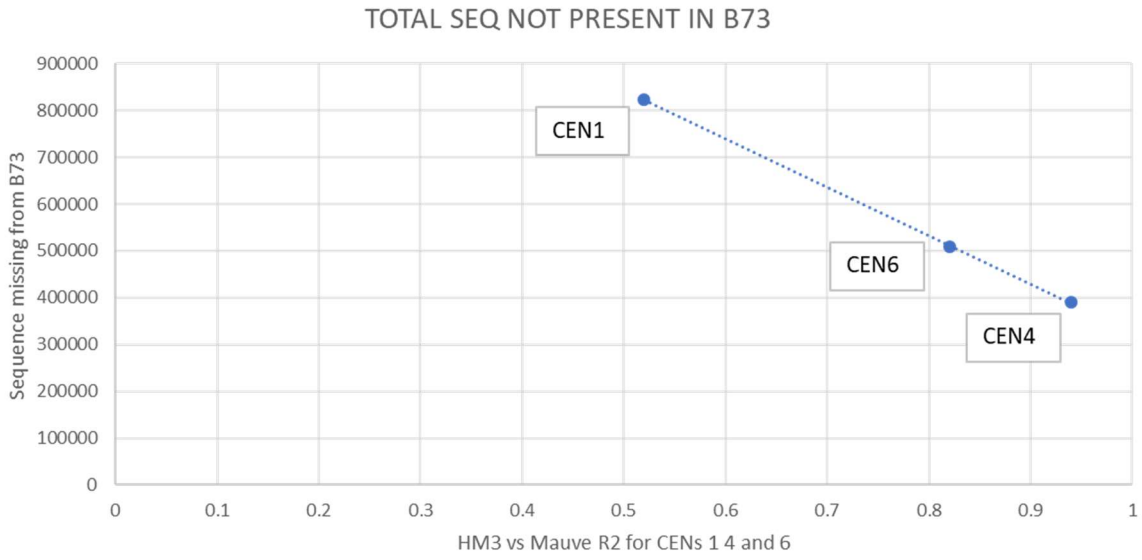
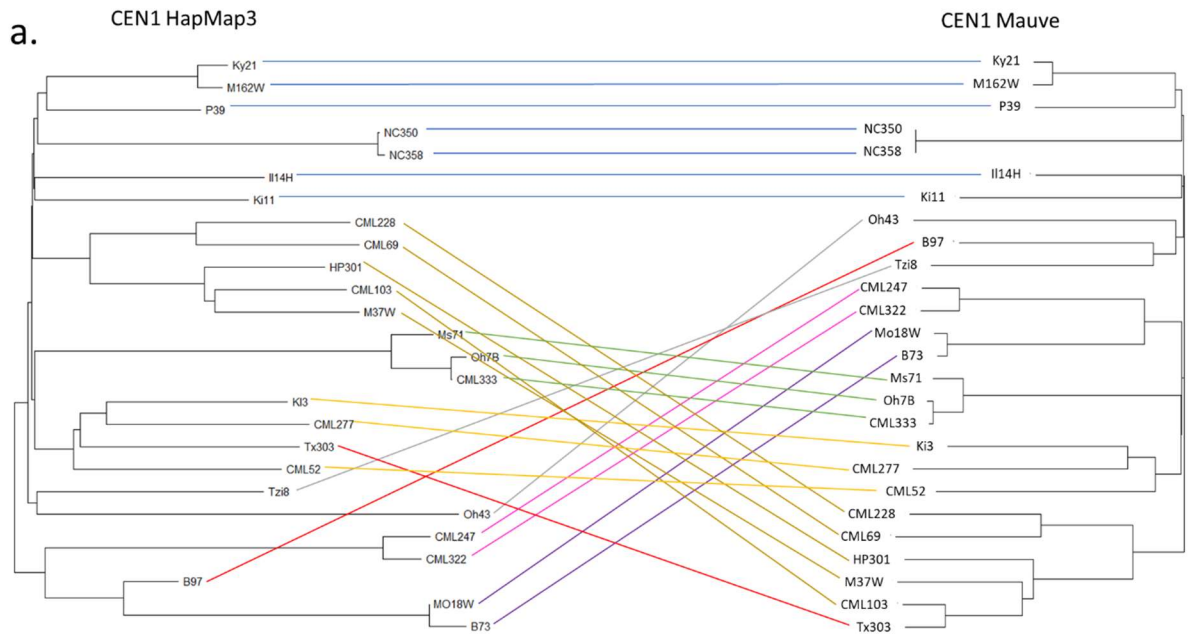
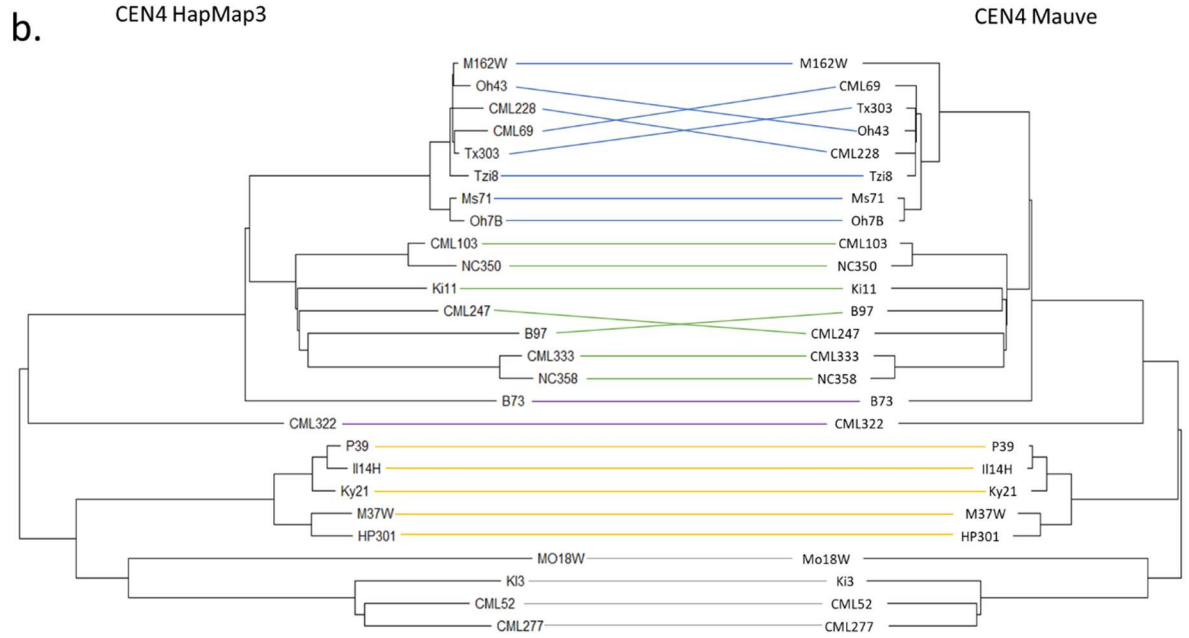


Figure S29. Correlation between sequence that is not present in B73 and the R2 value for Mauve/HapMap3 comparisons.

Missing sequence in B73 primarily corresponds to unique retrotransposon insertions not present in B73 but may also represent minor deletions/insertions. This value represents the difference between the alignment length and the B73 sequence length.


a.
CEN1 HapMap3
CEN1 Mauve
Ky21
M162W
P39
NC350
NC358
Il14H
Ki11
CML228
CML69
HP301
CML103
M37W
Ms71
Oh7B
CML333
Ki3
CML277
Tx303
CML52
Tzi8
Oh43
CML247
CML322
B97
MO18W
B73
Mo18W


b.
CEN4 HapMap3
CEN4 Mauve
M162W
Oh43
CML228
CML69
Tx303
Tzi8
Ms71
Oh7B
CML103
NC350
Ki11
CML247
B97
CML333
NC358
B73
CML322
P39
Il14H
Ky21
M37W
HP301
MO18W
Mo18W
Ki3
CML52
CML277

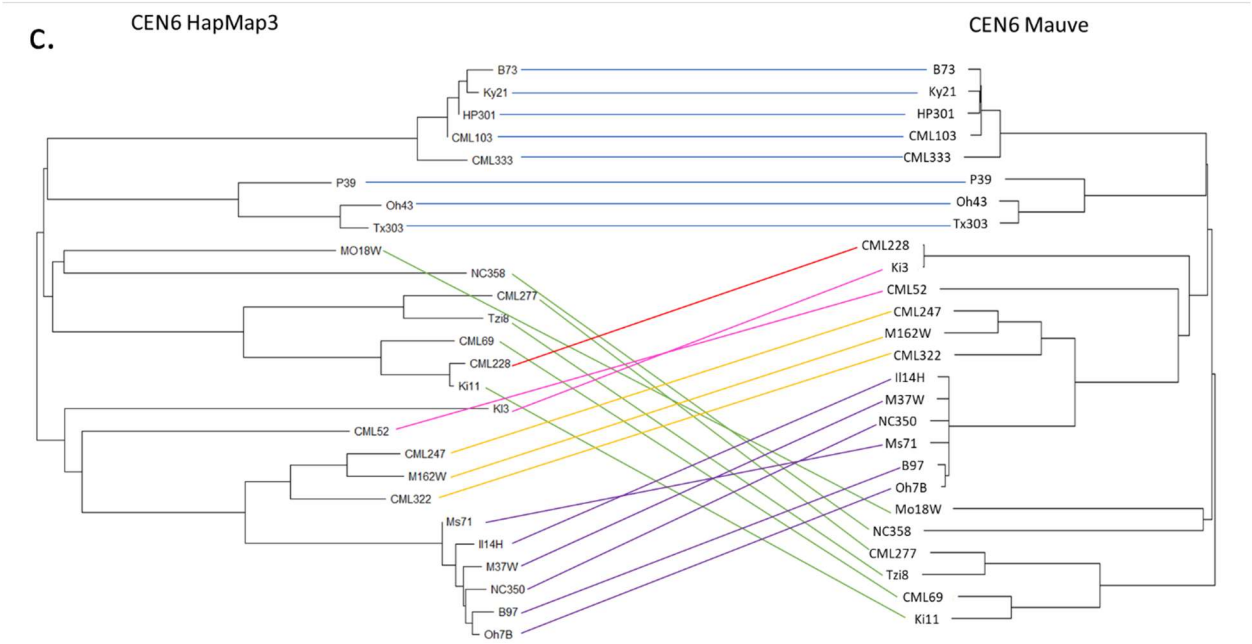


Figure S30. Comparison of phylogenetic trees for matching regions for HapMap3 and Mauve in the CEN1 (a), CEN4 (b.) and CEN6 (c) alignments.

Red lines connect NAM lines that have changed position in the tree between the physical maps and HapMap3 data. Discrepancies were found in the position of CML228 in CEN6 and, Tx303 and B97 in CEN1 between HapMap3 and the physical maps.

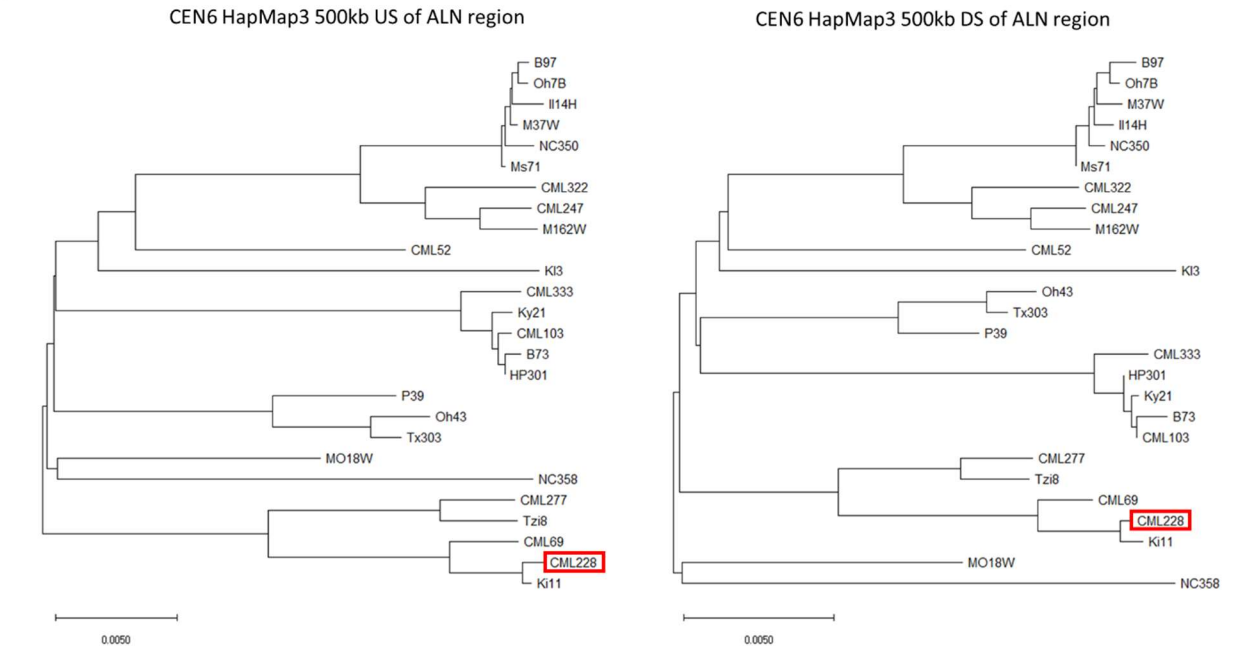


Figure S31. Comparison of trees corresponding to the 500kb upstream (US) and downstream (DS) of the CEN6 Mauve alignment region.

We observe that the position of CML228 remains the same relative to Ki11 in both trees suggesting the discrepancy with the physical map alignment is not a local issue.

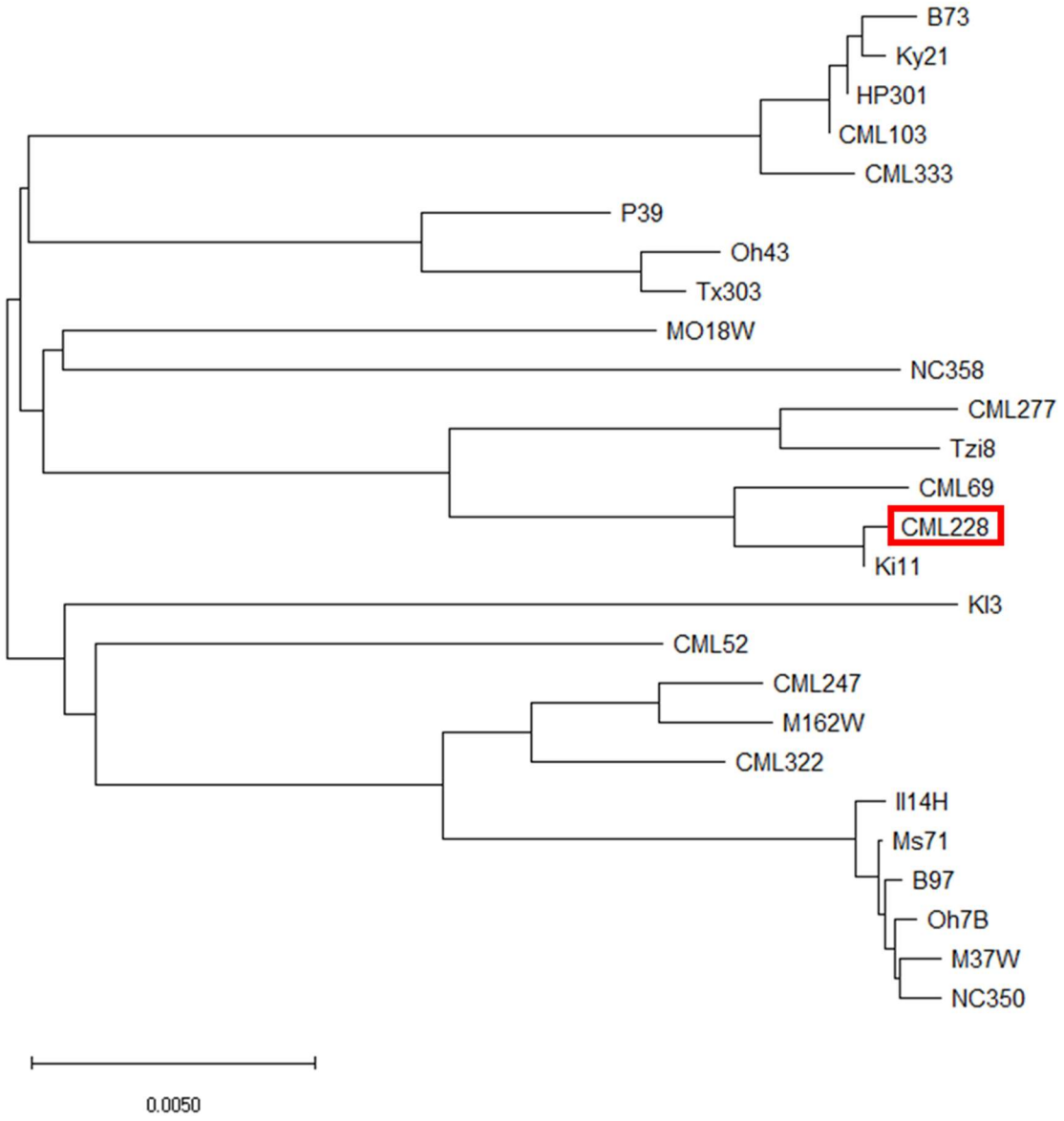


Figure S32. Phylogenetic tree of unimputed HapMap3 data over the CEN6 alignment region.
These data still show CML228 grouped with Ki11 suggesting that the discrepancy with the physical alignments is not due to SNP imputation.

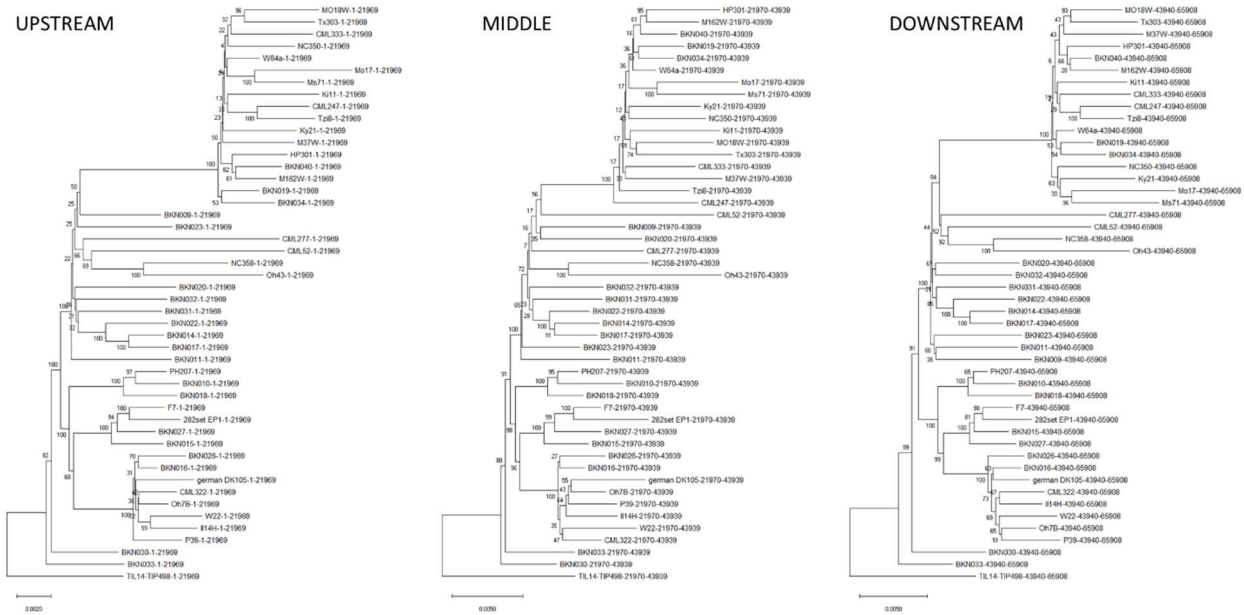


Figure S33. CEN2 HapMap3 Subtrees

(bootstrapped 1000 replicates) for HAP2 over three subregions of REGION13_SYN. Three trees were generated from this region to verify that there was no regional variation in tree structure across the region. Major clades stay together though there is variation in the distribution of lines within each major clade indicating the positions of lines within the major HAP2 clades are lower confidence. This is also reflected in the lower bootstrap values near the base of the major HAP2 subclades.

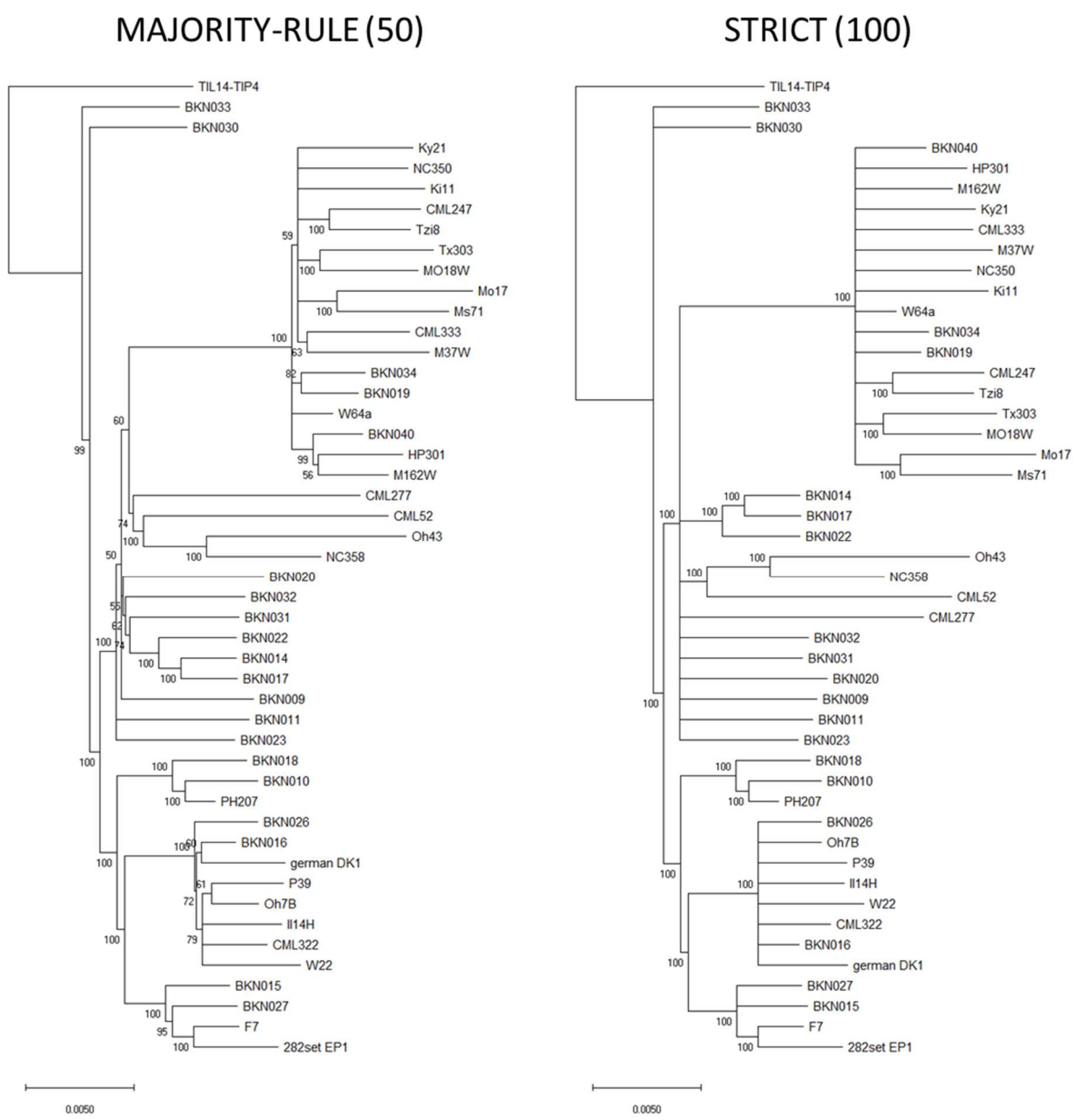


Figure S34. CEN2 Bootstrap consensus trees

(1000 replicates) for HAP2 over REGION13_SYN (HapMap3 data). Consensus thresholds of 50% (Majority-rule) and 100% (Strict) were used to collapse weak clades. Estimated branch lengths are shown over each branch and bootstrap values above the minimum threshold are positioned above the nodes.

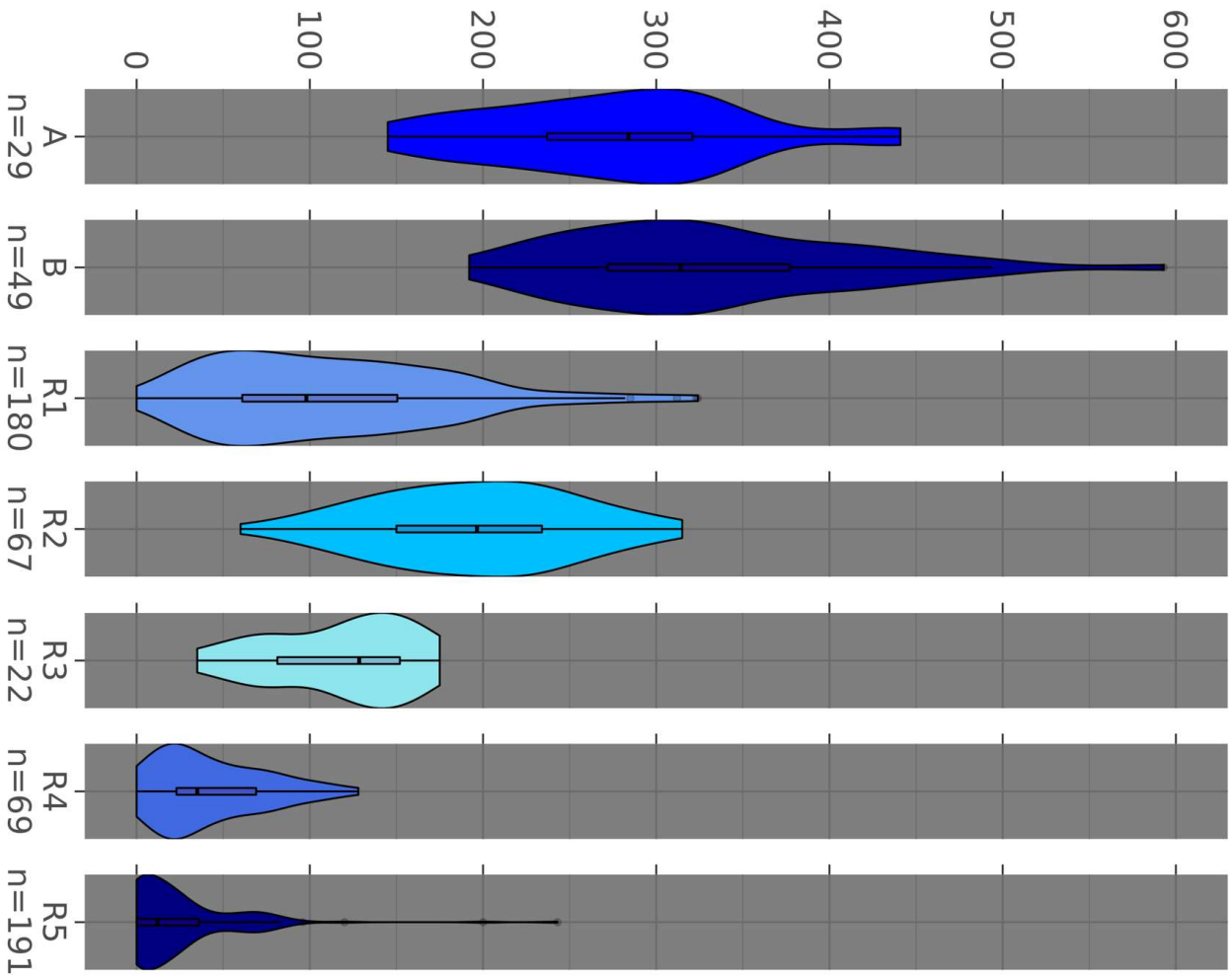


Figure S35. B73 v5 CR1 insertion date violin plot by subtype.
Each CR1 subtype A, B and recombinants R1-R5

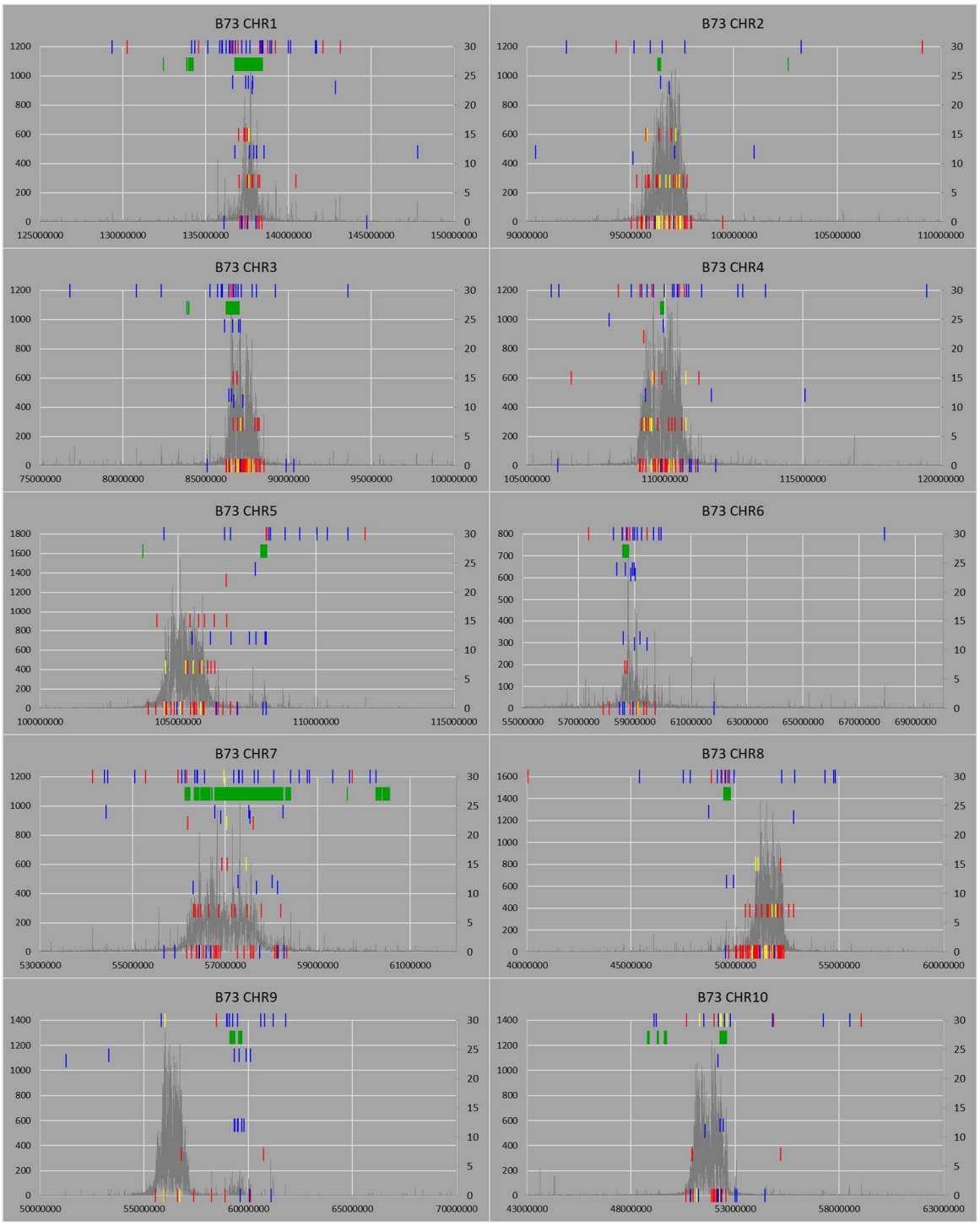


Figure S36. Distribution of UMRs across the B73 v5 genome.

The position and insertion date of CR2 elements and CR1 elements are marked in red and blue respectively on the secondary Y axis (right). Any element with an insertion date >=30ka appears at the top of the chart. Those CR2s containing a UMR are marked in yellow. The centC repeat is marked in green and cenH3 raw coverage is marked as a grey line on the primary y axis (left). We observe a strong correlation between UMRs and the active centromeres regardless of whether the chromosome contains an ancestral (CEN1) or neocentromere (CEN5).

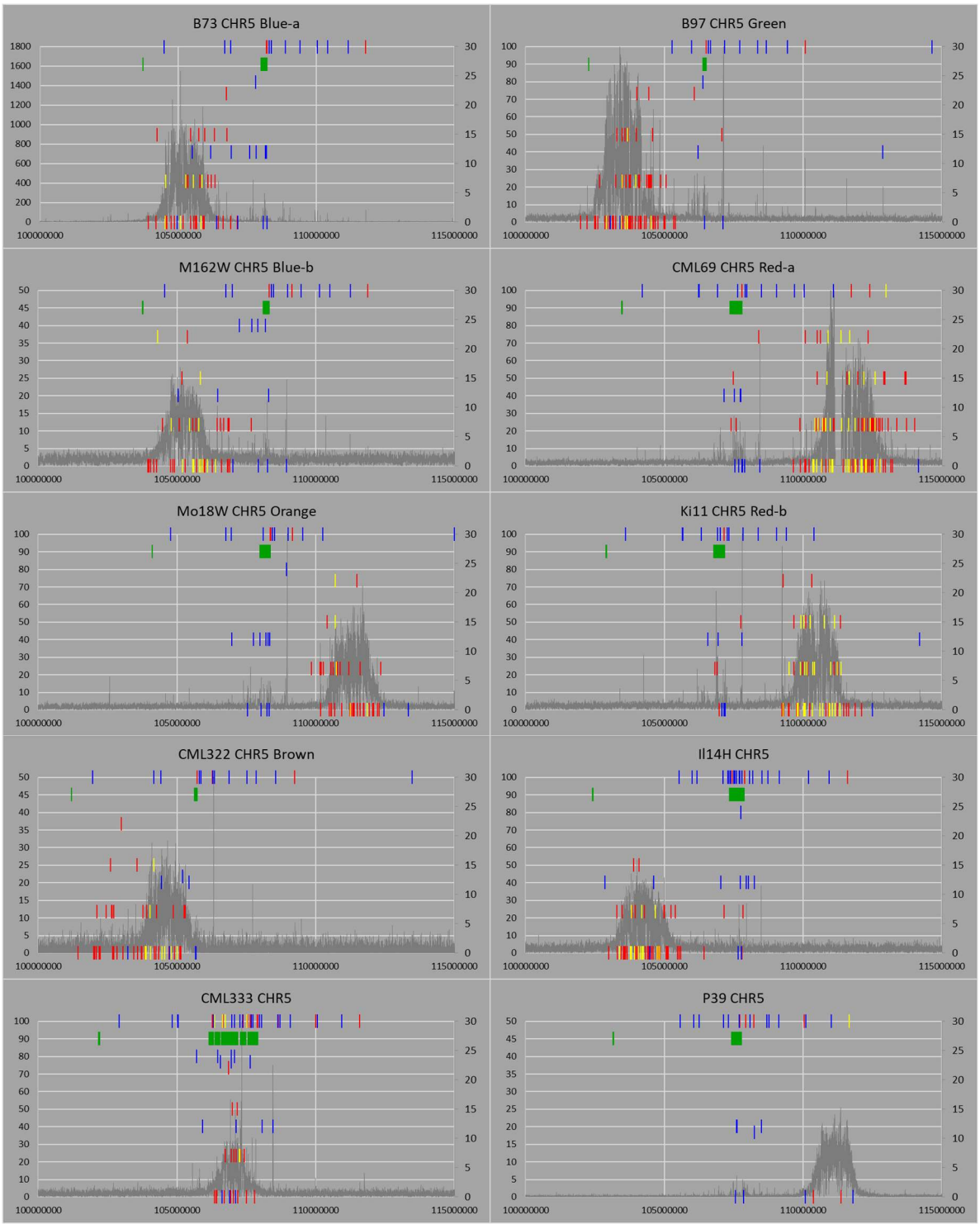


Figure S37. Distribution of UMRs across the CEN5 lineages.

The position and insertion date of CR2 elements and CR1 elements are marked in red and blue respectively on the secondary Y axis (right). Any element with an insertion date >=30ka appears at the top of the chart. Those CR2s containing a UMR are marked in yellow. The centC repeat is marked in green and cenH3 raw coverage is marked as a grey line on the primary y axis (left). We observe a strong correlation between UMRs and the active centromeres regardless of neocentromere position or lineage.

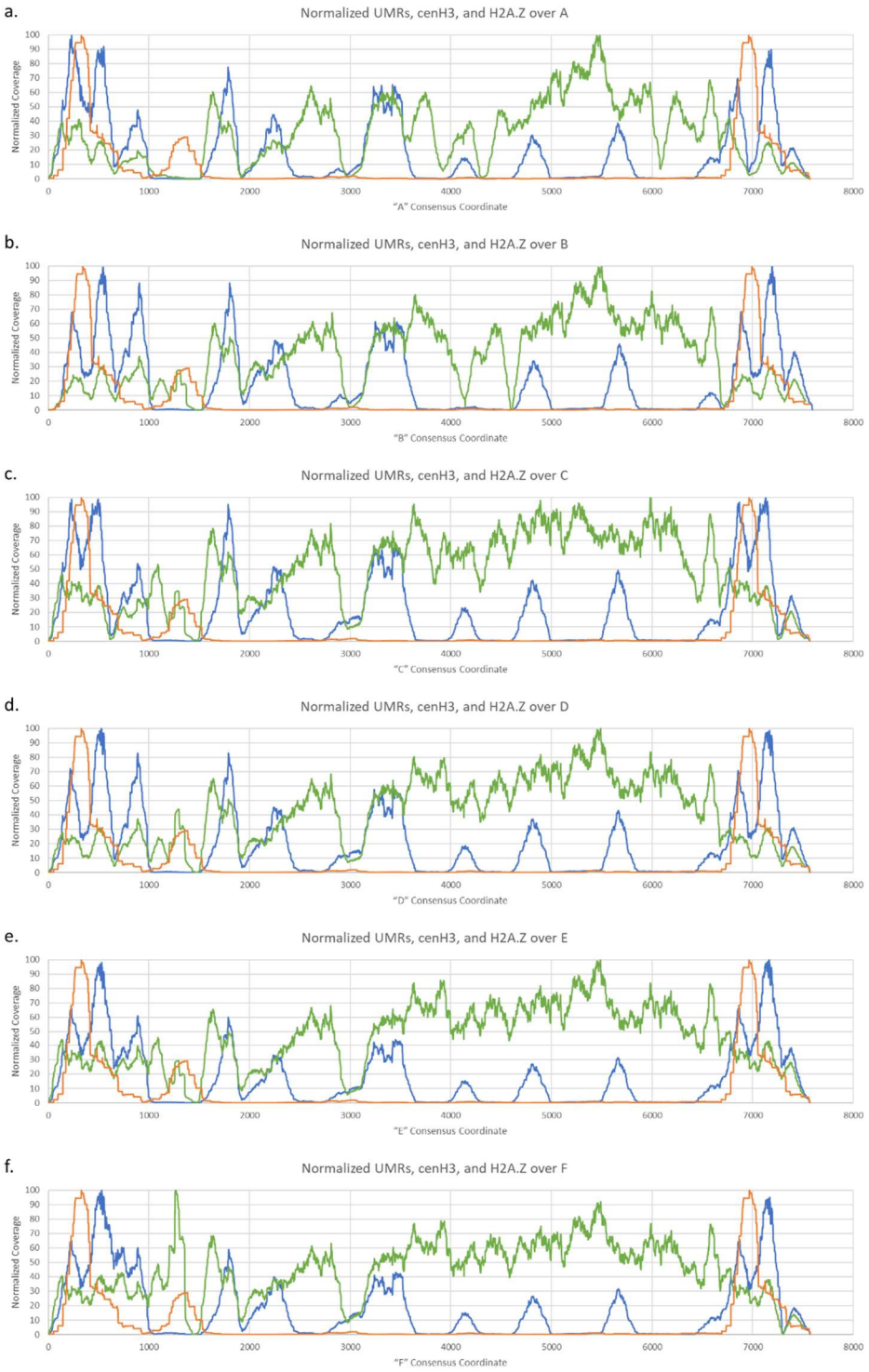


Figure S38. UMR, H2A.Z and cenH3 coverage for CR2 parental subtypes.
A schematic of major functional domains for CR2 parental subtypes A-F (**a.-f. respectively**) are shown above the respective charts. UMRs found within the terminal coordinates of CR2s across the NAM lines were extracted from their respective genomes and blasted to each parental subtype. The number of

UMRs aligned to each position was then counted and normalized to 100. UMR coverage (orange) displays three peaks one in each LTR and one in the UTR in CR2 which is consistent across all parents. H2A.Z and cenH3 coverage (B73) were obtained via bowtie mapping to the CR2 parental consensus sequences and normalized to 100. H2A.Z peaks (blue) are found in the LTRs and across the polyprotein. CenH3 coverage (green) is variable but much of the element is covered which is to be expected for centromere targeting retroelements.

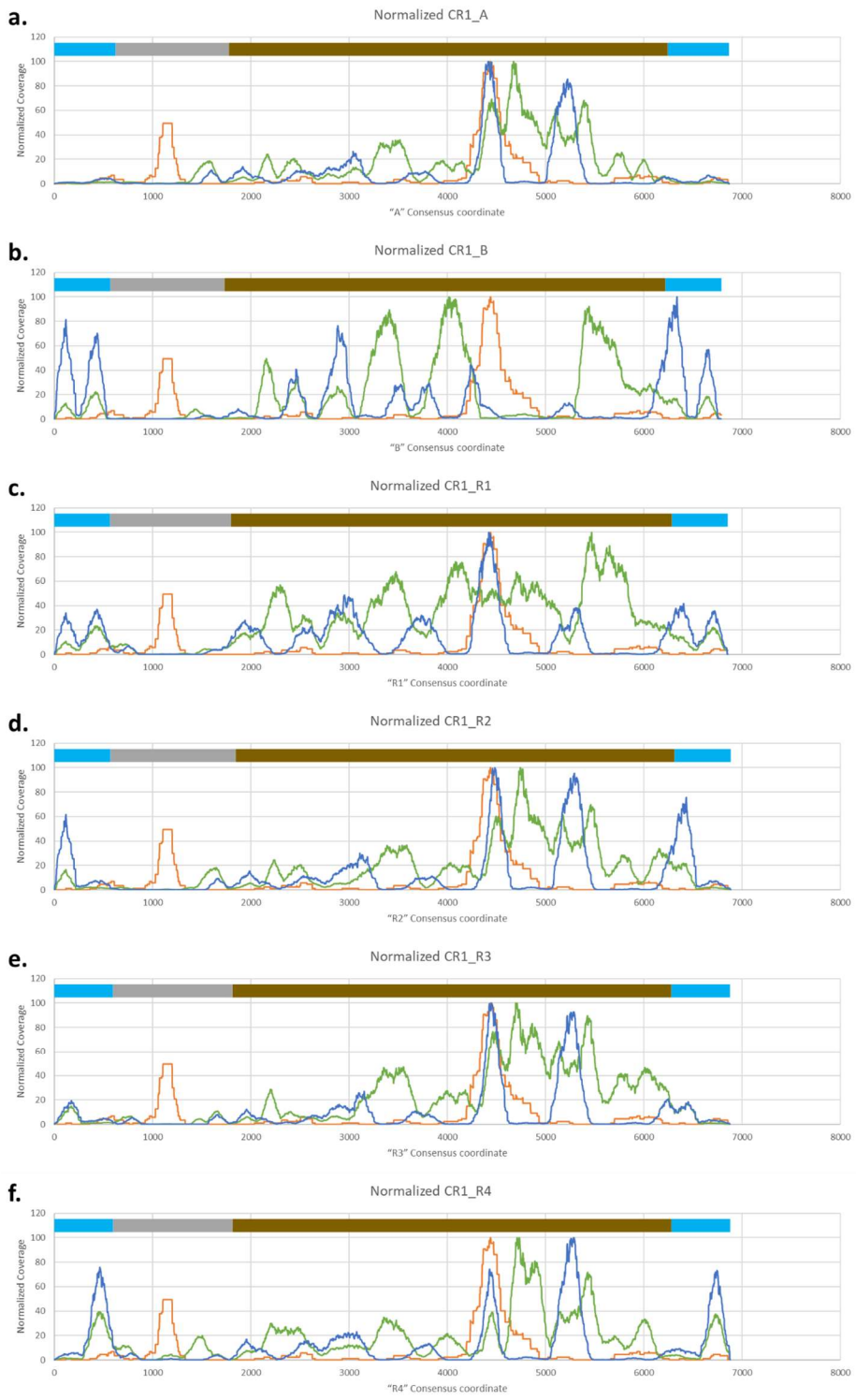


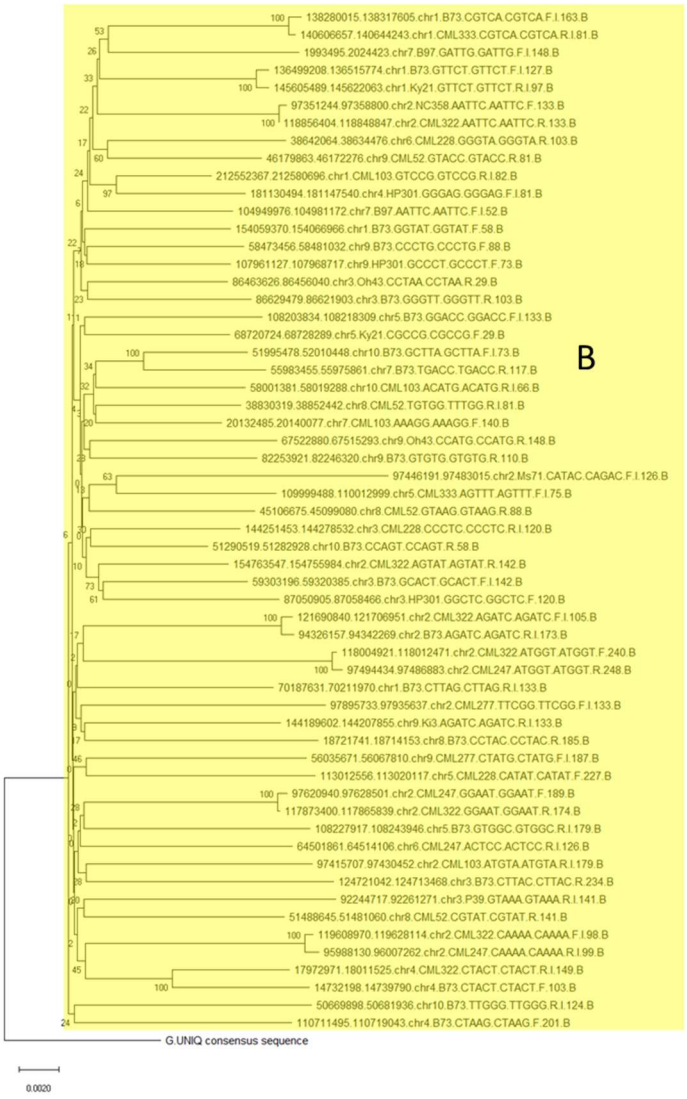
Figure S39. UMR, H2A.Z and cenH3 coverage for CR1 subtypes.

A schematic of major functional domains for CR1 subtypes A-R4 (**a.-f. respectively**) are shown above the respective charts. UMRs found within the terminal coordinates of CR1s across the NAM lines were extracted from their respective genomes and blasted to each parental subtype. The number of UMRs aligned to each position was then counted and normalized to 100. There are far fewer CR1s associated

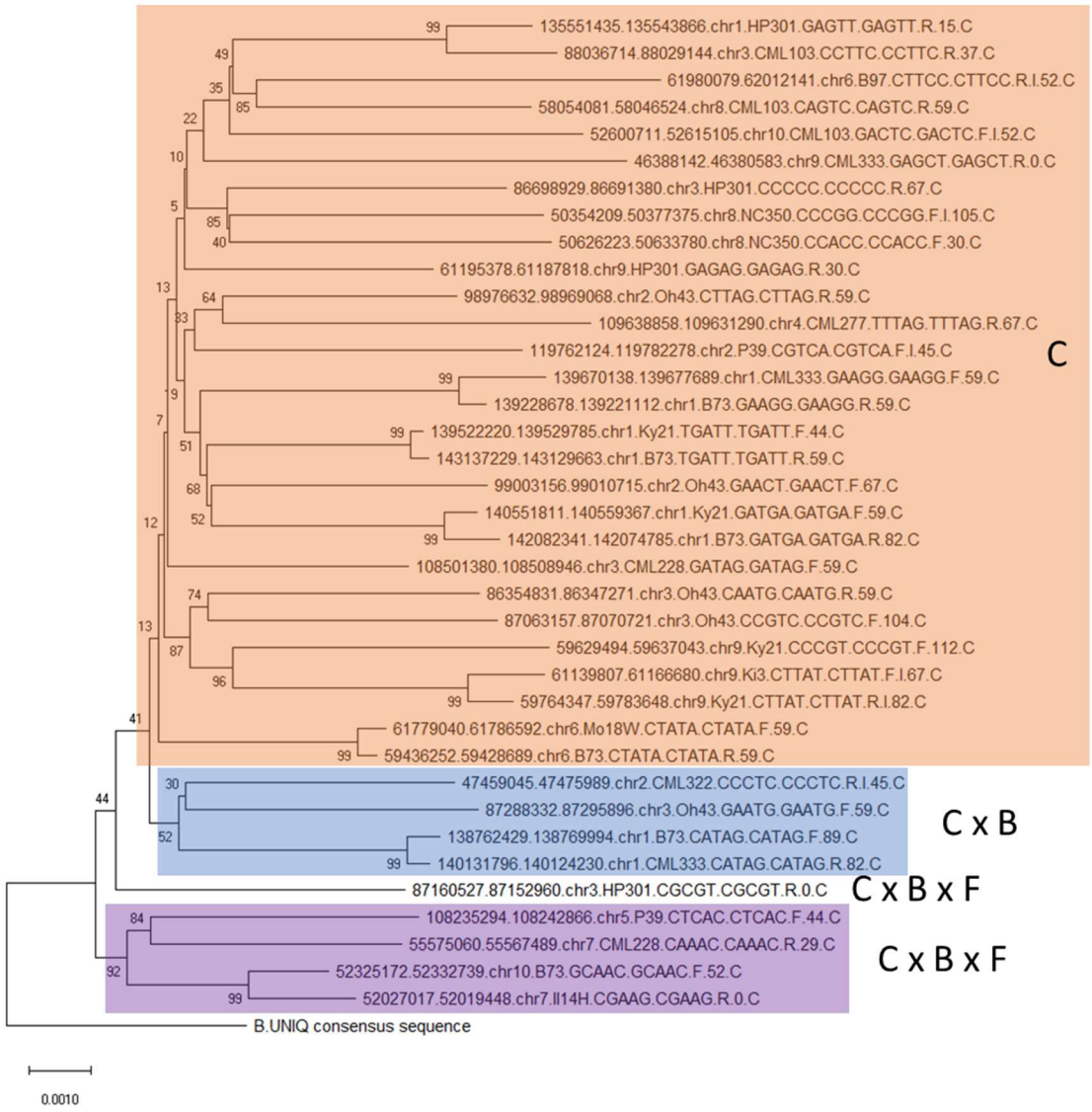
with UMRs than CR2s. UMR coverage (orange) displays three peaks one in each UTR and one in the PP in CR1 which is consistent across all subtypes. H2A.Z and cenH3 coverage (B73) were obtained via bowtie mapping to the CR1 subtype consensus sequences and normalized to 100. H2A.Z peaks (blue) are found in the LTRs and across the polyprotein. CenH3 coverage (green) is variable but much of the element is covered which is to be expected for centromere targeting retroelements. The cenH3 coverage is lower in CR1 due to the higher fraction of CR1 elements found outside active centromeres (*SI Appendix Table ST6*).

a.

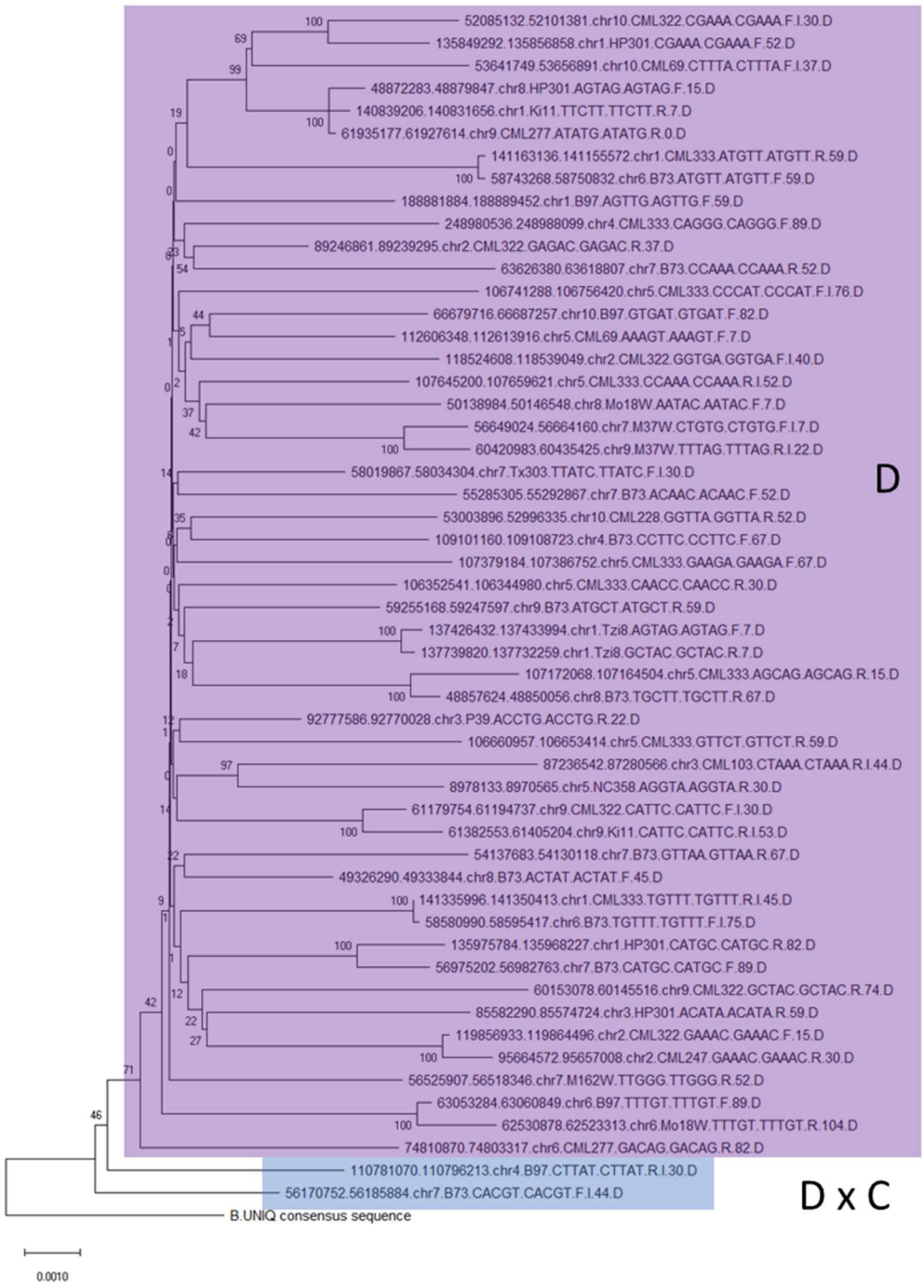
b.


138280015.138317605.chr1.B73.CGTCA.CGTCA.F.I.163.B
140606657.140644243.chr1.CML333.CGTCA.CGTCA.R.I.81.B
1993495.2024423.chr7.B97.GATTG.GATTG.F.I.148.B
136499208.136515774.chr1.B73.GTTCT.GTTCT.F.I.127.B
145605489.145622063.chr1.Ky21.GTTCT.GTTCT.R.I.97.B
97351244.97358800.chr2.NC358.AATTC.AATTC.F.133.B
118856404.118848847.chr2.CML322.AATTC.AATTC.R.133.B
38642064.38634476.chr6.CML228.GGGTA.GGGTA.R.103.B
46179863.46172276.chr9.CML52.GTACC.GTACC.R.81.B
212552367.212580696.chr1.CML103.GTCCG.GTCCG.R.I.82.B
181130494.181147540.chr4.HP301.GGGAG.GGGAG.F.I.81.B
104949976.104981172.chr7.B97.AATTC.AATTC.F.I.52.B
154059370.154066966.chr1.B73.GGTAT.GGTAT.F.58.B
58473456.58481032.chr9.B73.CCCTG.CCCTG.F.88.B
107961127.107968717.chr9.HP301.GCCCT.GCCCT.F.73.B
86463626.86456040.chr3.Oh43.CCTAA.CCTAA.R.29.B
86629479.86621903.chr3.B73.GGGTT.GGGTT.R.103.B
108203834.108218309.chr5.B73.GGACC.GGACC.F.I.133.B
68720724.68728289.chr5.Ky21.CGCCG.CGCCG.F.29.B
51995478.52010448.chr10.B73.GCTTA.GCTTA.F.I.73.B
55983455.55975861.chr7.B73.TGACC.TGACC.R.117.B
58001381.58019288.chr10.CML103.ACATG.ACATG.R.I.66.B
38830319.38852442.chr8.CML52.TGTGG.TTTGG.R.I.81.B
20132485.20140077.chr7.CML103.AAAGG.AAAGG.F.140.B
67522880.67515293.chr9.Oh43.CCATG.CCATG.R.148.B
82253921.82246320.chr9.B73.GTGTG.GTGTG.R.110.B
97446191.97483015.chr2.Ms71.CATAC.CAGAC.F.I.126.B
109999488.110012999.chr5.CML333.AGTTT.AGTTT.F.I.75.B
45106675.45099080.chr8.CML52.GTAAG.GTAAG.R.88.B
144251453.144278532.chr3.CML228.CCCTC.CCCTC.R.I.120.B
51290519.51282928.chr10.B73.CCAGT.CCAGT.R.58.B
154763547.154755984.chr2.CML322.AGTAT.AGTAT.R.142.B
59303196.59320385.chr3.B73.GCACT.GCACT.F.I.142.B
87050905.87058466.chr3.HP301.GGCTC.GGCTC.F.120.B
121690840.121706951.chr2.CML322.AGATC.AGATC.F.I.105.B
94326157.94342269.chr2.B73.AGATC.AGATC.R.I.173.B
118004921.118012471.chr2.CML322.ATGGT.ATGGT.F.240.B
97494434.97486883.chr2.CML247.ATGGT.ATGGT.R.248.B
70187631.70211970.chr1.B73.CTTAG.CTTAG.R.I.133.B
97895733.97935637.chr2.CML277.TTCGG.TTCGG.F.I.133.B
144189602.144207855.chr9.Ki3.AGATC.AGATC.R.I.133.B
18721741.18714153.chr8.B73.CCTAC.CCTAC.R.185.B
56035671.56067810.chr9.CML277.CTATG.CTATG.F.I.187.B
113012556.113020117.chr5.CML228.CATAT.CATAT.F.227.B
97620940.97628501.chr2.CML247.GGAAT.GGAAT.F.189.B
117873400.117865839.chr2.CML322.GGAAT.GGAAT.R.174.B
108227917.108243946.chr5.B73.GTGGC.GTGGC.R.I.179.B
64501861.64514106.chr6.CML247.ACTCC.ACTCC.R.I.126.B
97415707.97430452.chr2.CML103.ATGTA.ATGTA.R.I.179.B
124721042.124713468.chr3.B73.CTTAC.CTTAC.R.234.B
92244717.92261271.chr3.P39.GTAAA.GTAAA.R.I.141.B
51488645.51481060.chr8.CML52.CGTAT.CGTAT.R.141.B
119608970.119628114.chr2.CML322.CAAAA.CAAAA.F.I.98.B
95988130.96007262.chr2.CML247.CAAAA.CAAAA.R.I.99.B
17972971.18011525.chr4.CML322.CTACT.CTACT.R.I.149.B
14732198.14739790.chr4.B73.CTACT.CTACT.F.103.B
50669898.50681936.chr10.B73.TTGGG.TTGGG.R.I.124.B
110711495.110719043.chr4.B73.CTAAG.CTAAG.F.201.B
B
G.UNIQ consensus sequence
0.0020

c.


135551435.135543866.chr1.HP301.GAGTT.GAGTT.R.15.C
88036714.88029144.chr3.CML103.CCTTC.CCTTC.R.37.C
61980079.62012141.chr6.B97.CTTCC.CTTCC.R.I.52.C
58054081.58046524.chr8.CML103.CAGTC.CAGTC.R.59.C
52600711.52615105.chr10.CML103.GACTC.GACTC.F.I.52.C
46388142.46380583.chr9.CML333.GAGCT.GAGCT.R.0.C
86698929.86691380.chr3.HP301.CCCCC.CCCCC.R.67.C
50354209.50377375.chr8.NC350.CCCGG.CCCGG.F.I.105.C
50626223.50633780.chr8.NC350.CCACC.CCACC.F.30.C
61195378.61187818.chr9.HP301.GAGAG.GAGAG.R.30.C
98976632.98969068.chr2.Oh43.CTTAG.CTTAG.R.59.C
109638858.109631290.chr4.CML277.TTTAG.TTTAG.R.67.C
119762124.119782278.chr2.P39.CGTCA.CGTCA.F.I.45.C
C
139670138.139677689.chr1.CML333.GAAGG.GAAGG.F.59.C
139228678.139221112.chr1.B73.GAAGG.GAAGG.R.59.C
139522220.139529785.chr1.Ky21.TGATT.TGATT.F.44.C
143137229.143129663.chr1.B73.TGATT.TGATT.R.59.C
99003156.99010715.chr2.Oh43.GAACT.GAACT.F.67.C
140551811.140559367.chr1.Ky21.GATGA.GATGA.F.59.C
142082341.142074785.chr1.B73.GATGA.GATGA.R.82.C
108501380.108508946.chr3.CML228.GATAG.GATAG.F.59.C
86354831.86347271.chr3.Oh43.CAATG.CAATG.R.59.C
87063157.87070721.chr3.Oh43.CCGTC.CCGTC.F.104.C
59629494.59637043.chr9.Ky21.CCCGT.CCCGT.F.112.C
61139807.61166680.chr9.Ki3.CTTAT.CTTAT.F.I.67.C
59764347.59783648.chr9.Ky21.CTTAT.CTTAT.R.I.82.C
61779040.61786592.chr6.Mo18W.CTATA.CTATA.F.59.C
59436252.59428689.chr6.B73.CTATA.CTATA.R.59.C
47459045.47475989.chr2.CML322.CCCTC.CCCTC.R.I.45.C
87288332.87295896.chr3.Oh43.GAATG.GAATG.F.59.C
C x B
138762429.138769994.chr1.B73.CATAG.CATAG.F.89.C
140131796.140124230.chr1.CML333.CATAG.CATAG.R.82.C
87160527.87152960.chr3.HP301.CGCGT.CGCGT.R.0.C
C x B x F
108235294.108242866.chr5.P39.CTCAC.CTCAC.F.44.C
55575060.55567489.chr7.CML228.CAAAC.CAAAC.R.29.C
C x B x F
52325172.52332739.chr10.B73.GCAAC.GCAAC.F.52.C
52027017.52019448.chr7.II14H.CGAAG.CGAAG.R.0.C
B.UNIQ consensus sequence
0.0010

d.


52085132.52101381.chr10.CML322.CGAAA.CGAAA.F.I.30.D
135849292.135856858.chr1.HP301.CGAAA.CGAAA.F.52.D
53641749.53656891.chr10.CML69.CTTTA.CTTTA.F.I.37.D
48872283.48879847.chr8.HP301.AGTAG.AGTAG.F.15.D
140839206.140831656.chr1.Ki11.TTCTT.TTCTT.R.7.D
61935177.61927614.chr9.CML277.ATATG.ATATG.R.0.D
141163136.141155572.chr1.CML333.ATGTT.ATGTT.R.59.D
58743268.58750832.chr6.B73.ATGTT.ATGTT.F.59.D
188881884.188889452.chr1.B97.AGTTG.AGTTG.F.59.D
248980536.248988099.chr4.CML333.CAGGG.CAGGG.F.89.D
89246861.89239295.chr2.CML322.GAGAC.GAGAC.R.37.D
63626380.63618807.chr7.B73.CCAAA.CCAAA.R.52.D
106741288.106756420.chr5.CML333.CCCAT.CCCAT.F.I.76.D
66679716.66687257.chr10.B97.GTGAT.GTGAT.F.82.D
112606348.112613916.chr5.CML69.AAAGT.AAAGT.F.7.D
118524608.118539049.chr2.CML322.GGTGA.GGTGA.F.I.40.D
107645200.107659621.chr5.CML333.CCAAA.CCAAA.R.I.52.D
50138984.50146548.chr8.Mo18W.AATAC.AATAC.F.7.D
56649024.56664160.chr7.M37W.CTGTG.CTGTG.F.I.7.D
60420983.60435425.chr9.M37W.TTTAG.TTTAG.R.I.22.D
58019867.58034304.chr7.Tx303.TTATC.TTATC.F.I.30.D
55285305.55292867.chr7.B73.ACAAC.ACAAC.F.52.D
53003896.52996335.chr10.CML228.GGTTA.GGTTA.R.52.D
109101160.109108723.chr4.B73.CCTTC.CCTTC.F.67.D
107379184.107386752.chr5.CML333.GAAGA.GAAGA.F.67.D
106352541.106344980.chr5.CML333.CAACC.CAACC.R.30.D
59255168.59247597.chr9.B73.ATGCT.ATGCT.R.59.D
137426432.137433994.chr1.Tzi8.AGTAG.AGTAG.F.7.D
137739820.137732259.chr1.Tzi8.GCTAC.GCTAC.R.7.D
107172068.107164504.chr5.CML333.AGCAG.AGCAG.R.15.D
48857624.48850056.chr8.B73.TGCTT.TGCTT.R.67.D
92777586.92770028.chr3.P39.ACCTG.ACCTG.R.22.D
106660957.106653414.chr5.CML333.GTTCT.GTTCT.R.59.D
87236542.87280566.chr3.CML103.CTAAA.CTAAA.R.I.44.D
8978133.8970565.chr5.NC358.AGGTA.AGGTA.R.30.D
61179754.61194737.chr9.CML322.CATTC.CATTC.F.I.30.D
61382553.61405204.chr9.Ki11.CATTC.CATTC.R.I.53.D
54137683.54130118.chr7.B73.GTTAA.GTTAA.R.67.D
49326290.49333844.chr8.B73.ACTAT.ACTAT.F.45.D
141335996.141350413.chr1.CML333.TGTTT.TGTTT.R.I.45.D
58580990.58595417.chr6.B73.TGTTT.TGTTT.F.I.75.D
135975784.135968227.chr1.HP301.CATGC.CATGC.R.82.D
56975202.56982763.chr7.B73.CATGC.CATGC.F.89.D
60153078.60145516.chr9.CML322.GCTAC.GCTAC.R.74.D
85582290.85574724.chr3.HP301.ACATA.ACATA.R.59.D
119856933.119864496.chr2.CML322.GAAAC.GAAAC.F.15.D
95664572.95657008.chr2.CML247.GAAAC.GAAAC.R.30.D
56525907.56518346.chr7.M162W.TTGGG.TTGGG.R.52.D
63053284.63060849.chr6.B97.TTTGT.TTTGT.F.89.D
62530878.62523313.chr6.Mo18W.TTTGT.TTTGT.R.104.D
74810870.74803317.chr6.CML277.GACAG.GACAG.R.82.D
110781070.110796213.chr4.B97.CTTAT.CTTAT.R.I.30.D
56170752.56185884.chr7.B73.CACGT.CACGT.F.I.44.D
B.UNIQ consensus sequence
D
D x C
0.0010

e.

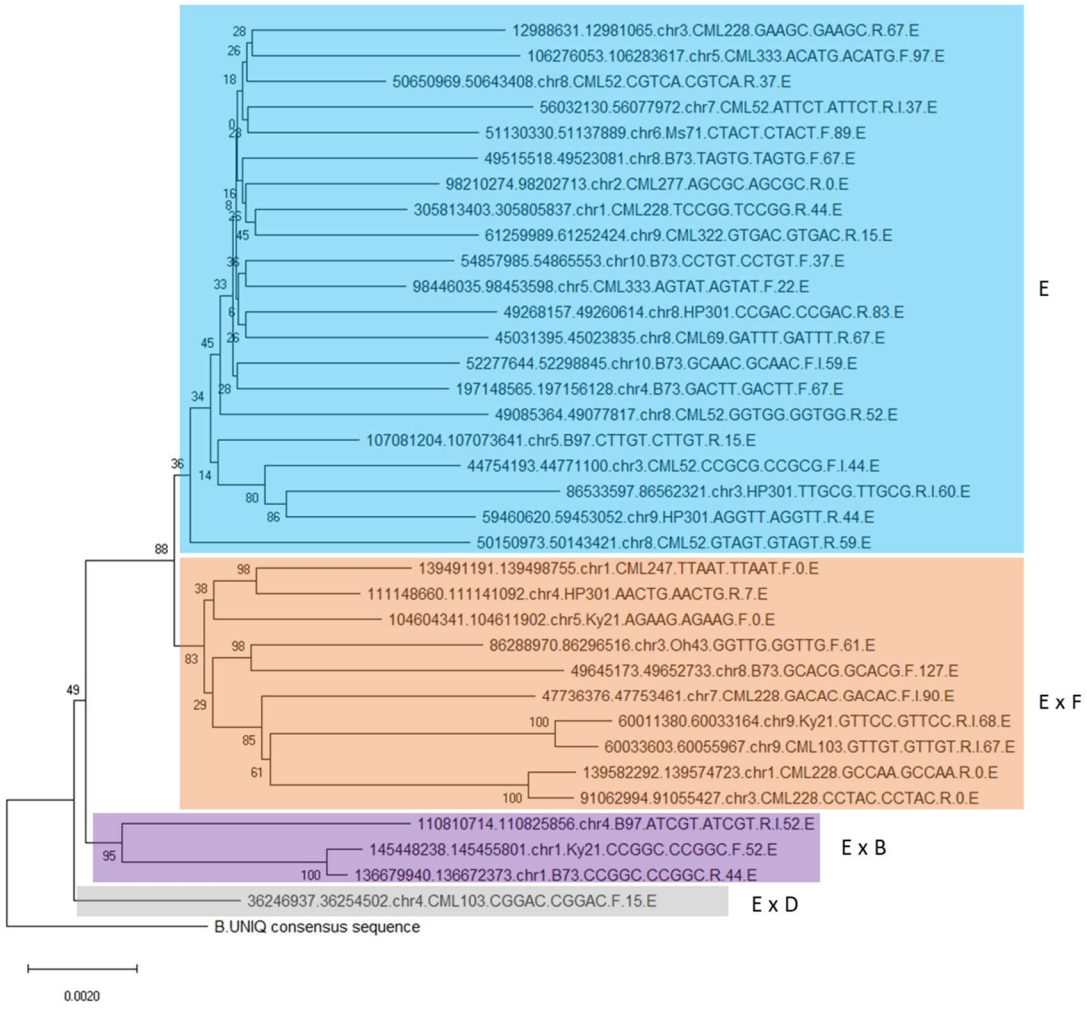


Figure S40. CR2 sub-recombinant phylogenetic trees.

Individual phylogenetic trees were generated from parental CR2 subtypes A-E (**a.-e.** respectively) (Neighbor joining, K2p, pairwise deletion, bootstrap 1000). Each tree is rooted using a distant CR2 parental consensus CR2. Ancestral parental haplotypes “A” and “B” were found to have no sub recombinant elements. However, trees for “C”, “D” and “E” subtypes indicated the presence of between 1 and 3 subrecombinants.

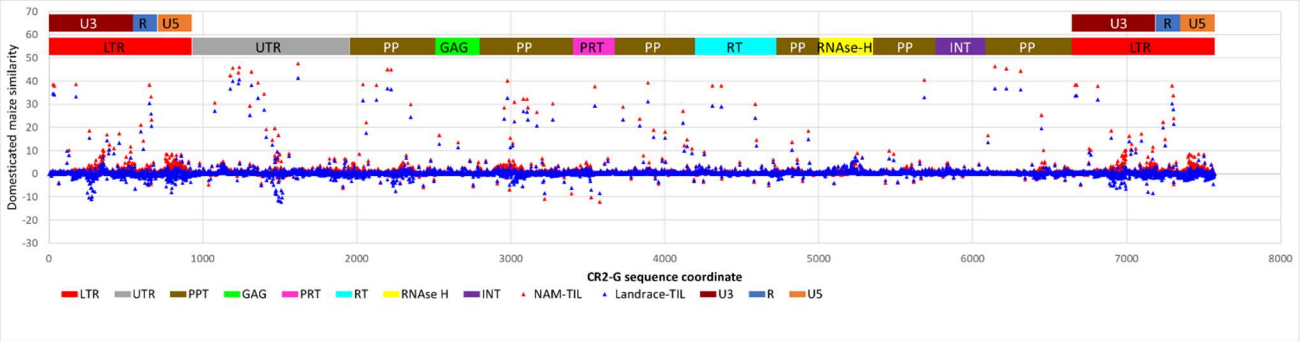
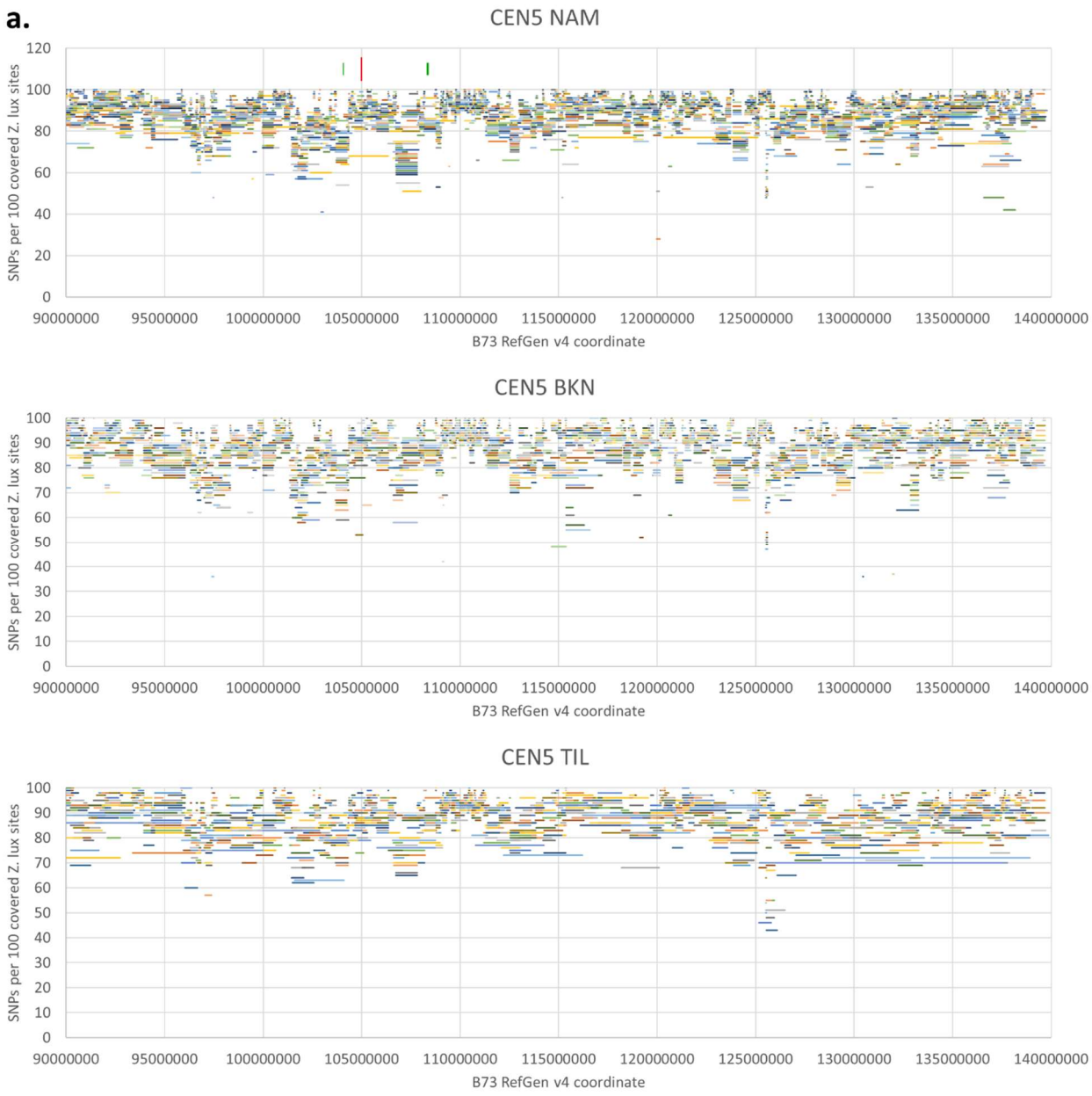
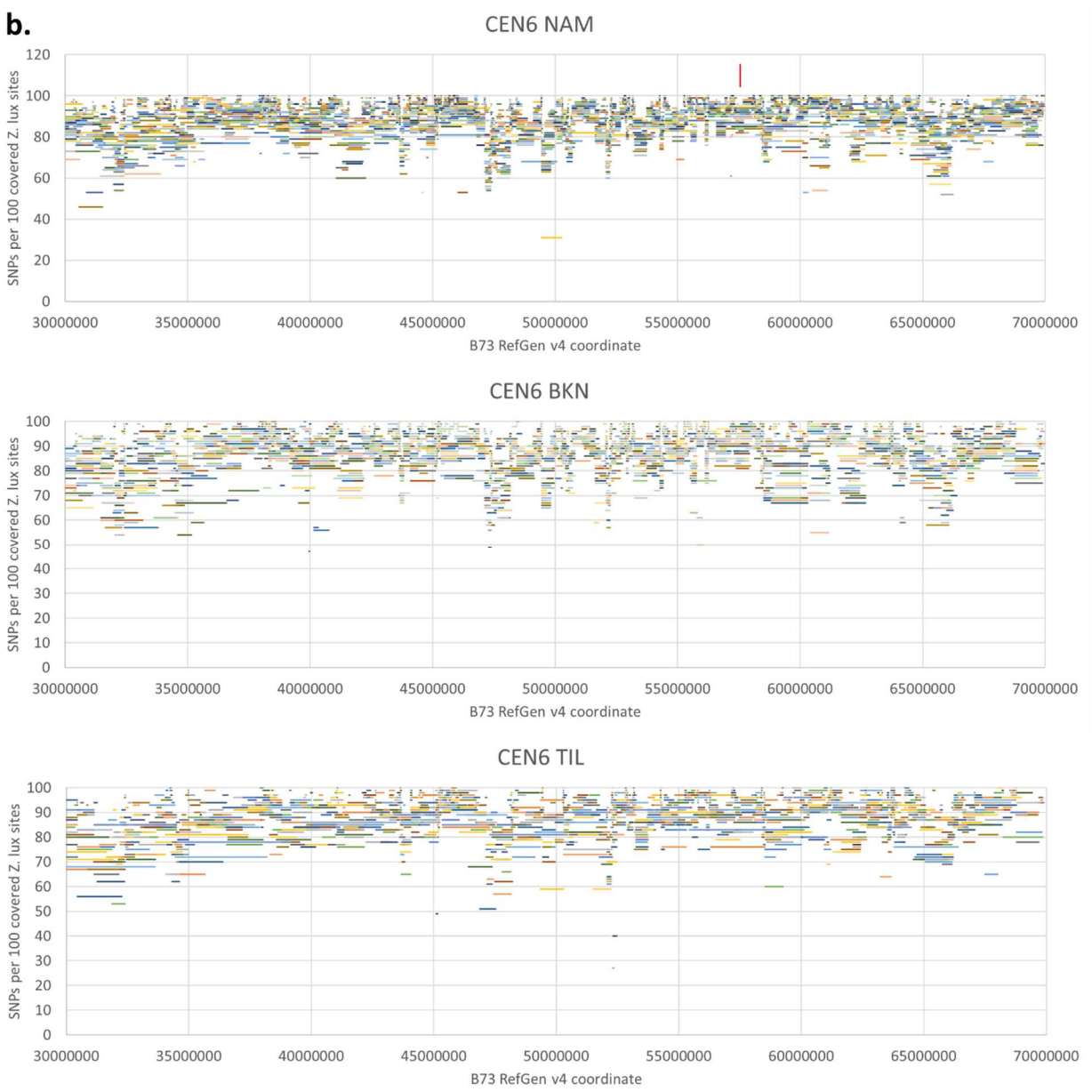


Figure S41. Enrichment of SNPs in NAM compared to teosinte.

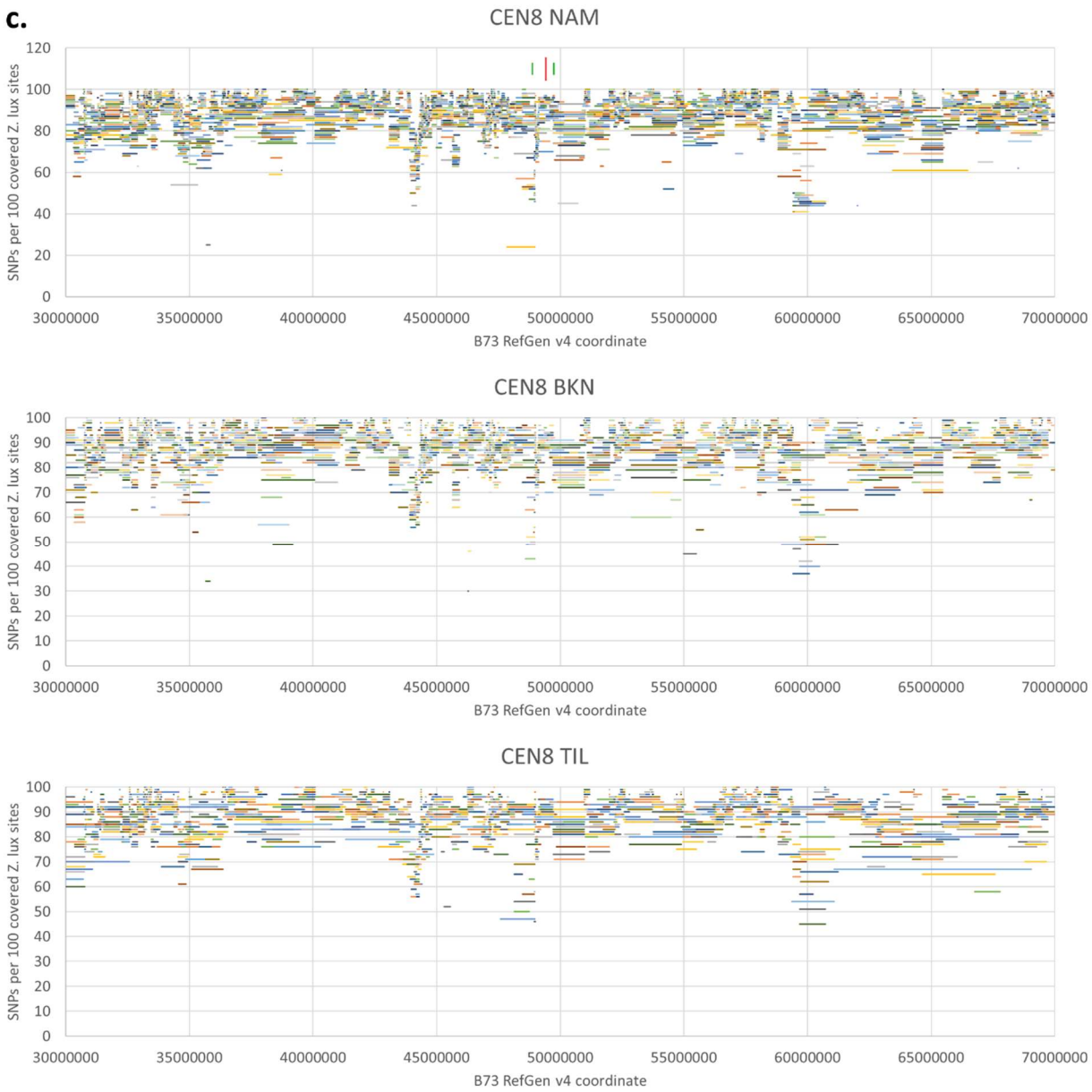
Enrichment for SNP positions in NAM lines and landraces compared to teosinte is marked with red and blue triangles respectively. Domesticated maize similarity is calculated for each TIL as the difference of the percentage of reads corresponding to the CR2-G SNP for the TIL and the average of the TIL percentages for the SNP position. Groups of lines were calculated as the difference between the mean for the groups (NAM, Landrace, or TIL) over the position (NAM-TIL and Landrace-TIL). Most positions hover at or around zero indicating no enrichment. The CR2 is annotated with the major regions (LTR, UTR and non-core domain polyprotein (PP)) and functional domains (GAG, protease (PRT), reverse transcriptase (RT), RNAse-H, and integrase (INT)) as well as the LTR subregions U3, R, and U5).

a.

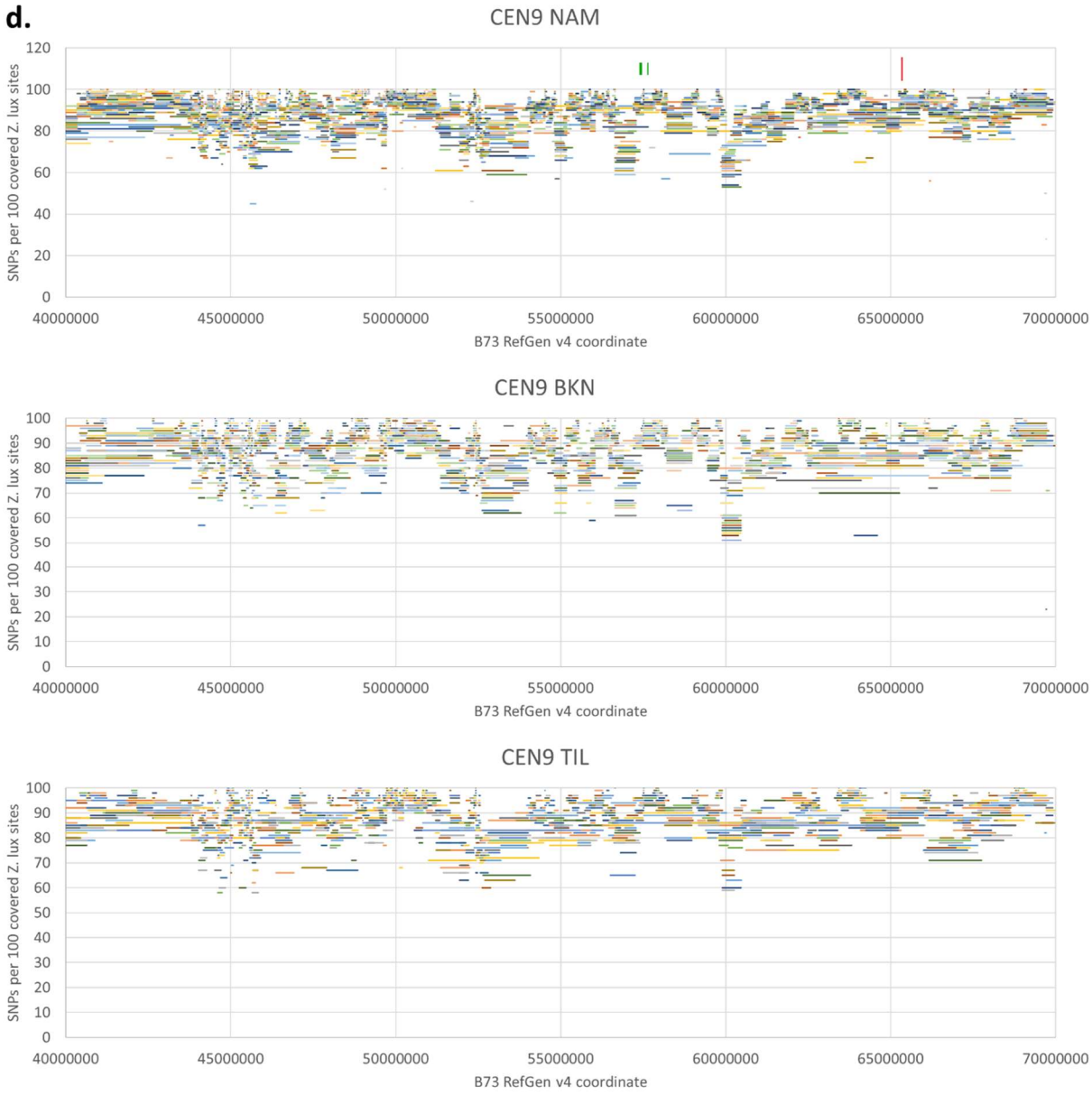

CEN5 NAM
SNPs per 100 covered Z. lux sites
0
20
40
60
80
100
120
90000000
95000000
100000000
105000000
110000000
115000000
120000000
125000000
130000000
135000000
140000000
B73 RefGen v4 coordinate
CEN5 BKN
SNPs per 100 covered Z. lux sites
0
10
20
30
40
50
60
70
80
90
100
90000000
95000000
100000000
105000000
110000000
115000000
120000000
125000000
130000000
135000000
140000000
B73 RefGen v4 coordinate
CEN5 TIL
SNPs per 100 covered Z. lux sites
0
10
20
30
40
50
60
70
80
90
100
90000000
95000000
100000000
105000000
110000000
115000000
120000000
125000000
130000000
135000000
140000000
B73 RefGen v4 coordinate


b.
CEN6 NAM
SNPs per 100 covered Z. lux sites
B73 RefGen v4 coordinate
CEN6 BKN
SNPs per 100 covered Z. lux sites
B73 RefGen v4 coordinate
CEN6 TIL
SNPs per 100 covered Z. lux sites
B73 RefGen v4 coordinate

c.


CEN8 NAM
SNPs per 100 covered Z. lux sites
120
100
80
60
40
20
0
30000000
35000000
40000000
45000000
50000000
55000000
60000000
65000000
70000000
B73 RefGen v4 coordinate
CEN8 BKN
SNPs per 100 covered Z. lux sites
100
90
80
70
60
50
40
30
20
10
0
30000000
35000000
40000000
45000000
50000000
55000000
60000000
65000000
70000000
B73 RefGen v4 coordinate
CEN8 TIL
SNPs per 100 covered Z. lux sites
100
90
80
70
60
50
40
30
20
10
0
30000000
35000000
40000000
45000000
50000000
55000000
60000000
65000000
70000000
B73 RefGen v4 coordinate

d.


CEN9 NAM
SNPs per 100 covered Z. lux sites
B73 RefGen v4 coordinate
CEN9 BKN
SNPs per 100 covered Z. lux sites
B73 RefGen v4 coordinate
CEN9 TIL
SNPs per 100 covered Z. lux sites
B73 RefGen v4 coordinate

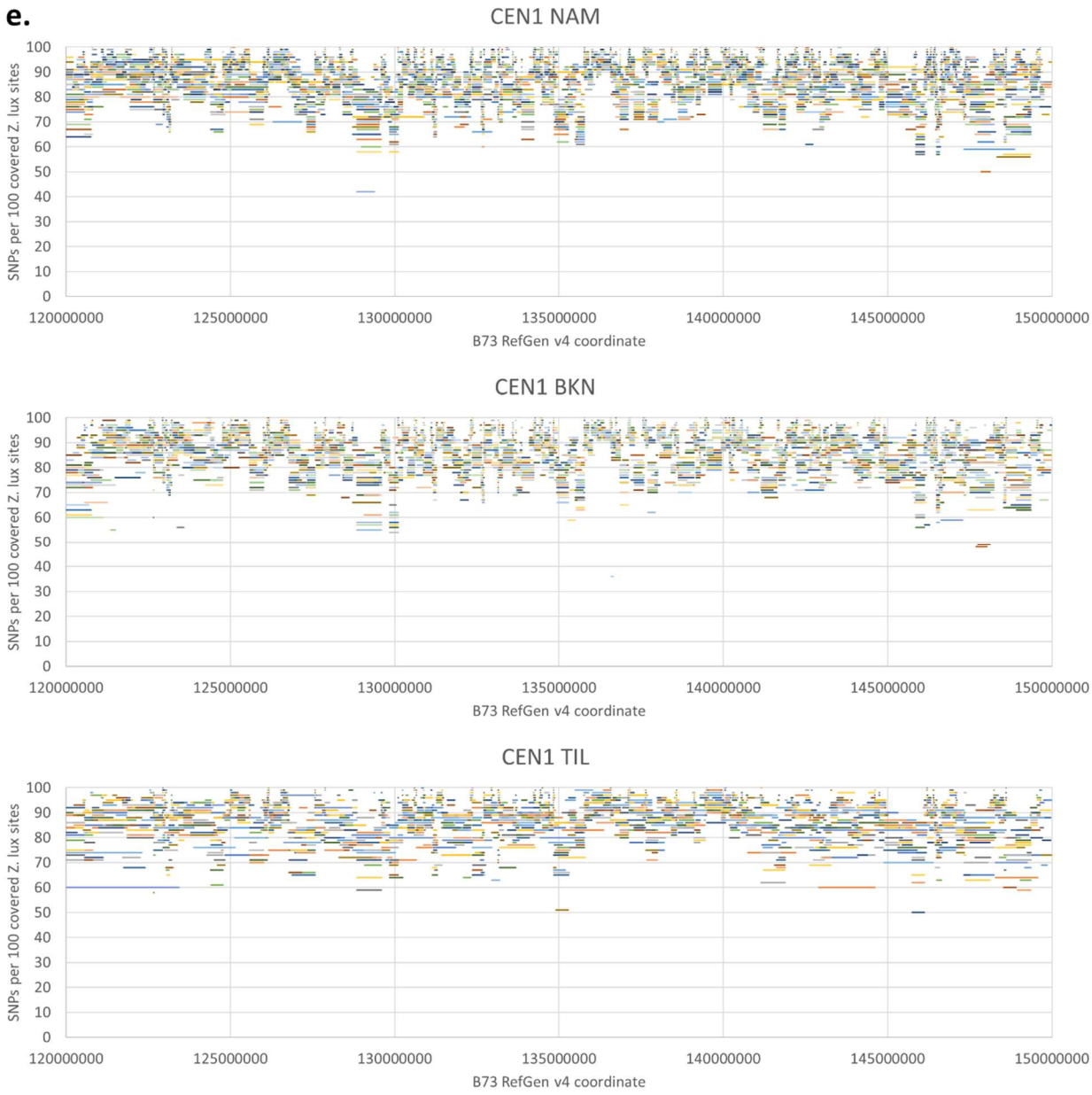


Figure S42. SNP density of NAM, TIL and landraces compared to Zea luxurians in CENs 1 (e), 5 (a), 6 (b), 8 (c), and 9 (d).

Each bar represents the number of SNP positions in a region with 100 positions covered by *Zea luxurians* GSS data in B73 RefGen v4. The CentC repeat regions are marked with green lines at the top of each NAM chart. The position of CR2 elements with both junctions shared by *Zea luxurians* that were not inserted into CentC (*SI Appendix Table ST21*) are marked with a red line. CEN1 is included as a negative control and contains no shared *Zea luxurians* junctions.

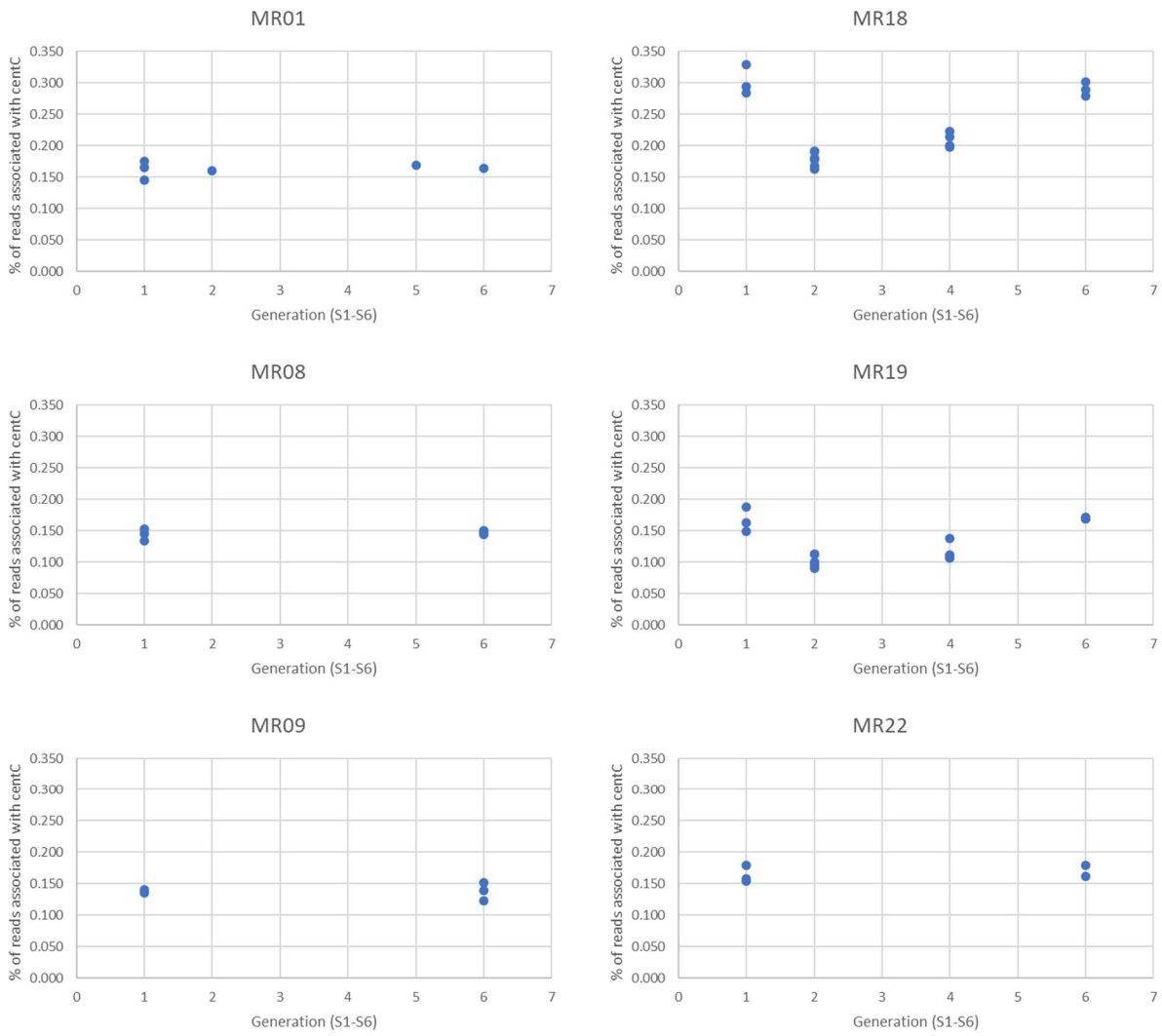


Figure S43. Percentage of nucleotides corresponding to centC in six cultivars across six generations of inbreeding.

No significant loss of centC was detected in any of the 6 cultivars though this may be due to data limitations, centromere positioning, or the number of generations of inbreeding.

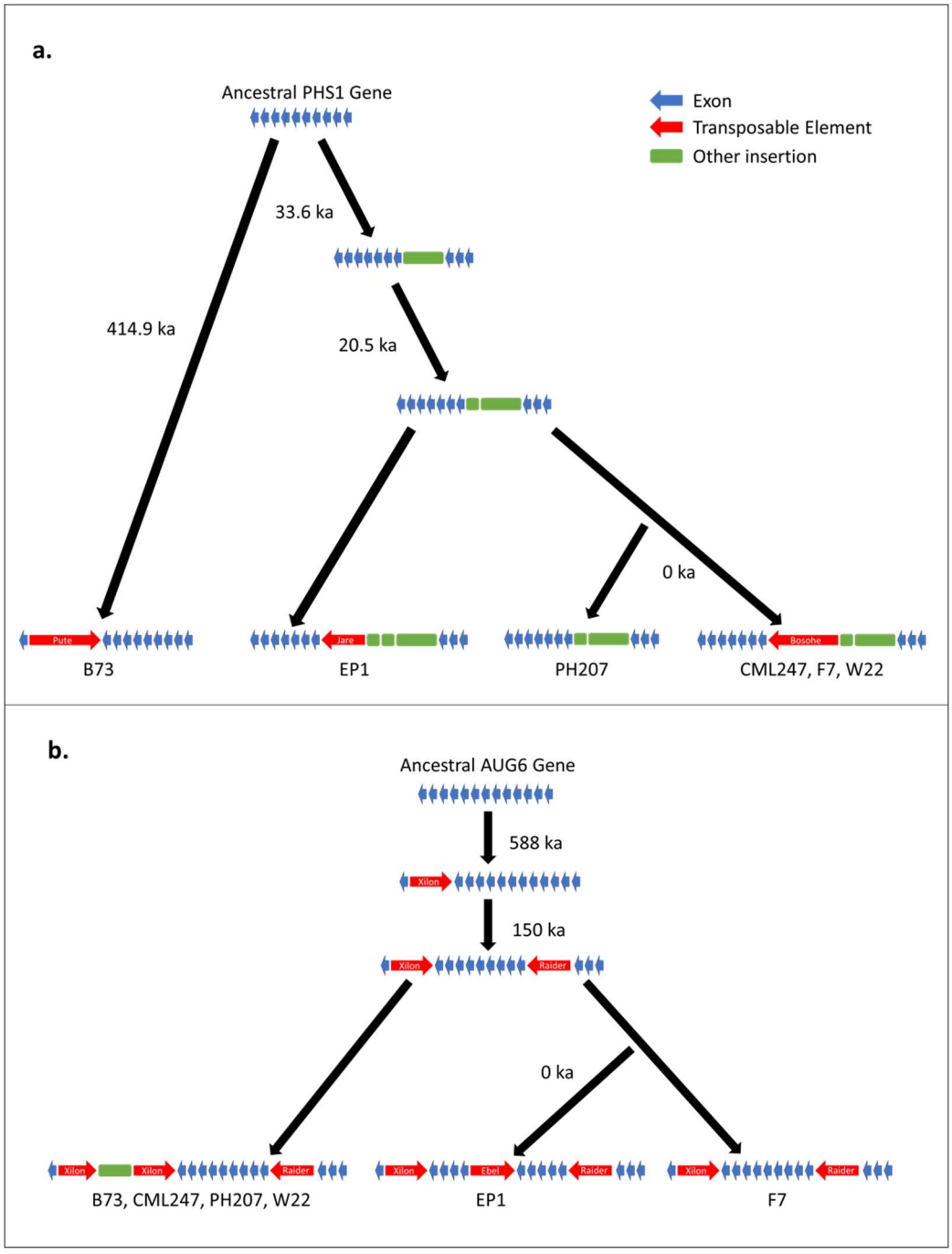


Figure S44. Schematic phylogeny of CEN5 AUG6 and PHS1 Genes.

Exons (not to scale) are shown as blue arrows in the direction of transcription. Transposable elements and their orientation are shown as red arrows. Non-TE insertions of unkown orientation are marked in green.

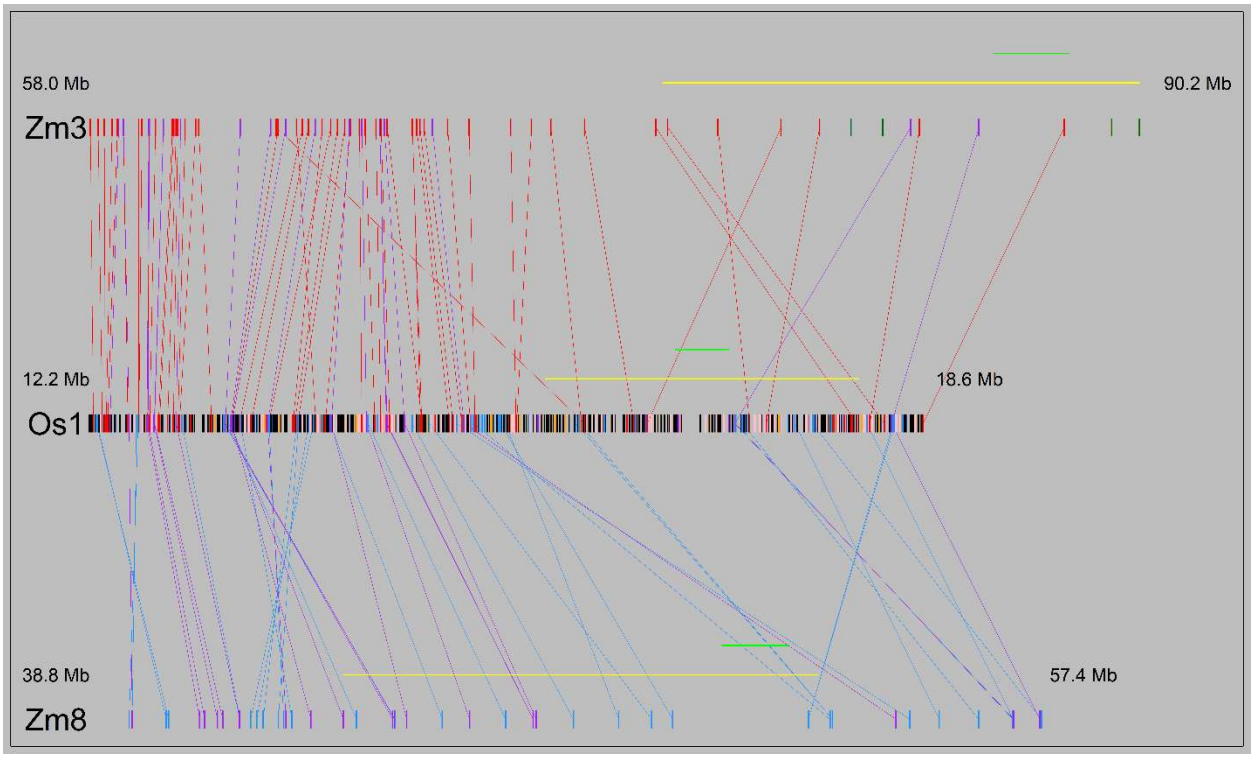


Figure S45. Syntenic regions of Os1, Zm3 and Zm8.

Centromeric regions are denoted by yellow lines, and active centromeres are marked with green lines. Genes with orthologs shared between Os1 and Zm3 are marked in red. Genes with orthologs shared between Os1 and Zm8 are marked in blue. Os1 genes with orthologs shared between both Zm3 and Zm8 are marked in purple. Os1 genes with no orthologs are marked as black. Os1 genes with Si5 orthologs only are marked in orange. Os1 genes with orthologs on a non-syntenic chromosomes are denoted in pink. Zm3 genes with orthologs on Os12 are marked in dark green.

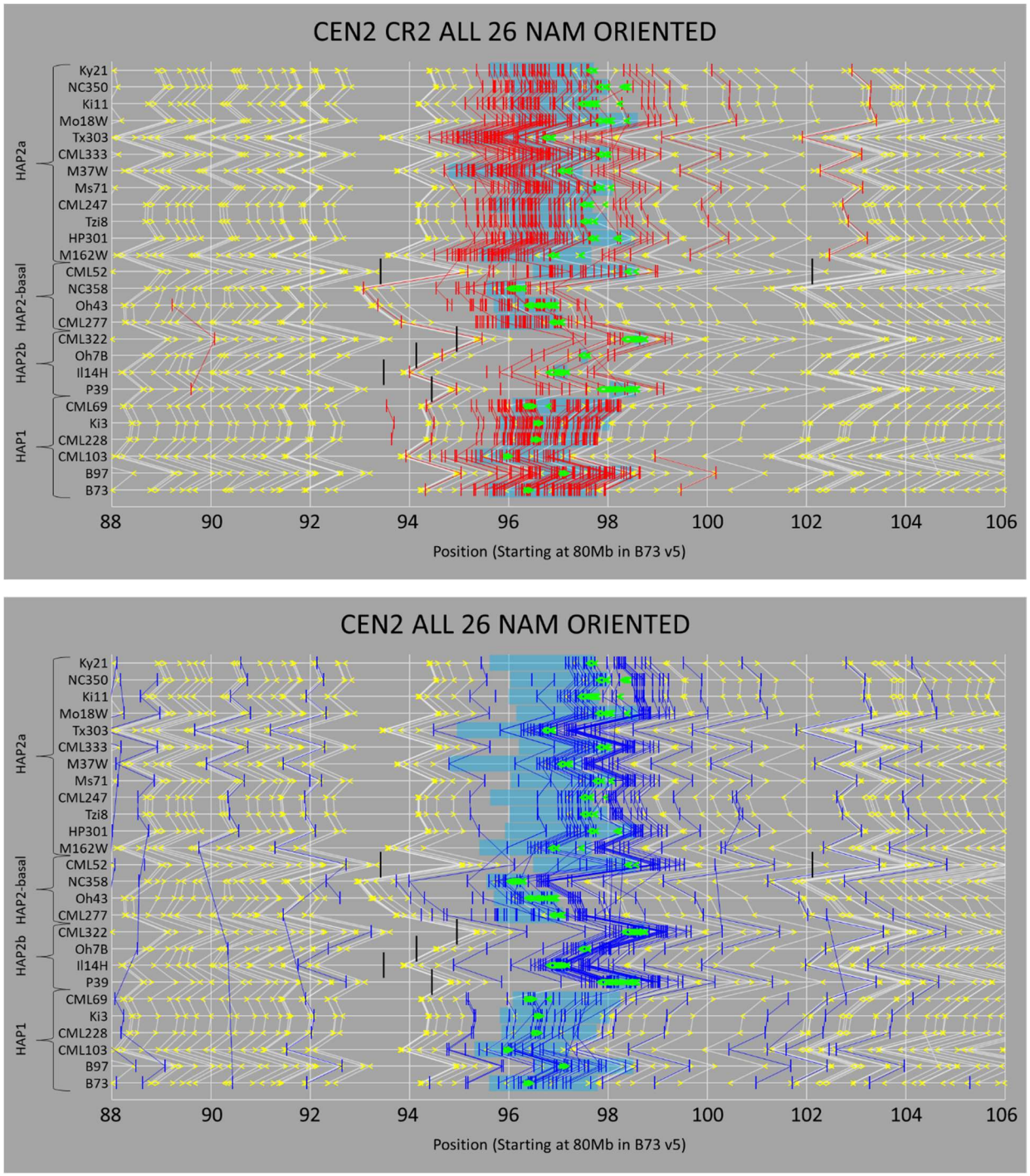


Figure S46. CEN2 CR2 and CR1 gene models and orthologs among the NAM lines.

Shared CR2s are shown as red tick marks connected by red lines (top). Shared CR1s are shown as red tick marks connected by blue lines (bottom). Most CR1 elements cluster around the ancestral centC locus while CR2 elements inhabit the post domestication active centromere positions. Orthologous genes among all 26 NAM lines were identified via blastn of B73 v5 gene models to the 26 physical maps and are shown as yellow markers connected by a white line. All NAM lines are zeroed at the start of Zm00001eb086050 (B73 v5 annotation). The arrows denote the orientation of the gene model. Active centromeres were identified via cenH3 ChIP-seq for each NAM line mapped to the respective physical

map. Active centromere regions (light blue shade) were manually defined using cenH3 coverage charts. The active CEN was defined as the region where cenH3 coverage is ~3X higher than background. CentC repeats identified with blast are marked with light green arrows. The inversion breakpoints in the HAP2b lines and the unique CML52 pericentric inversion breakpoints are marked with black lines. Blast results were sorted by line and coordinate and filtered to remove non-syntenic genes.

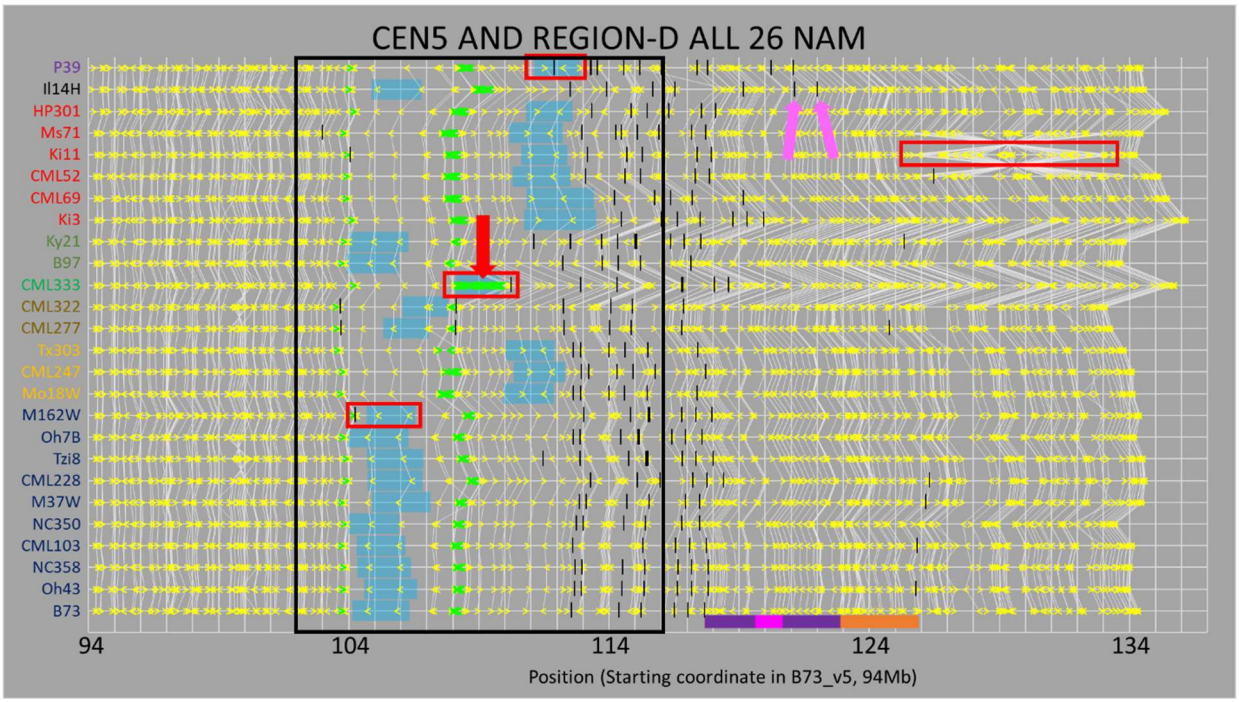


Figure S47. CEN5 Schematic overview, gene models, and orthologs among the NAM lines.
Orthologous genes among all 26 NAM lines were identified via blastn of B73 v5 gene models to the 26 physical maps and are shown as yellow markers connected by a white line. The arrows denote the orientation of the gene model. Active centromeres were identified via cenH3 ChIP-seq for each NAM line mapped to the respective physical map. Active centromere regions (light blue shade) were manually defined using cenH3 coverage charts. The active CEN was defined as the region where cenH3 coverage is ~3X higher than background. CentC repeats identified with blast are marked with light green arrows. The inversion breakpoints in the colored lineages are marked with black lines. The CEN5 proximal domestication locus region-D is shaded Purple pink and orange. The orange shade represents the selected region among improved maize and purple represents the selected region among domesticated maize (including HM3 landraces). P39 and Il14H have an introgression of distant sequence in region-D (Pink shade and arrows). There is only one line with an active CEN5M (CML333, Red box with red arrow). The closest recombination breakpoint to an active centromere was ___kb upstream of the M162W CEN5L. The centromeric recombination breakpoint in P39 was found to have occurred before the centromere occupied that locus due to the very recent formation of the P39 CEN5R (see Chapter 4 Figure 4.5 for estimated divergence of the flanking regions). The region investigated for shared CR content is boxed in black (*SI Appendix Figure S48*). There is a Ki11 specific ~7.5 Mb inversion in Ki11 DS of Region-D.

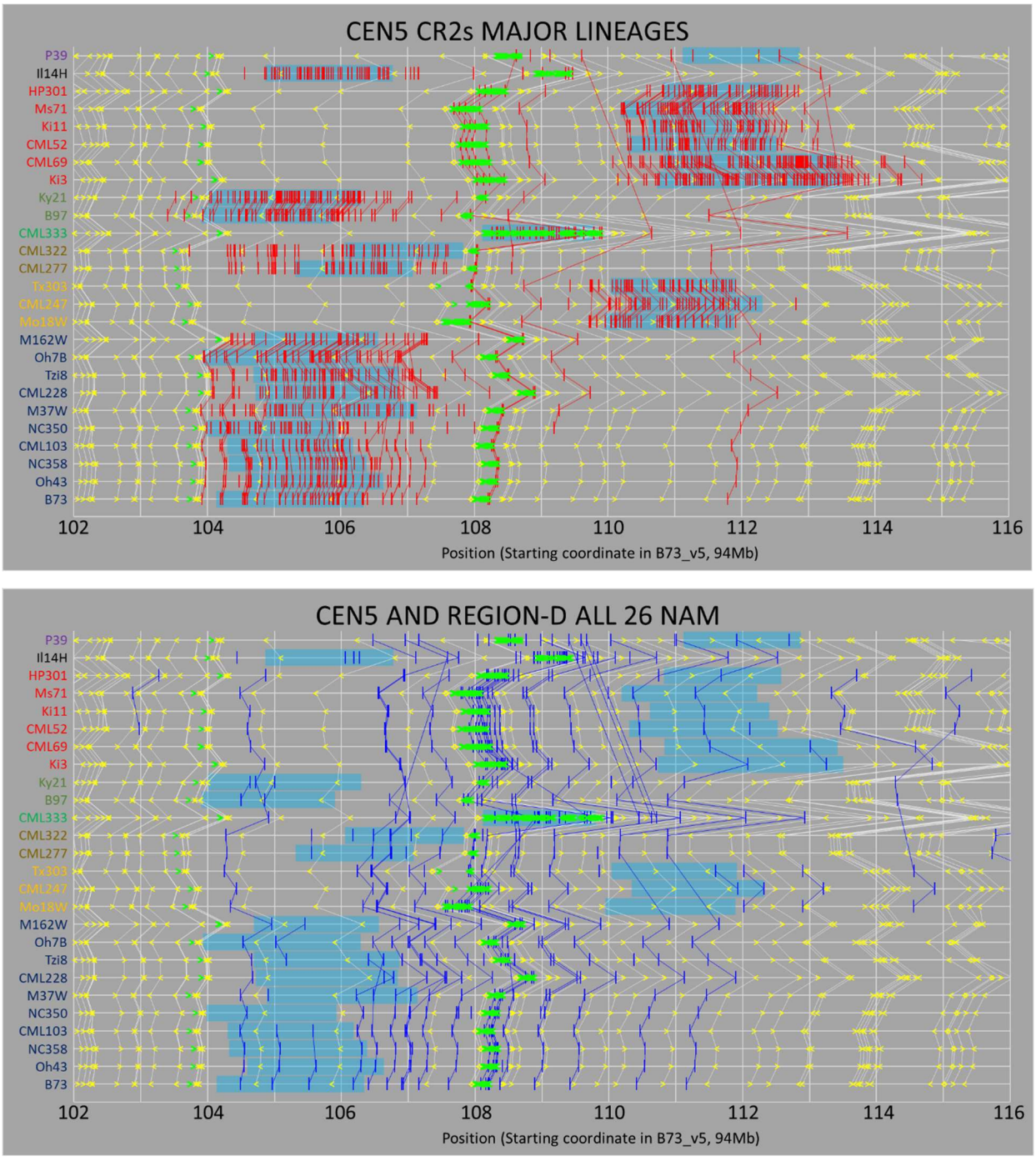


Figure S48. CEN5 CR2 and CR1 gene models and orthologs among the NAM lines.
Shared CR2s are shown as red tick marks connected by red lines (top). Shared CR1s are shown as red tick marks connected by blue lines (bottom). Most CR1 elements cluster around the ancestral centC locus while CR2 elements inhabit the post domestication active centromere positions. Orthologous genes among all 26 NAM lines were identified via blastn of B73 v5 gene models to the 26 physical maps and are shown as yellow markers connected by a white line. All NAM lines are zeroed at 102Mb (B73 v5). The arrows denote the orientation of the gene model. Active centromeres were identified via cenH3

ChIP-seq for each NAM line mapped to the respective physical map. Active centromere regions (light blue shade) were manually defined using cenH3 coverage charts. The active CEN was defined as the region where cenH3 coverage is ~3X higher than background. CentC repeats identified with blast are marked with light green arrows. Blast results were sorted by line and coordinate and filtered to remove non-syntenic genes.

Supplemental Tables

v5 Start	v5_End	Orientation	Alias	Gene_ID	Expressed (B73)	Region	Marker	Notes
103,867,079	103,867,914	-	Zm00001e013946	Zm00001eb235050	Y	CEN5L	A	bZIP-transcription factor 36
104,723,179	104,723,810	-	Zm00001e013948	Zm00001eb235070	Y	CEN5L	B	
105,855,811	105,856,493	-	Zm00001e013955	none	N	CEN5L	C	
107,269,317	107,269,923	-	Zm00001e013957	Zm00001eb235180	N	CEN5L	D	
107,929,306	107,980,823	-	Zm00001e013960	Zm00001eb235200	Y	Upstream of CEN5M		AUGMIN Subunit 6
107,971,926	107,972,277	+	Zm00001e013962	none	Y	Upstream of CEN5M		
107,972,729	107,973,102	+	Zm00001e013963	none	Y	Upstream of CEN5M		
108,409,233	108,439,785	+	Zm00001e013964	Zm00001eb235230	N	Downstream of CEN5M		Kelch domains
108,436,905	108,437,440	+	Zm00001e013965	none	N	Downstream of CEN5M		
108,542,805	108,572,765	+	Zm00001e013966	Zm00001eb235240	N	Downstream of CEN5M		Voltage gated CIC domains
108,945,298	108,980,844	+	Zm00001e013967	Zm00001eb235260	Y	Downstream of CEN5M		API5
109,476,497	109,487,597	+	Zm00001e013969	Zm00001eb235270	Y	Downstream of CEN5M		DUF3082 (domain of unknown function)
109,824,723	109,825,035	+	Zm00001e013970	none	N	Downstream of CEN5M		
110,102,282	110,119,169	+	Zm00001e013971	Zm00001eb235320	Y	CEN5R	E	P-loop_NTPase domain
110,808,014	110,809,764	+	Zm00001e013973	Zm00001eb235330	N	CEN5R	F	PLN3089 (hypothetical protein)
111,463,221	111,464,966	-	Zm00001e013976	Zm00001eb235340	N	CEN5R	G	
112,113,874	112,117,629	+	Zm00001e013978	Zm00001eb235350	N	CEN5R	H	

Table ST1. Details of expressed genes flanking CEN5M. B73 gene models corresponding to CEN5L markers A-D and CEN5R markers E-H

Line	Markers	Region_start	Region_stop	Total_H2A.Z_Nuc	Cov_sum_H2A.Z_Peaks
B73	**A-D**	103882536	107269317	20103	1636404
Oh43	**A-D**	103508445	106964785	14315	295776
NC358	**A-D**	102220075	105680390	16062	366907
CML103	**A-D**	103359347	106805865	15252	514012
NC350	**A-D**	102143136	105571338	16972	448254
M37W	**A-D**	104373074	107953392	17076	463334
CML228	**A-D**	107319939	111186083	14780	481176
Tzi8	**A-D**	102841912	106349766	13455	224544
Oh7B	**A-D**	102717921	106173830	17492	548609
M162W	**A-D**	103918630	107353132	16490	591492

Mo18W	**A-D**	104249354	107207409	24179	912050
CML247*	**A-D**	104244576	107274038	11284	368403
Tx303*	**A-D**	103362046	106339399	21381	716218
CML277	**A-D**	100422897	103789046	18383	525245
CML322	**A-D**	101395942	104815005	18009	345512
CML333	**A-D**	102351188	105381030	19374	656614
B97	**A-D**	102470602	105654677	20638	647532
Ky21	**A-D**	104308475	107613756	19598	848273
Ki3	**A-D**	104871084	107825549	11469	215002
CML69	**A-D**	103626649	106581062	21945	591090
CML52	**A-D**	105247675	108198976	22189	562349
Ki11	**A-D**	103050408	106004885	25206	1265642
Ms71	**A-D**	105515697	108474035	28584	984669
HP301	**A-D**	104025268	106974295	23960	579249
Il14H	**A-D**	102584863	106565095	21432	650325
P39	**A-D**	103322102	106611342	20341	562125
B73	**E-H**	110102282	112113874	26271	2247365
Oh43	**E-H**	109812606	111819089	18391	397279
NC358	**E-H**	108514791	110536843	19566	571939
CML103	**E-H**	109640524	111651927	18870	719898
NC350	**E-H**	108490647	110491325	14683	484394
M37W	**E-H**	110900368	112928597	19445	598671
CML228	**E-H**	114112838	116130464	16824	407998
Tzi8	**E-H**	109247559	111259560	10795	182939
Oh7B	**E-H**	109030998	111018261	23194	747261
M162W	**E-H**	110210103	112197474	23605	821545
Mo18W	**E-H**	110336147	112632120	14422	516605
CML247*	**E-H**	110554208	112849567	6408	214073
Tx303*	**E-H**	109579845	111847922	11640	346238

CML277	**E-H**	106590596	108537632	21762	741494
CML322	**E-H**	107593102	109515029	20084	470582
CML333	**E-H**	109818964	111913800	12789	438711
B97	**E-H**	108421389	110366838	23631	814010
Ky21	**E-H**	110388433	112463568	23515	1083727
Ki3	**E-H**	111026691	114595503	7178	109451
CML69	**E-H**	109759024	113310429	11415	327182
CML52	**E-H**	111355889	113840363	11281	279062
Ki11	**E-H**	109103763	111671240	14758	640116
Ms71	**E-H**	111636578	114098447	20570	583213
HP301	**E-H**	110139169	112595180	21803	510356
Il14H	**E-H**	109984907	111907776	18803	597247
P39	**E-H**	109871503	111991062	13838	374435

Table ST2. Coordinates and H2A.Z coverage data for the 26 NAM lines in the A-D region and the E-H region.

*Coverage data for CML247 and Tx303 were calculated using a fold change threshold of 3.5, all other NAM lines were calculated with a fold change threshold of 3 (*SI APPENDIX Figure S4*).

Line	Total Nuc Ratio	Cov Sum Ratio
B73	0.77	0.73
Oh43	0.78	0.74
NC358	0.82	0.64
CML103	0.81	0.71
NC350	1.16	0.93
M37W	0.88	0.77
CML228	0.88	1.18
Tzi8	1.25	1.23
Oh7B	0.75	0.73
M162W	0.70	0.72

Mo18W	1.68	1.77
CML247	1.76	1.72
Tx303	1.84	2.07
CML277	0.84	0.71
CML322	0.90	0.73
CML333	1.51	1.50
B97	0.87	0.80
Ky21	0.83	0.78
Ki3	1.60	1.96
CML69	1.92	1.81
CML52	1.97	2.02
Ki11	1.71	1.98
Ms71	1.39	1.69
HP301	1.10	1.13
Il14H	1.14	1.09
P39	1.47	1.50

Table ST3. Ratios for CEN5L and CEN5R among the NAM lines.

Deletion	Line	Start	End	Deletion size	Alignment
Blue	B73	105,427,987	105,427,988	deleted	BLUE_DELETION.fasta
Blue	Mo18W	105,597,397	105,710,931	113,534	BLUE_DELETION.fasta
Green	B97	103,405,815	103,405,816	deleted	GREEN_DELETION.fasta
Green	Ki11	103,826,830	104,125,907	299,077	GREEN_DELETION.fasta
Brown	CML322	104,147,092	104,147,705	deleted	BROWN_DELETION.fasta
Brown	B73	106,563,334	106,665,267	101,933	BROWN_DELETION.fasta

Table ST4. CEN5L deletion details.

ID	LINE	TYPE	READS	MAPPED_%
19325X1	M162W	cenH3 chIP DNA	79,979,199	42.37%
19325X2	Oh43	cenH3 chIP DNA	72,550,525	32.68%
19325X3	CML333	cenH3 chIP DNA	89,008,755	37.55%
19325X4	Ky21	cenH3 chIP DNA	106,721,470	39.31%
19325X5	M37W	cenH3 chIP DNA	76,339,414	39.44%
19325X6	CML69	cenH3 chIP DNA	79,820,672	35.42%
19325X7	P39	cenH3 chIP DNA	74,436,485	41.88%
19325X8	NC358	cenH3 chIP DNA	66,566,027	37.21%
19325X9	CML103	cenH3 chIP DNA	78,045,098	39.85%
19325X10	CML228	cenH3 chIP DNA	82,808,526	33.07%
19325X11	CML52	cenH3 chIP DNA	70,705,399	32.14%
19325X12	B97	cenH3 chIP DNA	72,465,157	39.62%
19325X13	Mo17	cenH3 chIP DNA	80,338,017	30.15%
19325X14	HP301	cenH3 chIP DNA	74,230,637	27.88%
19325X15	Mo18W	cenH3 chIP DNA	75,459,513	41.41%
19325X16	NC350	cenH3 chIP DNA	74,205,863	38.38%
19325X17	Tzi8	cenH3 chIP DNA	61,980,390	30.84%
19325X18	Ki3	cenH3 chIP DNA	79,494,797	34.01%
19325X19	Ki11	cenH3 chIP DNA	108,849,565	42.60%
19325X20	CML277	cenH3 chIP DNA	88,509,393	33.90%
19325X21	CML247	cenH3 chIP DNA	84,315,322	38.37%
19325X22	Tx303	cenH3 chIP DNA	56,818,190	26.99%
19325X23	Oh7B	cenH3 chIP DNA	86,922,514	40.66%

19325X24	**IL14H**	**cenH3 chIP DNA**	**90,969,275**	**37.82%**
19325X25	**CML322**	**cenH3 chIP DNA**	**67,598,535**	**35.34%**
19325X26	**Ms71**	**cenH3 chIP DNA**	**64,959,620**	**35.08%**
8464X1	**B73**	**cenH3 chIP DNA**	**192,541,415**	**37.91%**

Table ST5. H2A.Z Read count and mapping percentage.

CR	SUBTYPE	MAX H2A.Z OVER CONS	CONTROL REGION MAX H2A.Z	FOLD_DIF	ELEM # IN B73_v5	APPROX MAX COV PER ELEMENT
CR1	**A**	**2144**	**592**	**3.6**	**29**	**74**
CR1	**B**	**1441**	**592**	**2.4**	**49**	**29**
CR1	**R1**	**3723**	**592**	**6.3**	**180**	**21**
CR1	**R2**	**1923**	**592**	**3.2**	**67**	**29**
CR1	**R3**	**1923**	**592**	**3.2**	**22**	**87**
CR1	**R4**	**1785**	**592**	**3.0**	**69**	**26**
CR1	**R5**	**3728**	**592**	**6.3**	**191**	**20**
CR2	**CR2**	**10907**	**592**	**18.4**	**428**	**25**

Table ST6. Maximum H2A.Z coverage over CR1 subtypes and a Consensus CR2 element compared to the CEN5 H2A.Z control peak.

	REGION13_SYN_START	REGION13_SYN_END
B73 v5	97,489,935	99,668,255
B73 v4	97,329,566	99,494,071
HM3 REGION	97,329,570	99,494,065

Table ST7. Coordinates of Region13_SYN in B73 v4, v5 and HapMap3 SNP coordinates. This region was used for phylogenetic analyses of CEN2.

	US ANCHOR	DS_ANCHOR	v4_START	v4_END	v5_START	v5_END	HM3_START	HM3_END
CEN1	Zm00001d030470	Zm00001d030486	135,199,904	136,742,183	135,003,349	136,535,972	135,199,962	136,737,492
CEN4	Zm00001d050637	LTR Cluster	107,948,303	109,195,106	107,848,407	109,134,275	107,948,369	109,195,089
CEN6	Zm00001d035818	Zm00001d035829	52,783,826	53,795,747	59,482,974	60,494,932	52,783,974	53,795,740
CEN2_REGION13_SYN	US HAP1 INV	DS HAP1 INV	97,329,566	99,494,071	97,489,935	99,668,255	97,329,570	99,494,065

Table ST8. Coordinates and anchors for the regions compared between HapMap3 and the NAM physical maps. These positions in the context of B73 v5 can be see *SI Appendix Figure S15*.

Region	Max K2P	Estimated r[5]	Sequence Lengths from MAUVE alignments	Chromosome	Start	End	MAUVE Region Size
tb1[1]	0.001090670	*5.5E-08	47,728 - 47,740	1	270.460	270.560	100,000
Chr10LA.a[2]	0.001014910	5.1E-08	300,817 - 300,848	10	101.613	102.120	507,200
Chr10LA.b[2]	0.001674140	8.4E-08	83,768 - 83,791	10	102.273	102.433	160,700
Chr10LA.c[2]	0.001511050	7.6E-08	190,229 - 190,266	10	102.890	103.196	306,000
region D.a[3]	0.001225440	6.1E-08	76,004 - 76,016	5	118.131	120.070	1,939,000
region D.b[3]	0.001387260	6.9E-08	2,422,111 - 2,422,133	5	121.118	126.343	5,225,000
CEN1	0.001441510	7.2E-08	2,737,108 - 2,737,236	1	132.000	140.000	8,000,000
CEN4	0.001490820	7.5E-08	2,570,655 - 2,570,748	4	105.112	112.500	7,388,000
CEN6	0.001436050	7.1E-08	3,316,539 - 3,316,709	6	48.671	55.848	7,177,060
New rate[4]		7.3E-08					

Table ST9. Nucleotide substitution rate for dating maize centromere sequence. Substitution rates were calculated using the methods outlined in this paper for previously described domestication loci on chromosomes 1, 5 and 10, and the bottlenecked CENs 1, 4 and 6.

*The rate difference in the *tbl1* locus from Clark *et al.* 2005 and this calculation, are due to different approaches for calculation. Clark *et al.* calculations used the mean number of mutations per lineage, and it was assumed all lineages displayed a "star phylogeny" [149, 237]. Here, maximum K2P was used as a measure of divergence to account for the fact that groups of maize lineages do not diverge simultaneously (Schneider *et al.* 2016 Fig2).

[1]- The region upstream of the *tb1* gene includes regions used to previously estimate the nucleotide substitution rate of $3.3 * 10^{-8}$ [149].

[2]- The previously reported domestication locus on the long arm of chromosome 10 [150] was split into three parts separated by two highly diverse regions.

[3]- The domestication locus "Region D" on the long arm of chromosome 5 [24] was split into two parts separated by a diverse region present in the P39/IL14H haplotype previously reported as a deletion [24].

[4]- The new centromere substitution rate was calculated using the weighted average from CEN1, CEN4 and CEN6.

[5]- The nucleotide substitution rate *r* was calculated assuming a divergence date of 10,000 years, based on the assumption that the regions examined represent domestication loci (Schneider et al, 2006).

	Start	End	Size
DEL1	91,931,286	92,108,592	177,306
DEL2	93,199,123	94,220,767	1,021,644

Table ST10. B73 v5 coordinates of the deletions in NC358/Oh43 and the estimated deletion size.

	COUNT
TOTAL POSITIONS	3,621
UNINFORMATIVE	848
POSITONS_WITH_>40Ns	11
TIL10 UNIQ	14
TIL14_UNIQ	14
Oh43_UNIQ	19
NC358_UNIQ	13
BKN033_UNIQ	8
BKN030_UNIQ	24
TIL10_NC358_Oh43_SHARED	4
TIL14_BKN030_BKN033_SHARED	9

Table ST11. HapMap3 SNP counts for different comparisons across the CEN2 domestication region.

Dataset	Start	Stop
CML277 Physical Map	95,665,767	95,665,463
B73_v5 Physical Map	94,389,390	94,389,694
HapMap3	94,228,993	94,229,297

Table ST12. Coordinates of the CML277/BKN009 introgression over Zm00001eb087250.

Gene Model v5	Gene Model v4	H2A.Z	RNAseq*
Zm00001eb087260	Zm00001d004236	N	N
Zm00001eb087270	None	N	N
Zm00001eb087280	None	Y	N
Zm00001eb087310	Zm00001d004238	Y	N
Zm00001eb087320	None	N	N
Zm00001eb087340	Zm00001d004242	Y	N

Table ST13. B73 v5 gene models between the recombination locus R3 and the B73 active centromere.
*RNAseq data was obtained from Root, Shoot, Leaf and Tassel tissues.

	CML322	Oh7B	P39	Il14H
US breakpoint window start	94,133,209	93,083,532	95,269,874	93,804,896
US breakpoint window end	94,160,992	93,111,316	95,297,654	93,832,678
DS breakpoint window start	122,172,029	120,801,296	123,679,988	121,843,932
DS breakpoint window end	122,189,213	120,818,480	123,697,172	121,861,116

Table ST14. Physical map coordinates for the CEN2 HAP2B inversion breakpoints in each of the four physical maps containing the inversion.

Position	Syn/N-Syn	nt Change	AA Change	Hap-Changed
20	N	A-G	N-S	H1
39	S	C-T		H1
225	S	C-T		H3
234	S	C-T		H1/H3
237	S	C-T		H2
525	N	A-T	E-D	H2
593	N	G-A	S-N	H2
635	N	C-T	A-V*	H1
692	N	C-T	S-F*	H1
696	S	T-C		H1
746	N	C-A	P-H	H1
827	N	A-G	N-S	H1
886	N	C-A	L-I	H2
969	S	G-A		H1
1008	S	A-T		H3
H1	B73 haplotype			
H2	W22 haplotype			
H3	CML69 haplotype			
* indicates an AA change to one with different chemical properties				

Table ST15. SNP analysis of the PHS1 gene. The nucleotide change is based on the ancestral Sorghum bicolor ortholog, the amino acid change if applicable, and the haplotype in which the SNP occurred.

	Deletion Junction 1
US	TTATCGAGCG|TTTTTTTCCT
DS	TATCGAGCGT|TTTTTTTTCC
	Deletion Junction 2
US	AAAAAAACAGA|AAAAAAAAGA
DS	AAAAAAACAG|AAAAAAAAAG

Table ST16. Deletion junctions for P39 Monomers in the CEN2 CTR1 locus. Sequence highlighted in red was deleted from the P39 region.

SUBSECTION	V5 START	V5 END	B73_SEQ_SIZE	ALN length	R2	Slope	v4 START	v4 END	HM3 START	HM3 END
500	135,003,349	135,317,514	314,165	500,000	0.42	16.393	135,199,904	135,514,112	135,199,962	135,514,106
1000	135,317,515	135,684,565	367,050	500,000	0.57	18.363	135,514,113	135,881,151	135,514,133	135,881,091
1500	135,684,566	135,985,412	300,846	500,000	0.52	18.847	135,881,152	136,192,967	135,881,172	136,192,938
2000	135,985,413	136,337,661	352,248	500,000	0.51	20.416	136,192,968	136,543,880	136,192,987	136,543,819
END	136,337,662	136,535,972	198,310	323,308	0.55	17.808	136,543,881	136,742,183	136,543,921	136,737,492

Table ST17. Coordinates of CEN1 alignment subsections used to search for potential local divergence maxima. No anomalies associated with misalignments were detected in any subsection.

	Physical Map Divergence			HapMap3 Divergence		
	MAX	MEAN	MEDIAN	MAX	MEAN	MEDIAN
CEN1	11.1	9.2	9.8	13.8	9.1	9.3
CEN4	11.2	6.7	7.7	14.6	8.6	10.0
CEN6	10.2	7.9	8.8	12.0	8.9	10.0

Table ST18. Comparisons of divergence date (ka) for single haplotype centromeres CEN1 CEN4 and CEN6.

Deletion	Deletion coords (G consensus)	Deletion size	# of elements	Insertion date range (ka)	Interrupted domains	Frameshift	Junctions
1	5716 - 6570	854	5	323 - 75	INT	N	CCACAC|GGTTTG-ATTGCA|GATTTA
2	4201 - 4734	533	3	0	RT	N	GGTTAC|GTGCGT-CCACAG|GGAATT

Table ST19. CR2 deletion variant details

	A	B	C	D	E	F	G	FxB	GxF_1	Gx(FxB)	FxG	GxD	GxE	(GxD)x(FxB)	GxF_2	Gx(FxE)	FxE	(FxE)xC
A																		
B	53.8																	
C	76.8	38.2																
D	78.7	40.0	25.4															
E	82.4	44.6	28.2	30.0														
F	90.7	52.0	37.3	35.5	17.2													
G	105.5	70.3	55.6	53.8	34.5	19.1												
FxB	99.9	63.9	52.8	54.7	39.1	20.9	40.9											
GxF_1	103.6	68.5	53.8	51.9	32.7	15.4	7.2	37.3										
Gx(FxB)	101.8	67.6	52.8	51.0	31.8	16.3	8.1	32.7	13.6									
FxG	103.6	68.5	53.8	51.9	32.7	17.2	1.8	39.1	5.4	10.0								
GxD	95.3	58.4	43.7	41.9	24.5	16.3	13.6	38.2	13.6	12.7	11.8							
GxE	97.1	62.1	47.3	45.5	26.3	18.1	8.1	40.0	13.6	9.1	10.0	7.2						
(GxD)x(FxB)	99.0	61.1	46.4	48.2	32.7	26.3	40.0	48.2	38.2	37.3	38.2	26.3	31.8					
GxF_2	99.0	62.1	47.3	45.5	26.3	9.1	10.0	30.9	10.0	9.1	11.8	10.9	9.1	31.8				
Gx(FxE)	101.8	64.8	50.1	48.3	27.3	20.9	14.5	42.7	16.3	15.4	16.3	10.0	10.0	34.5	11.8			
FxE	101.8	63.9	49.2	47.3	22.7	25.4	40.9	47.3	39.1	40.0	39.1	32.7	34.5	40.9	34.5	26.3		
(FxE)xC	90.7	54.7	29.1	38.2	24.5	25.4	42.7	47.3	40.9	40.0	40.9	30.9	34.5	33.6	34.5	37.3	30.9	

	R5	R4	R3	R2	R1	B	A
R5							
R4	507.7						
R3	582.7	149.1					
R2	718.3	285.7	204.0				
R1	130.0	582.8	654.6	602.9			
B	364.1	769.7	827.5	715.2	245.8		
A	694.5	234.8	170.5	226.5	747.7	873.1	

Table ST20. Divergence (ka) between CR1 (bottom) and CR2 (top) elements

Context	Dataset	Junction ID
NESTED IN CR1	J178_3_input_seq_PI_422162/Dawe_ChIP_PI_422162	54643410_54643610_CML69_chr6_GGCCA_GGCCA_R_0
NESTED IN CR1	J178_3_input_seq_PI_422162/Dawe_ChIP_PI_422162	54650982_54651182_CML69_chr6_GGCCA_GGCCA_R_0
NESTED IN CR1	J178_3_input_seq_PI_422162/Dawe_ChIP_PI_422162	61435530_61435730_CML277_chr9_TGTAG_TGTAG_R_0
NESTED IN CR1	J178_3_input_seq_PI_422162/Dawe_ChIP_PI_422162	61443102_61443302_CML277_chr9_TGTAG_TGTAG_R_0
NESTED IN CR1	Dawe_ChIP_PI_422162	49623497_49623697_CML69_chr8_TCGTC_TCGTC_F_7
NESTED IN CR1	Dawe_ChIP_PI_422162	49631065_49631265_CML69_chr8_TCGTC_TCGTC_F_7

NESTED IN CR1	Dawe_ChIP_PI_422162	107164404_107164604_CML333_chr5_AGCAG_AGCAG_R_11
NESTED IN CR1	Dawe_ChIP_PI_422162	107171968_107172168_CML333_chr5_AGCAG_AGCAG_R_11

Table ST21. Z. luxurians putative shared CR2s

a.	CEN_COUNT	TOTAL_COUNT	%_IN_CEN
A	3	27	11
B	4	39	10
R1	83	350	24
R2	13	75	17
R3	3	34	9
R4	95	217	44
R5	455	974	47
TOTAL	656	1716	38

b.	CEN_COUNT	TOTAL_COUNT	%_IN_CEN
A	2	29	7
B	13	58	22
C	20	37	54
D	35	53	66
E	11	21	52
F	41	59	69
G	741	1025	72
FxB	142	195	73
GxF_1	199	285	70
Gx(FxB)	81	110	74
FxG	148	190	78
GxD	98	141	70
GxE	120	168	71
(GxD)x(FxB)	246	327	75

GxF_2	43	65	66
Gx(FxE)	449	611	73
FxE	75	110	68
(FxE)xC	81	120	68
UNK	1103	1515	73
TOTAL	3648	5119	71

Table ST22. CR counts in the active centromere by subtype for CR1 (Top) and CR2 (Bottom)

Start	End	Chromosome	5' TSD	3' TSD	Orientation	Haplotype	Insertion date (ka)	Line
46609100	46615946	chr8	CTTTT	CTTTT	R	R1	163	Il14H
52244421	52251260	chr8	GTCTG	GTCTG	F	R1	82	B73
104489262	104496133	chr5	TGAGT	TGAGT	R	R2	80	B73
84353211	84360097	chr3	CCTTT	CCTTT	R	R4	73	P39
105512077	105518943	chr5	CTTCT	CTTCT	F	R5	12	B73
112282859	112289705	chr5	TTGCC	TTGCC	F	R5	12	CML247
104609476	104616345	chr5	GGAGG	GGAGG	R	R5	12	Il14H
105053491	105060352	chr5	TGTAC	TGTAC	R	R5	12	M162W
112582770	112589645	chr5	CTAAG	CTAAG	F	R5	9	CML69
110059173	110066052	chr5	ATCAC	ATCAC	F	R5	6	CML228
104967564	104974437	chr5	GAGAG	GAGAG	R	R5	0	B73
103059864	103066731	chr5	ACTTT	ACTTT	F	R5	0	B97
103417441	103424306	chr5	GGCAG	GGCAG	F	R5	0	B97
109918433	109925306	chr5	CATTG	CATTG	F	R5	0	CML228
102242374	102249234	chr5	GAATA	GAATA	F	R5	0	CML277
103657424	103664282	chr5	GGTTC	GGTTC	F	R5	0	CML277
103863666	103870532	chr5	AGGTT	AGGTT	R	R5	0	CML322
104732877	104739743	chr5	CTCTG	CTCTG	R	R5	0	CML322
113310923	113317797	chr5	GCTAC	GCTAC	F	R5	0	CML52

104697486	104704359	chr5	AGCAT	AGCAT	F	R5	0	Il14H
104482281	104489127	chr5	CTTCC	CCACC	F	R5	0	Il14H
105411273	105418141	chr5	TATTG	TATTG	F	R5	0	M37W
111795284	111802149	chr5	CTGAG	CTGAG	R	R5	0	P39
51899371	51906231	chr8	ACAAA	ACAAA	F	R5	0	B73
51200127	51206989	chr8	CCTCA	CCTCA	F	R5	0	B73
48902284	48909146	chr8	ATGAC	ATGAC	R	R5	0	Ms71
47699732	47706608	chr8	CATAC	CATAC	F	R5	0	Ms71
48453932	48460799	chr8	TAATC	TAATC	F	R5	0	P39

Table ST23. CR1s in post-domestication neocentromeres

	S1			S2			S3			S4			S5			S6		
Cultivar	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 1**	**Rep 2**	**Rep 3**
MR01	x	x	x	x									x			x		
MR08	x	x	x													x	x	x
MR09	x	x														x	x	x
MR18	x	x	x	x	x	x				x	x					x	x	x
MR19	x	x	x	x	x	x				x	x	x				x	x	x
MR22	x	x	x													x	x	

Table ST24. Datasets unavailable for analysis of centC loss. Unsequenced datasets are marked with an X. Various replicates and generations are missing from each cultivar. The first and last generation are particularly sparse.

a.	Chromosome	Syntenic Region Coordinates (Mb)	Region size (Mb)	Centromeric Region (Mb)	Total Centromeric Genes	Expressed Genes
	Os1	12.2 - 18.6	6.4	15.7 - 18.1	74	N/A
	Zm3	75.6 - 90.2	14.6	75.6 - 90.2	37	13
	Zm8	43.2 - 58.0	14.8	43.2 - 52.9	48	24
	Si5	18.6 - 22.3	3.7	18.6 - 22.3	34	N/A
b.	**Chromosome**	**Total Genes**	**Artifacts*** **	**Genes Retained***	**Genes Moved into Syntenic Regions****	**Genes moved out of Syntenic regions**
	Os1	599	340	3	N/A	N/A
	Zm3	52	15	3	4	96
	Zm8	92	12	3	11	96
	Si5	60	26	3	7	45
c.	**Chromosome**	**Zm3 orthos**	**Zm8 orthos**	**Genes with no orthos**	**Genes with orthos**	**Genes Acquired**** **
	Os1	92	75	351	248	N/A
	Zm3	N/A	22	27	25	12
	Zm8	22	N/A	35	57	17
	Si5	5	8	27	33	2

Table ST25. Syntenic coordinates and genes of Os1, Zm3 and Zm8. Gene counts and region coordinates for Os1, Zm3 and Zm8.

*To be retained a gene must be in OsCEN1 and have orthologs on Zm3, Zm8 and Si5

**In Os1 a gene was considered moved if it had a single Zm ortholog that was not on Zm3, Zm8

**In Zm3 a gene was considered moved if it had a single Os ortholog that was not on Os1 or Os12

**In Zm8 a gene was considered moved if it had a single Os ortholog that was not on Os1

**In Si5 a gene was considered moved if it had a single Os ortholog that was not on Os1

***In Os1 A gene was considered an artifact if it had no orthologs and no domain

***In Zm a gene was considered an artifact if it had no orthologs no domain and no expression

***In Si a gene was considered an artifact if it had no orthologs and no domain

****Genes were considered aquired in Zm if they had no orthologs in either Os or Si and had a domain OR were shown to be expressed in RNAseq data

SUPPLEMENTAL REFERENCES

1. Smith, M.R., et al., *Global Expanded Nutrient Supply (GENuS) Model: A New Method for Estimating the Global Dietary Supply of Nutrients.* PLOS ONE, 2016. **11**(1): p. e0146976.
2. Piperno, D.R. and K.V. Flannery, *The earliest archaeological maize (Zea mays L.) from highland Mexico: new accelerator mass spectrometry dates and their implications.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(4): p. 2101-2103.
3. Matsuoka, Y., et al., *A single domestication for maize shown by multilocus microsatellite genotyping.* Proceedings of the National Academy of Sciences, 2002. **99**(9): p. 6080.
4. Bennetzen, J., et al., *Genetic Evidence and the Origin of Maize.* Latin American Antiquity, 2001. **12**(1): p. 84-86.
5. Eyre-Walker, A., et al., *Investigation of the bottleneck leading to the domestication of maize.* Proceedings of the National Academy of Sciences, 1998. **95**(8): p. 4441.
6. Hilton, H. and B.S. Gaut, *Speciation and domestication in maize and its wild relatives: evidence from the globulin-1 gene.* Genetics, 1998. **150**(2): p. 863-872.
7. Bourguiba, H., et al., *Loss of genetic diversity as a signature of apricot domestication and diffusion into the Mediterranean Basin.* BMC plant biology, 2012. **12**: p. 49-49.
8. Reif, J.C., et al., *Wheat genetic diversity trends during domestication and breeding.* Theoretical and Applied Genetics, 2005. **110**(5): p. 859-864.
9. Tenaillon, M.I., et al., *Selection Versus Demography: A Multilocus Investigation of the Domestication Process in Maize.* Molecular Biology and Evolution, 2004. **21**(7): p. 1214-1225.
10. Doebley, J. and A. Stec, *Genetic analysis of the morphological differences between maize and teosinte.* Genetics, 1991. **129**(1): p. 285-295.
11. Doebley, J. and A. Stec, *Inheritance of the morphological differences between maize and teosinte: comparison of results for two F2 populations.* Genetics, 1993. **134**(2): p. 559-570.
12. Doebley, J., et al., *Genetic and morphological analysis of a maize-teosinte F2 population: implications for the origin of maize.* Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(24): p. 9888-9892.
13. Dorweiler, J., et al., *Teosinte glume architecture 1: A Genetic Locus Controlling a Key Step in Maize Evolution.* Science, 1993. **262**(5131): p. 233.
14. Hufford, M.B., et al., *Comparative population genomics of maize domestication and improvement.* Nature Genetics, 2012. **44**(7): p. 808-811.
15. Mahtani, M.M. and H.F. Willard, *Physical and genetic mapping of the human X chromosome centromere: repression of recombination.* (1088-9051 (Print)).
16. Stitzer, M.C. and J. Ross-Ibarra, *Maize domestication and gene interaction.* New Phytologist, 2018. **220**(2): p. 395-408.
17. Nannas, N.J. and R.K. Dawe, *Genetic and genomic toolbox of Zea mays.* Genetics, 2015. **199**(3): p. 655-669.
18. G.W., B., *A Possible Influence of the Spindle Fibre on Crossing-Over in Drosophila.* PNAS, A Possible Influence of the Spindle Fibre on Crossing-Over in Drosophila. **18**: p. 160-165.
19. ALBERT LEVAN, K.F.a.A.A.S., *NOMENCLATURE FOR CENTROMERIC POSITION ON CHROMOSOMES*, L. INSTITUTE OF GENETICS, SWEDEN, AND ROSWELL PARK MEMORIAL INSTITUTE, Editor. 1964: BUFFALO, N.Y., U.S.A.
20. Neumann, P., et al., *Stretching the Rules: Monocentric Chromosomes with Multiple Centromere Domains.* PLoS Genet, 2012. **8**(6): p. 1 -13
21. Kiyotaka Nagaki, K.K., and Minoru Murata, *Visualization of Diffuse Centromeres with Centromere-Specific Histone H3 in the Holocentric Plant Luzula nivea.* the Plant Cell, 2005. **17**: p. 1886-1893.

22. Zachary Lippman1*, A.-V.r.G., Michael Black3†,, et al., *Role of transposable elements in heterochromatin and epigenetic control.* Nature, 2004. **430**(6998): p. 467-71.
23. Andreas Houben, T.W., Rieko Furushima-, G.P. Shimogawara, Gottfried KuÈ nzel, Ingo, and S.a.K.c. Fukui, *The cell cycle dependant phosphorylation of histone H3 is correlated with the condensation of plant mitotic chromosomes.* The Plant Journal, 1999. **16**(6): p. 675 -679.
24. Schneider, K.L., et al., *Inbreeding drives maize centromere evolution.* Proc Natl Acad Sci U S A, 2016. **113**(8): p. E987-96.
25. Jiming Jiang, J.A.B., Wayne A. Parrott and R. Kelly Dawe, *A molecular view of plant centromeres.* Trends in Plant Science, 2003. **18**(12): p. 570-575.
26. Bodor, D.L., et al., *The quantitative architecture of centromeric chromatin.* eLife, 2014. **3**: p. e02137.
27. Tachiwana, H., et al., *Crystal structure of the human centromeric nucleosome containing CENP-A.* Nature, 2011. **476**: p. 232.
28. Dalal, Y., et al., *Structure, dynamics, and evolution of centromeric nucleosomes.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(41): p. 15974-15981.
29. Henikoff, S. and T. Furuyama, *The unconventional structure of centromeric nucleosomes.* Chromosoma, 2012. **121**(4): p. 341-352.
30. Black, B.E. and D.W. Cleveland, *Epigenetic Centromere Propagation and the Nature of CENP-A Nucleosomes.* Cell, 2011. **144**(4): p. 471-479.
31. Smith, M.M., *Centromeres and variant histones: what, where, when and why?* Current Opinion in Cell Biology, 2002. **14**(279-285).
32. Hasson, D., et al., *The octamer is the major form of CENP-A nucleosomes at human centromeres.* Nature structural & molecular biology, 2013. **20**(6): p. 687-695.
33. Annika Guse, C.W.C., Ben Moree, Colin J. Fuller & Aaron F. Straight, *In vitro centromere and kinetochore assembly on defined chromatin templates.* Nature, 2011. **477**(15): p. 354-358.
34. Fukagawa, T. and William C. Earnshaw, *The Centromere: Chromatin Foundation for the Kinetochore Machinery.* Developmental Cell, 2014. **30**(5): p. 496-508.
35. Blower, M.D., B.A. Sullivan, and G.H. Karpen, *Conserved Organization of Centromeric Chromatin in Flies and Humans.* Developmental cell, 2002. **2**(3): p. 319-330.
36. Nishino, T., et al., *CENP-T-W-S-X forms a unique centromeric chromatin structure with a histone-like fold.* Cell, 2012. **148**(3): p. 487-501.
37. Yu, H.-G., et al., *The plant kinetochore.* Trends in Plant Science, 2000. **5**(12): p. 543-547.
38. Braselton, J.P. and C.C. Bowen, *The Ultrastructure of the Kinetochores of Lilium Longiflorum During the first Meiotic Division.* Caryologia, 1971. **24**(1): p. 49-58.
39. Baskin, T.I. and W.Z. Cande, *The Structure and Function of the Mitotic Spindle in Flowering Plants.* Annual Review of Plant Physiology and Plant Molecular Biology, 1990. **41**(1): p. 277-315.
40. Dawe, R.K., et al., *A Maize Homolog of Mammalian CENPC Is a Constitutive Component of the Inner Kinetochore.* The Plant Cell, 1999. **11**(7): p. 1227.
41. Roudier, F., et al., *Integrative epigenomic mapping defines four main chromatin states in Arabidopsis.* The EMBO journal, 2011. **30**(10): p. 1928-1938.
42. Fransz, P., R. ten Hoopen, and F. Tessadori, *Composition and formation of heterochromatin in Arabidopsis thaliana.* Chromosome Research, 2006. **14**(1): p. 71-82.
43. Chodavarapu, R.K., et al., *Relationship between nucleosome positioning and DNA methylation.* Nature, 2010. **466**(7304): p. 388-392.
44. Bauwens, S., et al., *Distribution of the rDNA and three classes of highly repetitive DNA in the chromatin of interphase nuclei of Arabidopsis thaliana.* Chromosoma, 1991. **101**(1): p. 41-48.

45. Nasmyth, K. and C.H. Haering, *Cohesin: Its Roles and Mechanisms.* Annual Review of Genetics, 2009. **43**(1): p. 525-558.
46. Gent, J.I., et al., *Strong epigenetic similarity between maize centromeric and pericentromeric regions at the level of small RNAs, DNA methylation and H3 chromatin modifications.* Nucleic acids research, 2012. **40**(4): p. 1550-1560.
47. Yuan, G.-C., et al., *Genome-Scale Identification of Nucleosome Positions in <em>S. cerevisiae</em&gt.* Science, 2005. **309**(5734): p. 626.
48. Raisner, R.M., et al., *Histone Variant H2A.Z Marks the 5′ Ends of Both Active and Inactive Genes in Euchromatin.* Cell, 2005. **123**(2): p. 233-248.
49. Santisteban, M.S., T. Kalashnikova, and M.M. Smith, *Histone H2A.Z Regulates Transcription and Is Partially Redundant with Nucleosome Remodeling Complexes.* Cell, 2000. **103**(3): p. 411-422.
50. Meneghini, M.D., M. Wu, and H.D. Madhani, *Conserved Histone Variant H2A.Z Protects Euchromatin from the Ectopic Spread of Silent Heterochromatin.* Cell, 2003. **112**(5): p. 725-736.
51. Zilberman, D., et al., *Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks.* Nature, 2008. **456**(7218): p. 125-129.
52. Dion, M.F., et al., *Dynamics of Replication-Independent Histone Turnover in Budding Yeast.* Science, 2007. **315**(5817): p. 1405.
53. Mizuguchi, G., et al., *ATP-Driven Exchange of Histone H2AZ Variant Catalyzed by SWR1 Chromatin Remodeling Complex.* Science, 2004. **303**(5656): p. 343.
54. Zemach, A., et al., *Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation.* Science, 2010. **328**(5980): p. 916.
55. Conerly, M.L., et al., *Changes in H2A.Z occupancy and DNA methylation during B-cell lymphomagenesis.* Genome research, 2010. **20**(10): p. 1383-1390.
56. Murphy, P.J., et al., *Placeholder Nucleosomes Underlie Germline-to-Embryo DNA Methylation Reprogramming.* Cell, 2018. **172**(5): p. 993-1006.e13.
57. Yamada, S., et al., *The histone variant H2A.Z promotes initiation of meiotic recombination in fission yeast.* (1362-4962 (Electronic)).
58. Choi, K., et al., *Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters.* Nature Genetics, 2013. **45**(11): p. 1327-1336.
59. Shi, J., et al., *Widespread gene conversion in centromere cores.* PLoS Biol, 2010. **8**(3): p. e1000327.
60. Bennetzen, J.M.a.J.L., *Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice.* Proc Natl Acad Sci U S A, 2006. **103**(2): p. 383-388.
61. Wolfgruber, T.K., et al., *High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events.* Front Plant Sci, 2016. **7**: p. 308.
62. Tanksley, S.D., et al., *High density molecular linkage maps of the tomato and potato genomes.* Genetics, 1992. **132**(4): p. 1141-1160.
63. Jacobs, P.A., *Mutation rates of structural chromosome rearrangements in man.* American journal of human genetics, 1981. **33**(1): p. 44-54.
64. Murat, F., et al., *Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution.* Genome research, 2010. **20**(11): p. 1545-1557.
65. Bowers, J.E., et al., *Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(37): p. 13206-13211.
66. Marshall, O.J., et al., *Neocentromeres: New Insights into Centromere Structure, Disease Development, and Karyotype Evolution.* American Journal of Human Genetics, 2008. **82**(2): p. 261-282.

67. Buckler, E.S., et al., *Meiotic drive of chromosomal knobs reshaped the maize genome.* Genetics, 1999. **153**(1): p. 415-426.
68. Alonso, A., et al., *A paucity of heterochromatin at functional human neocentromeres.* Epigenetics & Chromatin, 2010. **3**: p. 6-6.
69. Zhang, B., et al., *Formation of a functional maize centromere after loss of centromeric sequences and gain of ectopic sequences.* Plant Cell, 2013. **25**(6): p. 1979-89.
70. Nasuda, S., et al., *Stable barley chromosomes without centromeric repeats.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(28): p. 9842-9847.
71. Yan, H., et al., *Intergenic Locations of Rice Centromeric Chromatin.* PLOS Biology, 2008. **6**(11): p. e286.
72. Nagaki, K., et al., *Sequencing of a rice centromere uncovers active genes.* Nat Genet, 2004. **36**(2): p. 138-145.
73. Yasuhara, J.C. and B.T. Wakimoto, *Oxymoron no more: the expanding world of heterochromatic genes.* Trends in Genetics, 2006. **22**(6): p. 330-338.
74. Miller, J.T., et al., *Retrotransposon-related DNA sequences in the centromeres of grass chromosomes.* Genetics, 1998. **150**(4): p. 1615-1623.
75. Benjamin Gorinsek, F.G., and Dusan Kordis, *Evolutionary Genomics of Chromoviruses in Eukaryotes.* Molecular Biology and Evolution, 2004. **21**(5): p. 781-798.
76. Pavel Neumann, A.N., Andrea Koblížková, Eduard Kejnovský, Eva Hřibová3, Roman Hobza, Alex Widmer, Jaroslav Doležel, Jiří Macas, *Plant centromeric retrotransposons: a structural and cytogenetic perspective.* Mobile DNA, 2011. **2**(4): p. 1-16.
77. Gernot G. Presting, L.M., Jorg Fuchs and Ingo Schubert, *A Ty3/Gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes.* The Plant Journal, 1998. **16**(6): p. 721 - 728.
78. Gao, X., et al., *Chromodomains direct integration of retrotransposons to heterochromatin.* Genome research, 2008. **18**(3): p. 359-369.
79. Sharma, A. and G.G. Presting, *Evolution of centromeric retrotransposons in grasses.* Genome Biol Evol, 2014. **6**(6): p. 1335-52.
80. Phillip SanMiguel, Brandon S. Gaut, Alexander Tikhonov, Yuko Nakajima & Jeffrey L. Bennetzen, *The paleontology of intergene retrotransposons of maize.* Nature Genetics, 1998. **20**: p. 43-45.
81. Cédric Feschotte, N.J.a.S.R.W., *Plant transposable elements: Where genetics meets genomics.* Nature Reviews Genetics, 2002. **3**: p. 329-341.
82. Sharma, A., K.L. Schneider, and G.G. Presting, *Sustained retrotransposition is mediated by nucleotide deletions and interelement recombinations.* Proc Natl Acad Sci U S A, 2008. **105**(40): p. 15470-4.
83. Hu, W.-S. and H.M. Temin, *Retroviral Recombination and Reverse Transcription.* Science, 1990. **250**(4985): p. 1227-1233.
84. Anupma Sharma, T.K.W.a.G.G.P., *Tandem repeats derived from centromeric retrotransposons.* Biomed Central Genomics, 2013. **14**(142): p. 1 - 10.
85. Bennetzen, J.L., J. Ma, and K.M. Devos, *Mechanisms of Recent Genome Size Variation in Flowering Plants.* Annals of Botany, 2005. **95**(1): p. 127-132.
86. Zhang, Q.-J. and L.-Z. Gao, *Rapid and Recent Evolution of LTR Retrotransposons Drives Rice Genome Evolution During the Speciation of AA-Genome Oryza Species.* G3: Genes|Genomes|Genetics, 2017. **7**(6): p. 1875-1885.
87. BELYAYEV, A., *Bursts of transposable elements as an evolutionary driving force.* Evolutionary Biology, 2014. **27**: p. 2573-2584.
88. Ma, J., K.M. Devos, and J.L. Bennetzen, *Analyses of LTR-Retrotransposon Structures Reveal Recent and Rapid Genomic DNA Loss in Rice.* Genome Research, 2004. **14**(5): p. 860-869.

89. Vitte, C., O. Panaud, and H. Quesneville, *LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss.* BMC Genomics, 2007. **8**: p. 218-218.
90. Tian, Z., et al., *Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons?* Genome Research, 2009. **19**(12): p. 2221-2230.
91. Devos, K.M., J.K.M. Brown, and J.L. Bennetzen, *Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis.* Genome Research, 2002. **12**(7): p. 1075-1079.
92. McClintock, B., *Mutable loci in maize.* Carnegie Institution of Washington Yearbook, 1951. **50**: p. 174-181.
93. McClintock, B., *Controlling elements and the gene.* Cold Spring Harbor symposia on quantitative biology, 1956. **21**: p. 197.
94. Walsh, A.M., et al., *Widespread horizontal transfer of retrotransposons.* Proceedings of the National Academy of Sciences, 2013. **110**(3): p. 1012.
95. Jordan, I.K., L.V. Matyunina, and J.F. McDonald, *Evidence for the recent horizontal transfer of long terminal repeat retrotransposon.* Proceedings of the National Academy of Sciences, 1999. **96**(22): p. 12621.
96. Schubert, R.t.H.R.M.J.D.L.M.I., *Evolutionary conservation of kinetochore protein sequences in plants.* Chromosoma, 2000. **109**: p. 482-489.
97. Stephen Henikoff, K.A., Harmit S. Malik, *The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA.* Science, 2001. **293**: p. 1098-1102.
98. Gershenson, S., *A New Sex-Ratio Abnormality in DROSOPHILA OBSCURA.* Genetics, 1928. **13**(6): p. 488-507.
99. Chmátal, L., et al., *Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice.* (1879-0445 (Electronic)).
100. Rhoades, M.M., *PREFERENTIAL SEGREGATION IN MAIZE.* Genetics, 1942. **27**(4): p. 395.
101. Dawe, R.K. and W.Z. Cande, *Induction of centromeric activity in maize by suppressor of meiotic drive 1.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(16): p. 8512-8517.
102. Edward S. Buckler IV, T.L.P.-D., Carlyn S. Keith Buckler, R. Kelly Dawe, John F. Doebley and Timothy P. Holtsford, *Meiotic Drive of Chromosomal Knobs Reshaped the Maize Genome.* Genetics, 1999. **153**: p. 415-426.
103. Amor, D.J. and K.H.A. Choo, *Neocentromeres: role in human disease, evolution, and centromere study.* American journal of human genetics, 2002. **71**(4): p. 695-714.
104. Schnable, P.S., et al., *The B73 Maize Genome: Complexity, Diversity, and Dynamics.* Science, 2009. **326**(5956): p. 1112.
105. Jiao, Y., et al., *Improved maize reference genome with single-molecule technologies.* Nature, 2017. **546**(7659): p. 524-527.
106. Hufford, M.B., et al., *De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes.* Science, 2021. **373**(6555): p. 655.
107. Kurtz, S., et al., *Versatile and open software for comparing large genomes.* Genome biology, 2004. **5**(2): p. R12-R12.
108. Wolfgruber, T.K. and G.G. Presting, *JunctionViewer: customizable annotation software for repeat-rich genomic regions.* BMC bioinformatics, 2010. **11**: p. 23-23.
109. Ananiev, E.V., R.L. Phillips, and H.W. Rines, *Chromosome-specific molecular organization of maize (Zea mays L.) centromeric regions.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(22): p. 13073-13078.

110. Sharma, A. and G.G. Presting, *Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity.* Molecular Genetics and Genomics, 2008. **279**(2): p. 133-147.
111. Sharma, A. and G.G. Presting, *Evolution of centromeric retrotransposons in grasses.* Genome biology and evolution, 2014. **6**(6): p. 1335-1352.
112. Page, B.T., M.K. Wanous, and J.A. Birchler, *Characterization of a maize chromosome 4 centromeric sequence: evidence for an evolutionary relationship with the B chromosome centromere.* Genetics, 2001. **159**(1): p. 291-302.
113. Kearse, M., et al., *Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data.* Bioinformatics, 2012. **28**(12): p. 1647-1649.
114. Albert, P.S., et al., *Diversity of Chromosomal Karyotypes in Maize and Its Relatives.* Cytogenetic and Genome Research, 2010. **129**(1-3): p. 6-16.
115. Ogiyama, Y., et al., *Epigenetically induced paucity of histone H2A.Z stabilizes fission-yeast ectopic centromeres.* Nature Structural & Molecular Biology, 2013. **20**(12): p. 1397-1406.
116. Shang, W.-H., et al., *Chromosome engineering allows the efficient isolation of vertebrate neocentromeres.* Developmental cell, 2013. **24**(6): p. 635-648.
117. Maggert, K.A. and G.H. Karpen, *The activation of a neocentromere in Drosophila requires proximity to an endogenous centromere.* Genetics, 2001. **158**(4): p. 1615-1628.
118. Olszak, A.M., et al., *Heterochromatin boundaries are hotspots for de novo kinetochore formation.* Nature Cell Biology, 2011. **13**(7): p. 799-808.
119. Nishimura, K., et al., *3D genomic architecture reveals that neocentromeres associate with heterochromatin regions.* The Journal of cell biology, 2019. **218**(1): p. 134-149.
120. Sura, W., et al., *Dual Role of the Histone Variant H2A.Z in Transcriptional Regulation of Stress-Response Genes.* The Plant Cell, 2017. **29**(4): p. 791-807.
121. Presting, G.G., *Centromeric retrotransposons and centromere function.* Current Opinion in Genetics & Development, 2018. **49**: p. 79-84.
122. Bauwens, S., et al., *Distribution of the rDNA and three classes of highly repetitive DNA in the chromatin of interphase nuclei of Arabidopsis thaliana.* Chromosoma, 2004. **101**: p. 41-48.
123. Zlatanova, J. and A. Thakar, *H2A.Z: View from the Top.* Structure, 2008. **16**(2): p. 166-179.
124. Neumann, P., H. Yan, and J. Jiang, *The centromeric retrotransposons of rice are transcribed and differentially processed by RNA interference.* Genetics, 2007. **176**(2): p. 749-761.
125. Lisch, D., *Epigenetic Regulation of Transposable Elements in Plants.* Annual Review of Plant Biology, 2009. **60**(1): p. 43-66.
126. Lermontova, I., et al., *Loading of Arabidopsis centromeric histone CENH3 occurs mainly during G2 and requires the presence of the histone fold domain.* The Plant cell, 2006. **18**(10): p. 2443-2451.
127. Potok, M.E., et al., *Arabidopsis SWR1-associated protein methyl-CpG-binding domain 9 is required for histone H2A.Z deposition.* Nature Communications, 2019. **10**(1): p. 3352.
128. Stellfox, M.E., A.O. Bailey, and D.R. Foltz, *Putting CENP-A in its place.* Cellular and molecular life sciences : CMLS, 2013. **70**(3): p. 387-406.
129. Talbert, P.B. and S. Henikoff, *Environmental responses mediated by histone variants.* Trends in Cell Biology, 2014. **24**(11): p. 642-650.
130. Kumar, S.V. and P.A. Wigge, *H2A.Z-Containing Nucleosomes Mediate the Thermosensory Response in Arabidopsis.* Cell, 2010. **140**(1): p. 136-147.
131. Berriri, S., S.N. Gangappa, and S.V. Kumar, *SWR1 Chromatin-Remodeling Complex Subunits and H2A.Z Have Non-overlapping Functions in Immunity and Gene Regulation in Arabidopsis.* Molecular plant, 2016. **9**(7): p. 1051-1065.

132. Smith, A.P., et al., *Histone H2A.Z regulates the expression of several classes of phosphate starvation response genes but not as a transcriptional activator.* Plant physiology, 2010. **152**(1): p. 217-225.
133. Laspisa, D.J., K.L. Schneider, and G.G. Presting, *Improved Centromere Assemblies for RefGen_v4.* bioRxiv, 2022: p. 2022.03.30.486462.
134. Burrack, L.S. and J. Berman, *Neocentromeres and epigenetically inherited features of centromeres.* Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 2012. **20**(5): p. 607-619.
135. Ishii, K., et al., *Heterochromatin Integrity Affects Chromosome Reorganization After Centromere Dysfunction.* Science, 2008. **321**(5892): p. 1088-1091.
136. Topp, C.N. and R.K. Dawe, *Reinterpreting pericentromeric heterochromatin.* Current Opinion in Plant Biology, 2006. **9**(6): p. 647-653.
137. Glémin, S. and T. Bataillon, *A comparative view of the evolution of grasses under domestication.* (1469-8137 (Electronic)).
138. Haudry, A., et al., *Grinding up wheat: a massive loss of nucleotide diversity since domestication.* (0737-4038 (Print)).
139. Sanou, H., et al., *Loss of Genetic Diversity of Jatropha curcas L. through Domestication: Implications for Its Genetic Improvement.* Crop Science, 2015. **55**(2): p. 749-759.
140. Dixon, T.J., et al., *Shifts in genetic diversity during domestication of Black Tiger shrimp, Penaeus monodon, monitored using two multiplexed microsatellite systems.* Aquaculture, 2008. **283**(1-4): p. 1-6.
141. Choi, K., *Advances towards Controlling Meiotic Recombination for Plant Breeding.* (0219-1032 (Electronic)).
142. Lyzenga, W.J., C.J. Pozniak, and S. Kagale, *Advanced domestication: harnessing the precision of gene editing in crop breeding.* Plant Biotechnology Journal, 2021. **19**(4): p. 660-670.
143. May, K.M., et al., *The parental origin of the extra X chromosome in 47,XXX females.* (0002-9297 (Print)).
144. Sears, D.D., et al., *Cis-acting determinants affecting centromere function, sister-chromatid cohesion and reciprocal recombination during meiosis in Saccharomyces cerevisiae.* (0016-6731 (Print)).
145. Nambiar, M. and G.R. Smith, *Repression of harmful meiotic recombination in centromeric regions.* (1096-3634 (Electronic)).
146. Underwood, C.J., et al., *Epigenetic activation of meiotic recombination near Arabidopsis thaliana centromeres via loss of H3K9me2 and non-CG DNA methylation.* (1549-5469 (Electronic)).
147. Darling, A.E., B. Mau, and N.T. Perna, *progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement.* PLOS ONE, 2010. **5**(6): p. e11147.
148. Kumar, S., et al., *MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis.* Bioinformatics (Oxford, England), 2012. **28**(20): p. 2685-2686.
149. Clark, R.M., S. Tavare, and J. Doebley, *Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus.* Mol Biol Evol, 2005. **22**(11): p. 2304-12.
150. Tian, F., N.M. Stevens, and E.S. Buckler, *Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10.* Proceedings of the National Academy of Sciences, 2009. **106**(Supplement 1): p. 9979.
151. Andorf, C.M., et al., *MaizeGDB update: new tools, data and interface for the maize model organism database.* Nucleic Acids Research, 2016. **44**(D1): p. D1195-D1201.
152. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.* Molecular biology and evolution, 2018. **35**(6): p. 1547-1549.

153. Bukowski, R., et al., *Construction of the third-generation Zea mays haplotype map.* GigaScience, 2018. **7**(4): p. 1-12.
154. Fang, Z., et al., *Megabase-scale inversion polymorphism in the wild ancestor of maize.* (1943-2631 (Electronic)).
155. Fuentes, R.R., et al., *Domestication Shapes Recombination Patterns in Tomato. LID - 10.1093/molbev/msab287 [doi] LID - msab287.* (1537-1719 (Electronic)).
156. Zapata, L., et al., *Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms.* Proceedings of the National Academy of Sciences, 2016. **113**(28): p. E4052-E4060.
157. Bansal, V., A. Bashir, and V. Bafna, *Evidence for large inversion polymorphisms in the human genome from HapMap data.* Genome research, 2007. **17**(2): p. 219-230.
158. Munté, A., et al., *Chromosomal inversion polymorphism leads to extensive genetic structure: a multilocus survey in Drosophila subobscura.* Genetics, 2005. **169**(3): p. 1573-1581.
159. Sun, S., et al., *Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes.* Nature Genetics, 2018. **50**(9): p. 1289-1295.
160. Lemmon, Z.H., et al., *The Role of cis Regulatory Evolution in Maize Domestication.* PLOS Genetics, 2014. **10**(11): p. e1004745.
161. Danilevskaya, O.N., X. Meng, and E.V. Ananiev, *Concerted modification of flowering time and inflorescence architecture by ectopic expression of TFL1-like genes in maize.* Plant physiology, 2010. **153**(1): p. 238-251.
162. Beissinger, T.M., et al., *Recent demography drives changes in linked selection across the maize genome.* Nature Plants, 2016. **2**(7): p. 16084.
163. Petit, M., et al., *Mobilization of retrotransposons in synthetic allotetraploid tobacco.* New Phytologist, 2010. **186**(1): p. 135-147.
164. O'Neill, R.J.W., M.J. O'Neill, and J.A.M. Graves, *Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid.* Nature, 1998. **393**(6680): p. 68-72.
165. Cavrak, V.V., et al., *How a retrotransposon exploits the plant's heat stress response for its activation.* PLoS genetics, 2014. **10**(1): p. e1004115-e1004115.
166. McClintock, B., *The origin and behavior of mutable loci in maize.* Proceedings of the National Academy of Sciences, 1950. **36**(6): p. 344.
167. Griffith, J.L., et al., *Functional genomics reveals relationships between the retrovirus-like Ty1 element and its host Saccharomyces cerevisiae.* Genetics, 2003. **164**(3): p. 867-879.
168. Yang, N. and H.H. Kazazian, *L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells.* Nature Structural & Molecular Biology, 2006. **13**(9): p. 763-771.
169. Xie, Z., et al., *Genetic and functional diversification of small RNA pathways in plants.* PLoS biology, 2004. **2**(5): p. E104-E104.
170. Xie, M. and B. Yu, *siRNA-directed DNA Methylation in Plants.* Current genomics, 2015. **16**(1): p. 23-31.
171. Hufford, M.B., et al., *De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes.* bioRxiv, 2021: p. 2021.01.14.426684.
172. Katoh, K. and D.M. Standley, *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.* Molecular Biology and Evolution, 2013. **30**(4): p. 772-780.
173. Tenaillon, M.I., et al., *Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and Zea luxurians.* Genome Biology and Evolution, 2011. **3**: p. 219-229.

174. Chia, J.-M., et al., *Maize HapMap2 identifies extant variation from a genome in flux.* Nature Genetics, 2012. **44**(7): p. 803-807.
175. Bukowski, R., et al., *Construction of the third generation Zea mays haplotype map.* GigaScience, 2017: p. gix134-gix134.
176. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Meth, 2012. **9**(4): p. 357-359.
177. Feng, Y.X., et al., *The genomic RNA in Ty1 virus-like particles is dimeric.* Journal of virology, 2000. **74**(22): p. 10819-10821.
178. Ross-Ibarra, J., M. Tenaillon, and B.S. Gaut, *Historical divergence and gene flow in the genus Zea.* Genetics, 2009. **181**(4): p. 1399-1413.
179. El Baidouri, M., et al., *Widespread and frequent horizontal transfers of transposable elements in plants.* Genome research, 2014. **24**(5): p. 831-838.
180. Koo, D.-H., et al., *Distinct DNA methylation patterns associated with active and inactive centromeres of the maize B chromosome.* Genome research, 2011. **21**(6): p. 908-914.
181. Du, C., Z. Swigoňová, and J. Messing, *Retrotranspositions in orthologous regions of closely related grass species.* BMC Evolutionary Biology, 2006. **6**(1): p. 62.
182. Estep, M.C., J.D. DeBarry, and J.L. Bennetzen, *The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution.* Heredity, 2013. **110**(2): p. 194-204.
183. Sharma, A., T.K. Wolfgruber, and G.G. Presting, *Tandem repeats derived from centromeric retrotransposons.* BMC Genomics, 2013. **14**(1): p. 142.
184. Schotanus, K., V. Yadav, and J. Heitman, *Epigenetic dynamics of centromeres and neocentromeres in Cryptococcus deuterogattii.* PLOS Genetics, 2021. **17**(8): p. e1009743.
185. Barra, V. and D. Fachinetti, *The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA.* Nature Communications, 2018. **9**(1): p. 4340.
186. van Heerwaarden, J., et al., *Genetic signals of origin, spread, and introgression in a large sample of maize landraces.* Proceedings of the National Academy of Sciences, 2011. **108**(3): p. 1088.
187. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Research, 1999. **27**(2): p. 573–580.
188. Richard, G.F., A. Kerrest, and B. Dujon, *Comparative genomics and molecular dynamics of DNA repeats in eukaryotes.* Microbiol Mol Biol Rev, 2008. **72**(4): p. 686-727.
189. Dawe, R.K., *Centromere renewal and replacement in the plant kingdom.* 102, 2005. **33**(11573–11574).
190. Melters, D.P., et al., *Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution.* Genome Biol, 2013. **14**(1): p. R10.
191. Plohl, M., et al., *Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin.* Gene, 2008. **409**(1-2): p. 72-82.
192. Ross-Ibarra, P.B.K.D.J.G.-L.G.M.M.J.S.R.K.D.J., *Diversity and evolution of centromere repeats in the maize genome.* Chomosoma, 2014. **124**: p. 57-65.
193. FRE´DE´RIC PAˆQUES, W.-Y.L., AND JAMES E. HABER, *Expansions and Contractions in a Tandem Repeat Induced by Double-Strand Break Repair.* Molecular and Cellular Biology, 1998. **18**(4): p. 2045–2054.
194. M. Katharine Rudd, G.A.W., and Huntington F. Willard, *The evolutionary dynamics of a-satellite.* Genome Research, 2006. **16**: p. 88-96.
195. Walker, P.M.B., *Origin of Satellite DNA.* Nature, 1971. **229**(5283): p. 306-308.
196. Phillip SanMiguel, A.T., Young-Kwan Jin,, et al., *Nested retrotransposons in the intergenic regions of the maize genome.* Science, 1996. **274**: p. 765-768.

197. Zuccolo, A., et al., *Transposable element distribution, abundance and role in genome size variation in the genus Oryza.* BMC Evolutionary Biology, 2007. **7**(1): p. 152.
198. Jr.*, H.H.K., *Mobile elements: Drivers of genome evolution.* Science, 2004. **303**: p. 1626-1632.
199. McDonald, J.F., *Evolution and consequences of transposable elements* Current Opinion in Genetics and Development 1993. **3**: p. 855-864.
200. Lerat, E., *Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs.* Nature Heredity, 2009. **104**: p. 520-533.
201. Chang, W. and A.H. Schulman, *BARE retrotransposons produce multiple groups of rarely polyadenylated transcripts from two differentially regulated promoters.* The Plant Journal, 2008. **56**(1): p. 40-50.
202. Sabot, F. and A.H. Schulman, *Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome.* Heredity, 2006. **97**: p. 381.
203. Jääskeläinen, M., et al., *Retrotransposon BARE-1: expression of encoded proteins and formation of virus-like particles in barley cells.* The Plant Journal, 1999. **20**(4): p. 413-422.
204. Pochart, P., et al., *Biochemical and electron microscope analyses of the DNA reverse transcripts present in the virus-like particles of the yeast transposon Ty1. Identification of a second origin of Ty1DNA plus strand synthesis.* Nucleic acids research, 1993. **21**(15): p. 3513-3520.
205. Gómez-Orte, E., C.M. Vicient, and J.A. Martínez-Izquierdo, *Grande retrotransposons contain an accessory gene in the unusually long 3'-internal region that encodes a nuclear protein transcribed from its own promoter.* Plant Molecular Biology, 2013. **81**(6): p. 541-551.
206. Engelman, L.K.a.A., *Retroviral Integrase Proteins and HIV-1 DNA Integration.* Journal of Biological Chemistry, 2012. **287**: p. 40858-40866.
207. Lanchy, J.-M., J.D. Ivanovitch, and J.S. Lodmell, *A structural linkage between the dimerization and encapsidation signals in HIV-2 leader RNA.* RNA (New York, N.Y.), 2003. **9**(8): p. 1007-1018.
208. Schmidt, T., *LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes.* Plant Molecular Biology, 1999. **40**(6): p. 903-910.
209. Xiong, Y. and T.H. Eickbush, *Origin and evolution of retroelements based upon their reverse transcriptase sequences.* The EMBO Journal, 1990. **9**(10): p. 3353-3362.
210. Kumar, A., *The evolution of plant retroviruses: moving to green pastures.* Trends in Plant Science, 1998. **3**(10): p. 371-374.
211. Kubis, S.E., et al., *The genomic organization of non-LTR retrotransposons (LINEs) from three Beta species and five other angiosperms.* Plant Molecular Biology, 1998. **36**(6): p. 821-831.
212. Schmidt, T. and J.S. Heslop-Harrison, *Genomes, genes and junk: the large-scale organization of plant chromosomes.* Trends in Plant Science, 1998. **3**(5): p. 195-199.
213. Wessler, S.R., *Transposable elements and the evolution of eukaryotic genomes.* Proceedings of the National Academy of Sciences, 2006. **103**(47): p. 17600-17601.
214. Feschotte, C. and E.J. Pritham, *DNA Transposons and the Evolution of Eukaryotic Genomes.* Annual Review of Genetics, 2007. **41**(1): p. 331-368.
215. Lu, C., et al., *Miniature Inverted–Repeat Transposable Elements (MITEs) Have Been Accumulated through Amplification Bursts and Play Important Roles in Gene Expression and Species Diversity in Oryza sativa.* Molecular Biology and Evolution, 2012. **29**(3): p. 1005-1017.
216. Oki, N., et al., *A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp. japonica.* Genes & Genetic Systems, 2008. **83**(4): p. 321-329.
217. Lisch, D., *Mutator transposons.* Trends in Plant Science, 2002. **7**(11): p. 498-504.
218. Cresse, A.D., et al., *Mu1-Related Transposable Elements of Maize Preferentially Insert into Low Copy Number DNA.* Genetics, 1995. **140**(1): p. 315-324.

219. Raizada, M.N., G.-L. Nan, and V. Walbot, *Somatic and Germinal Mobility of the RescueMu Transposon in Transgenic Maize.* The Plant Cell, 2001. **13**(7): p. 1587-1608.
220. Kapitonov, V.V. and J. Jurka, *Rolling-circle transposons in eukaryotes.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(15): p. 8714-8719.
221. Kapitonov, V.V. and J. Jurka, *Helitrons on a roll: eukaryotic rolling-circle transposons.* Trends in Genetics, 2007. **23**(10): p. 521-529.
222. Lu, S., et al., *CDD/SPARCLE: the conserved domain database in 2020.* (1362-4962 (Electronic)).
223. Stefan Kurtz, A.P., Arthur L Delcher, Michael Smoot, and C.A.a.S.L.S. Martin Shumway, *Versatile and open software for comparing large genomes.* Genome Biol, 2004. **5**(R12): p. 1-9.
224. Bennetzen, H.W.a.J.L., *Centromere retention and loss during the descent of maize from a tetraploid ancestor.* PNAS, 2012. **109**(51): p. 21004–21009.
225. Woodhouse, M.R., et al., *qTeller: a tool for comparative multi-genomic gene expression analysis.* Bioinformatics, 2021.
226. Arnaud Ronceret, M.-P.D., Inna N. Golubovskaya, and Wojciech P. Pawlowski,, *PHS1 regulates meiotic recombination and homologous chromosome pairing by controlling the transport of RAD50 to the nucleus.* PNAS, 2009. **106**(47): p. 20121-20126.
227. CLEGG, B.S.G.A.M.T., *Molecular evolution of alcohol dehydrogenase 1 in members of the grass family.* PNAS, 1991. **88**: p. 2060-2064.
228. Roessler, K., et al., *The genome-wide dynamics of purging during selfing in maize.* Nature Plants, 2019. **5**(9): p. 980-990.
229. Schnable, J.C., N.M. Springer, and M. Freeling, *Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(10): p. 4069-4074.
230. Song, K., et al., *Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution.* Proceedings of the National Academy of Sciences of the United States of America, 1995. **92**(17): p. 7719-7723.
231. Kersting, A., E. Bornberg-Bauer, and J. Gu, *Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes.* Nucleic acids research, 2011. **39**: p. 6029-43.
232. Wang, X., et al., *Genome Alignment Spanning Major Poaceae Lineages Reveals Heterogeneous Evolutionary Rates and Alters Inferred Dates for Key Evolutionary Events.* Molecular Plant, 2015. **8**(6): p. 885-898.
233. Bennetzen, J.L., et al., *Reference genome sequence of the model plant Setaria.* Nature Biotechnology, 2012. **30**: p. 555.
234. Muyle, A., et al., *Gene capture by transposable elements leads to epigenetic conflict in maize.* Molecular Plant, 2021. **14**(2): p. 237-252.
235. Jiang, N., et al., *Pack-MULE transposable elements mediate gene evolution in plants.* Nature, 2004. **431**(7008): p. 569-573.
236. Xie, Q., et al., *LNK1 and LNK2 are transcriptional coactivators in the Arabidopsis circadian oscillator.* The Plant cell, 2014. **26**(7): p. 2843-2857.
237. Slatkin, M. and R.R. Hudson, *Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations.* Genetics, 1991. **129**(2): p. 555.