# Higher order repeat structures reflect diverging evolutionary paths in maize centromeres and knobs

Rebecca D. Piri[1], M. Cinta Romay[2], Edward S. Buckler[2,3,4], R. Kelly Dawe[1,5,6] *

[1] Institute of Bioinformatics, University of Georgia, Athens, GA 30602
[2] Institute for Genomic Diversity, Cornell University, Ithaca, NY USA 14853
[3] Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY USA 14853
[4] USDA-ARS; Ithaca, NY, USA 14853
[5] Department of Genetics, University of Georgia, Athens GA 30602, USA
[6] Department of Plant Biology, University of Georgia, Athens GA 30602, USA

*Corresponding author:
R. Kelly Dawe
Department of Genetics
B414A Davison Life Sciences
University of Georgia, Athens, GA 30602
kdawe@uga.edu

Keywords:

Satellite DNA, HOR, centromere, maize, knobs, meiotic drive

**Abstract**

Background:

Highly repetitive tandem repeat arrays, known as satellite DNAs, are frequently found in low recombination regions such as centromeres. Satellite arrays often contain complex internal structures known as higher order repeats (HORs) that may have functional significance. Maize is unusual in having satellites in two different genomic contexts: centromeres, which interact with kinetochore proteins, and knobs, which are subject to meiotic drive when abnormal chromosome 10 is present. Whether HOR patterns exist in maize centromeres or knobs and how the patterns might relate to function is unknown.

Results:

Here, we generated 13 repeat-sensitive genome assemblies of maize and its recent ancestor, teosinte. We developed a local, binned approach to categorize HORs. Our findings reveal that HORs are ubiquitous in maize, but are generally low-frequency with small patterns, rather than the large, continuous HOR blocks found in humans and Arabidopsis. While centromeric CentC arrays contain majority HOR content, some of which is conserved in teosinte, the patterns are primarily locally-confined and unrelated to the active centromeres, as marked by Centromeric Histone H3. Knobs, on the other hand, have a more active HOR landscape. Large knobs contain megabase-scale repeat units, or similarity blocks, with conserved HORs. The large-scale repeat units may facilitate unequal crossing over events that enable rapid expansion, and possibly contain functional motifs that are recognized by trans-acting factors that mediate meiotic drive.

Conclusions:

HORs are present in all satellites in maize. HOR content is not associated with centromere function, but knobs contain conserved HOR patterns within similarity blocks that may facilitate meiotic drive.

## Background

Long tandem repeat arrays, called satellites, frequently occur in centromeric and subtelomeric regions of eukaryotic chromosomes. The weakened selection in these low recombination regions provide a haven for repetitive DNA, allowing it to replicate and spread with lowered risk of being purged [1]. Despite the prevalence of satellite DNAs, their origin, sequence arrangements, mechanisms of accumulation, and potential functional significance are often obscure. Recent advances in sequencing technologies have begun to reveal the internal structures of some of the longest repeat arrays in both animals and plants. In human centromeres, there is a core of alpha satellites with highly-homogenized HOR patterns that are associated with active kinetochores marked by the histone H3 variant CENP-A/CENH3 [2,3]. The HORs in core regions can differ dramatically among individuals, both in copy number and in distinct monomer patterns [4]. Outside the central core are divergent layers, representing obsolete centromere cores, that have accumulated variants and transposable elements over time. This dynamic, where highly-homogenized HOR patterns are actively evolving in close contact with the kinetochore, has been described as kinetochore selection on alpha satellite HORs [5]. A similar connection has been observed in Arabidopsis, although the HOR pattern layers and monomer variants are not as well-defined [6]. Arabidopsis CENH3 frequently associates with satellite CEN180 and the highest-frequency HOR patterns occur in the center of the arrays where CENH3 is most abundant.

Maize also has a centromere-associated satellite called CentC [7]. However, the association between CentC and CENH3 is polymorphic – functional centromere regions can be completely decoupled from CentC arrays in some chromosomes [8–10]. Additionally, maize harbors another class of widespread satellites known as knob repeats. Knobs are large, heterochromatic satellite arrays on chromosome arms, comprising two repeats, knob180 and TR1 [11,12]. In the presence of a chromosomal variant known as abnormal chromosome 10 (Ab10), they are subject to meiotic drive. Ab10 encodes two specialized kinesin proteins (KINDR and TRKIN) and several large knobs [13,14]. During meiosis, the kinesins associate with arrays of knob repeats (KINDR with knob180 repeats, and TRKIN with TR1 repeats) and move to the spindle poles more quickly than centromeres, allowing knobs to dictate chromosome movement. As a result, when Ab10 is heterozygous, the haplotype is preferentially passed on through the female germline (up to 83%) [15,16]. Importantly, knobs on other chromosome arms can take advantage of the trans-acting kinesin proteins and show the same levels of meiotic drive when they are heterozygous [16,17]. The signature of this drive system is

present in all maize genomes and their ancestors– knob repeat arrays dispersed across all chromosome arms, although only the largest arrays in mid-arm positions are known to show strong drive [16]. Knobs are defined by sequence, which is inert and densely repetitive, similar in structure to the satellite DNA associated with centromeres.

For the study of repeat arrays, maize presents an interesting test case where there are both centromeric satellites and meiotically-driven knob satellites. These repeat arrays can contribute as much as ~500 Mb to the maize genome, though the amounts vary from line to line [13]. Recent genome assemblies have included long spans of the major maize satellites [10,18], yet there has been no comprehensive analysis of their internal makeup, or their conservation among lines. Based on the emerging data from human and Arabidopsis, we anticipated that both centromeric and knob arrays would contain HORs, and that the most homogenous arrays would occur in regions of functional relevance. Yet, given that many maize centromeres have shifted away from CentC altogether [8–10], we anticipated that the HOR patterns associated with CentC may be functionally irrelevant, and therefore might be sparse or degraded. In contrast, given that knobs are in areas of high recombination and under active selection for meiotic drive, we anticipated that the HOR patterns in knobs would be pronounced, especially in larger knobs that are more strongly driven.

To evaluate the presence of HORs, we used PacBio HiFi sequence data to assemble 13 genomes – 10 of maize and 3 of its recent, wild ancestor, teosinte. After discovering that software used to identify and describe HORs in human and Arabidopsis did not perform well in maize, we developed new methods, which use similar monomer identification and clustering methods to previous tools, such as AlphaCENTAURI and HORmon, but allow greater flexibility in pattern identification [19,20]. We found that locally-confined, small HOR patterns are widespread in maize satellites, but that they tend to be low-frequency and irregular. In centromeres, the location, density, or conservation of these patterns are not correlated to the current active centromere position. Rather, these relatively small and infrequent CentC patterns may be a signature of recurrent breakage and repair that is innate to satellite DNA.

In knobs, we found high-frequency HOR patterns that are punctuated regularly along satellite arrays, about 1 Mb apart, intermingling with the local HOR patterns. The long distance between repeating HOR patterns indicates they are likely driven by an alternative mechanism than local HOR's – unequal crossing over. Further, we found that these patterns are shared among three of the largest knob180 arrays, indicating there is a mechanism for sharing sequences among these distinct satellite loci, possibly gene conversion or rolling-circle replication, consistent with classic models of centromere evolution [21,22]. Due to their

consistency within and among knobs, we postulate that the conserved HOR patterns may represent functional units that are under selection for meiotic drive.

## Results

### *Repeat-Sensitive HiFi Assemblies in Maize*

We initially assessed older genome assemblies from the maize pangenome, generated with PacBio CLR (long single reads), for their utility in interpreting satellites [10]. But we discovered poor agreement between satellite content in input reads and assemblies, indicating assemblies did not accurately represent the repeats [10,23] (Table S1-2). The older assemblies also had poor read alignment with new HiFi reads, possibly due to overpolishing with Illumina reads, relatively error-rich reads used to generate the assemblies, or misassembly (Figure S1-2). So, we opted to use new assemblies based on HiFi reads (circular consensus reads), which are more accurate than the sequencing used in the older assemblies.

For this study, we generated assemblies for 13 inbred lines – 3 from inbred lines previously used in genetic studies (B73, B73-Ab10, and Mo17) [10,18,24], 7 diverse inbreds from public breeding programs, and 3 teosinte lines from the PanAnd Project [25], representing ancestral variation. The final assemblies ranged from 2.186 Gb for B73 to 2.766 Gb for the teosinte inbred TIL01. Compared to repeat content in the raw sequence data, these HiFi-based assemblies appear to have a more faithful representation of satellite content (Table S1-2). The total satellite repeat content varied from 6.64 Mb in B73 to 300.48 Mb in TIL01 (Figure 1a).

There were gaps in all assemblies. Total gaps ranged from 46 in Mo17 (0 in satellite arrays) to 288 in Tx779 (11 in satellite arrays) (Table S3). For in-depth assessments of centromeric repeats, Mo17 was used due to its gapless arrays and availability of ChIP-seq data [18]. For knob repeats, CG108 was used, due to its relatively few N-gaps genome-wide (Table S3) and its gapless, large knob repeat arrays on chromosomes 7 and 8 (Figure 1d).

### *Repeat Array Positions are Highly Conserved*

Maize has four distinct classes of satellites related to centromeres: two associated with canonical centromeres, CentC (156bp), it's primary centromere-associated repeat, and Cent4 (741 bp), a pericentromeric repeat linked to centromere 4; and two associated with knobs, knob180 (180bp) and TR1 (358bp) [7,9,11,26]. All four classes of satellite are organized into tandem repeat arrays, where multiple copies occur together in head-to-tail orientation. Previous

pangenomic studies demonstrated that the major satellite arrays positions are well conserved within maize [10].

Within the 13 genomes used for this study, 156 distinct satellite array positions– 151 on normal chromosomes, and 5 specific to the Ab10 haplotype– were identified (Figure 1d). Of these positions, 78 are private, meaning they only occur in one line, and 30 are shared among all genomes (Figure 1c ). The high proportion of positional conservation is true even for small arrays that are not expected to be functional. The smallest array in a fully conserved position is only ~1.5 kb, containing only 9 CentC monomers. Within each genome, there are a total of 71 to 113 satellite arrays, with knob180 consistently having the most arrays and Cent4 consistently having the least arrays (Figure 1b).

Although there is some variation in the number of arrays per genome, variation in array number is not sufficient to account for the overall variation in satellite content. Rather, a handful of very large arrays drive overall satellite content variation (Figure 1b-d). The variation in length at one site can be substantial. At one CentC array position on chr10, the largest centromeric array assembled without gaps is 5.7 Mb in Mo17 and the smallest is 7.7 kb in B73 – a length difference of ~5 Mb, and a relative difference of >700x. In the largest knobs, size can vary >46 Mb at a single locus– for example, one of the largest gapless arrays on chromosome 7 is 51 Mb in CG119 but its homolog in CG44 is only ~9% of its size at 4.8 Mb.

*Local HOR Classification*

Methods for HOR identification can be categorized into two groups: first, algorithms that use high-copy, periodic k-mers to identify novel repeats and HORs de novo (i.e. TRASH, SRF toolkit); and second, methods that utilize previously-characterized satellite sequences to identify patterns of monomer subtypes (i.e. AlphaCENTAURI, HORmon, HiCAT) [19,20,23,27–29]. Current software is largely based on highly regular HOR structures, such as those observed in human and Arabidopsis functional centromere cores. They lack power to identify HOR patterns that are low-frequency and irregular, like those in maize. SRF toolkit is the only current algorithm that was benchmarked with maize, and it only identified two major satellite repeats in B73 – a CentC variant and a 4-mer knob180 unit, neither of which formed repeat arrays [23].

To improve HOR identification in maize, we adapted methods from algorithms that identify monomer subtype patterns. But, to better identify small, locally-confined patterns, only single 10 kb regions were analyzed at a time. Additionally, expectations of periodicity of the pattern, meaning that a pattern exists in multiple units tandemly with even spacing among all occurrences, were removed. Satellite arrays were separated into non-overlapping 10 kb bins

(Figure 2). Within a single bin, full-length monomers were identified with HMMER nhmmer and compared all to all with BLAT [28,30]. Similarity between monomer sequences was calculated as a Jaccard Index (# identical bp/(total length of both sequences - # identical bp)), which represents both sequence and length similarity. Pairwise similarity scores were used to generate networks, with nodes representing distinct monomer sequences and edges present between two nodes representing similarity. Networks were regenerated for all Jaccard similarity scores between .90 to .99 in .01 increments, where the nodes remained the same, but similarity cutoffs to determine edges became progressively more stringent, similar to the processes of alphaCENTAURI, HORmon, and HiCAT.

The repeat structure for the bin was then predicted for all thresholds with an LDA model, using summary information describing the network structure. Important network summary statistics included the following: proportion of monomers in the largest cluster (# monomers in largest cluster / # of total monomers), proportion of monomers in the second largest cluster (# monomers in second largest cluster / # of total monomers), number of unconnected clusters relative to number of monomers (number of monomers that do not share similarity with any other monomers / # of total monomers), average pairwise Jaccard Index, and proportion of monomers collapsed into the most prevalent subtype (or most common distinct sequence) (# monomers identical to most common sequence / # of total monomers).

The LDA model classified the repeat structures in each bin at every threshold in one of three categories: HOR, where most monomers belong to two or more similarly-sized clusters in the network; Order, where most monomers are connected together in a single cluster; and Disorder, where most monomers do not belong to clusters. For each bin, the similarity threshold classified with the highest posterior probability was used for further analysis, deemed the “optimal clustering threshold”, and adjacent bins with consistent classifications and thresholds were merged. See methods for more information on the LDA model and training data.

For bins with predicted HOR structures, monomers were labeled by their cluster identity. For example, all the monomers in the largest cluster were labeled as “A”, monomers from the second largest cluster were labeled as “B”, and so on using both upper and lower case characters (A-Y, a-y) and numbers 0-9, if needed. Monomers that were unconnected were labeled at “Z”, representing a private cluster. Monomer patterns could then be represented as character strings, with each monomer represented by its corresponding cluster character. The pattern string was then decomposed into k-mers of various lengths to identify repeating patterns of >=3 monomer subtypes.

The purity of an HOR region was calculated as the number of monomers that are in recognizable HOR patterns, divided by the total number of monomers. Over 98% of predicted HOR bins were confirmed to have at least 10% content of identifiable HOR patterns of at least 3 monomers, occurring >=2 times within the bin. By these criteria, a majority of all satellites were classified as HOR, ranging from a minimum of 48% of bins to a maximum of 97% (Figure 3a). Proportionately, assembled centromere-related satellites contain slightly more HOR content compared to knobs. Of these patterns, most are relatively small and low frequency (Figure 3b-c).

*Homologous Arrays Share Sequence Similarity*

Homologous arrays, which are conserved in positions relative to core genes, are also generally conserved in sequence. Comparing the whole satellite array sequence, excluding any non-satellite DNA or TEs, homologous arrays have an average Jaccard similarity of .82 (Figure 3d, Table S4). This similarity score captures both sequence and length (copy number) divergence and varies slightly among satellites– ranging from average similarity of .96 for Cent4 positions to .74 for TR1 positions. When comparing maize lines to other maize lines, the arrays tend to be more similar, consistent with the fact they are from one subspecies (*Zea mays* ssp. *mays*) with limited genetic diversity due to the domestication bottleneck [31–34]. When considering teosinte compared to teosinte, the results are more of a mixed bag– some satellites are more similar and some are less. This is perhaps not surprising as two of the teosinte lines are from *Zea mays* ssp. *parviglumis* (TIL01 and TIL11) and the third (TIL25) is from *Zea mays* ssp. *mexicana*.

Within a homologous group of arrays, smaller arrays are generally more similar to each other than larger arrays (Figure 3e). Smaller arrays also have less relative HOR content (Figure 3d-e). In fact, the smallest arrays, where the largest homolog is <=10kb, are completely devoid of HORs. In larger arrays, repeat content is significantly correlated with greater HOR content (.32 correlation with a p-value of .00012), meaning larger repeat arrays contain relatively higher proportions of HOR content (Figure 3d). In turn, this greater HOR content is significantly correlated with lower average similarity among homologous arrays (-.43 correlation with a p-value < .0001), meaning greater HOR content is related to greater divergence (Figure 3e). This trend likely reflects a tendency for longer arrays to undergo rapid expansion and contraction events (Jaccard similarity is normalized by the length of both arrays).

*Shared HOR Classification*

Patterns were then compared among 10 kb bins to capture HORs beyond local regions, including non adjacent bins on the same array and those in homologous arrays in other inbreds (Figure 4, step 1). To do this, all monomers from HORs were converted to consensus sequences (i.e. pattern ABCABC was converted to consensus monomers A, B, and C). Then, consensus monomers were compared all-to-all. A network was created, where each node represents a single consensus monomer, and two nodes are connected by an edge if their sequences are at least as similar as their shared optimal clustering threshold (Figure 4, step 3). Meaning, if patterns ABC and DEF with optimal clustering threshold of .95 are being compared, A, B, and C are expected to cluster separately at the .95 cut off. However, if these patterns share recent evolutionary history, A may cluster with D, B with E, and C with F. This step is necessary since local HORs were initially labeled based only on patterns within their local 10 kb bin. By reclustering, we could relabel HORs based on larger-scale comparisons and generate a key to translate among regions.

Shared HORs indicate maintained or recently duplicated patterns. Of all HOR regions in Mo17, our model assembly for centromeres, only 14% have shared HOR patterns with a homologous array in at least one of the other 12 genomes (Table S5). HOR regions with at least one shared pattern include 13% of CentC, 88% of Cent4, 8% of knob180, and 62% of TR1 HORs. A similar pattern is true in CG108, our model assembly for knobs. In CG108, 19% of HOR patterns are shared with another genome, including 13% of CentC, 88% of Cent4, 13% of knob180, and 59% of TR1 HORs. The data suggest that while CentC and knob180 are more prevalent satellites and are related to more active centromeric and neocentromeric function, their HOR patterns are less conserved in evolutionary time.

*Centromeric Satellite HORs are Not Related to Function in Mo17*

Of the 13 maize lines used here for genome analysis, CENH3 ChIP-seq data are available for only one, Mo17 [18]. CENH3 centers around the major CentC array on a subset of the chromosomes (1, 4, 7, and 9), is partially associated with CentC on other chromosomes (2, 3, 6 and 10), and is completely decoupled from CentC on others (8 and 5) (Figure 5, S5-6) [18]. We observed no obvious correlation between HORs and CENH3 localization. CentC arrays contain HORs that are shared in evolutionary time with homologs regardless of whether they are associated with CENH3. Six centromeres also contain HOR patterns that occur multiple times within a single array. However these patterns only occur in 2-4 regions that are generally close in proximity, and are also not correlated with the active centromere location.

Additionally, the HOR patterns present are not more frequent or of higher purity towards the center of the array. While HOR patterns are not present at the very edges of arrays, there is no obvious HOR core, unlike human and Arabidopsis centromeric repeat arrays.

*Knobs Contain Conserved Patterns*

Large knobs are characterized by megabase-scale periodicities. These large-scale patterns, roughly .48-1.5 Mb, are defined by regions of high similarity to the knob180 consensus followed by regions with low similarity to the consensus, visible as vertical stripes of low-similarity monomers (Figure 6, 7). Within a single unit, the regions of low homology to the knob consensus tend to have degraded and truncated knob repeats, while the regions of high consensus have more intact arrays and HORs (Figure 6a). When 1 Mb units of these patterns are aligned to each other in a traditional all-by-all dot plot, there is strong homology along the diagonal, suggesting that they are related by descent (Figure 7). We refer to the megabase-scale patterns within knobs as similarity blocks.

The similarity blocks are also typified by shared HOR patterns. In the fully assembled knob on the long arm of chromosome 7 in CG108, there are 417 distinct HOR patterns that occur in more than one similarity block, six of which occur in at least 30 blocks across the array (Figure 6c). These shared patterns exhibit a skipping pattern, where the conserved patterns are spread across the array in regular intervals, and found in most but not all similarity blocks. The same shared HOR patterns are found on the other two large knobs in CG108, which are on chromosomes 6 (6Mb) and 8 (28Mb). There are 677 HOR patterns that are shared among at least two of the knobs. The 10 highest-frequency, shared HOR patterns are composed of 7 high-frequency monomers, which are named here as knob180 A-F (Figure 7b, Table 3, S4).

## Discussion

In this study, we developed a novel HOR identification pipeline for maize that revealed a majority of satellite content, both in centromeres and knobs, are composed of higher-order repeat patterns. Unlike HOR patterns previously described in humans and Arabidopsis [3,6], maize HORs are primarily locally-confined, meaning most occurrences of a single HOR pattern are contained within a region of ~10 kb. These HOR patterns in maize are also fairly small and low-frequency– with average pattern sizes of ~4 monomers with ~3 occurrences.

While pervasive in the satellite space, the presence of HOR patterns does not seem to be related to active centromere activity in maize. In human centromeres, there are older layers

of HOR patterns that are present on either side of the centromere core. Drawing a line to connect regions with shared HOR patterns, human centromeres look like a layered onion [3]. Such layers are absent in maize centromeres. While shared patterns among regions in the same array exist in some centromeres, they are still locally constrained and do not exhibit any obvious patterns (Figure 5). Meaning, two bins with shared patterns are near each other, rather than existing on opposite sides of the array in the onion-like layers present in human centromeres. These results suggest that maize centromeric satellites undergo different selection than human centromeric satellites, which have evidence of kinetochore selection on pure HOR patterns [5].

The presiding evolutionary model for the origin and maintenance of HOR patterns is breakage induced replication (BIR), which proposes that repeats take advantage of the inherent instability and breakage at centromeres, co-opting repair mechanisms to expand existing patterns, generate new repeat variants, and exchange repeats among different arrays or chromosomes [35–37]. Under this model, HOR pattern expansions are frequent but relatively small, as strand repair is expected to be primarily over local distance (i.e. on the scale of a single HOR pattern length, upwards of a few dozen monomers). Experimentally, this model is supported by HOR copy number changes over 20 somatic cell divisions in a sensitized human cell line, but the frequency of the events within an organismal germ line is unknown [38]. This model unifies evidence of replication fork stalling and collapse at centromeres (reviewed in [35]) and high levels of centromeric rearrangements and instability [39–41] with observed repeat structures, as well as proposes a feasible colonization mechanism across different arrays via BIR. The local HOR patterns in maize centromeres, which are only present in longer arrays, seem to suit expectations that late-replicating, highly repetitive regions of the genome are enriched for DSBs, which can be repaired out-of-register and result in short-range tandem duplications [39,42]. However, there is no evidence that the kinetochore applies selection for enriched BIR in centromeric cores.

On the other hand, knobs do have signatures of selection. We have described large-scale patterns of similarity in some of the largest knobs that contain conserved HOR patterns (Figure 6,7). The shared HOR patterns among knob regions are dispersed at semi- regular intervals, punctuated with local patterns in between – like skips across the array (Figure 6). Mechanistically, the scale of these long-range skips seems unlikely to be driven by BIR due to the large distance between homologous patterns. Rather, because knobs are located on chromosome arms where recombination is high [43], unequal crossover may be a more apt explanation [44]. The unequal crossover model is based on classic literature showing that two

tandemly duplicated genes can either contract to one copy or expand to three copies as a direct result of crossing over [45,46]. Although there are no known genes in knobs, there is evidence that repetitive knob DNA pairs during meiosis and undergoes crossing over. Stack and colleagues demonstrated that foci of staining for MLH1, a key protein required for meiotic crossover [47], regularly occur within paired knobs at the pachytene substage of meiosis [48]. During pairing in repeat-dense regions, alignment can easily slip out-of-register and result in unequal crossing over. In the largest knobs, misalignment of similarity blocks could occur on a megabase scale, resulting in major shifts in total knob size. Therefore, within knobs unequal crossing over may drive large expansion events, while BIR drives an abundance of smaller-scale, local HOR patterns. Drive may positively select on both expansion types, as both mechanisms can increase the array sizes (although on different scales), as well as select for sequences specific to the high-frequency monomers.

In lines carrying abnormal chromosome 10, KINDR associates with knob180 knobs to initiate neocentromeric activity during cell division [13]. Larger knobs are known to drive better than smaller knobs, possibly indicating better binding affinity to the protein [49]. While the binding specificity of KINDR to knob repeats is well documented, how that interaction is achieved is unknown. An unknown linker protein is hypothesized to interact directly with the DNA sequence and recruit KINDR [16]. The distribution of the conserved knob repeats and HORs containing those repeats leads us to believe that they may function as the binding sites for the postulated linker protein (Figure 6, Table S5). Ongoing selection on these sequences for meiotic drive would also explain how they maintain their identity and structure over such long distances in multiple loci.

## Conclusions

We have demonstrated that the maize satellite landscape is replete with low-frequency, low-periodicity HORs that are not detectable using approaches developed for other model organisms (e.g [50–53]). These HOR patterns are consistent with random breakage and repair by BIR. Unlike in humans and Arabidopsis [3,6], there is no apparent enrichment of HORs in regions occupied by kinetochores, suggesting little if any functional relationship between HORs and centromere function. In contrast, knobs have repeat structures suggestive of functions related to meiotic drive. There are large, megabase-scale similarity blocks that may facilitate rapid expansion and contraction of knob size by unequal crossing over. Within the similarity blocks are conserved HOR patterns that may serve as binding sites for meiotic drive proteins.

## Methods

*Data Sources*

Raw reads from the maize NAM pangenome resource [10], including Illumina and PacBio CLR data, were retrieved from PRJEB31061 and PRJEB32225. Maize pangenome assemblies were downloaded from MaizeGDB.org. Raw reads, including PacBio CLR, Illumina, and ONT, and the previous assembly of B73-Ab10 [24] were collected from PRJEB35367. Raw PacBio HiFi reads for B73-Ab10 were collected from NCBI BioProject PRJEB35367. Raw PacBio HiFi reads from Mo17 were collected from PRJNA751841 [18]. Raw PacBio HiFi reads for B73 were collected from SRR11606869 [54]. Raw PacBio HiFi reads for TIL01, TIL11, and TIL25 were collected from PanAnd project data, Bioproject PRJEB50280.

For all other assemblies, DNA extraction and genome sequencing was performed at the Arizona Genomics Institute (The University of Arizona). Genomic DNA was extracted with a modified CTAB method [55]. High molecular-weight DNA was quality checked with Qubit HS (Invitrogen) and Femto Pulse Systems (Agilent) and 10 µg DNA were sheared to appropriate size range (15-20 kb) using Megaruptor 3 (Diagenode). PacBio HiFi sequencing libraries were constructed using SMRTbell Express Template Prep Kit 2.0. The library was size-selected on a Pippin HT (Sage Science) using the S1 marker with a cutoff at 15 kb. The sequencing libraries were sequenced on a PacBio Sequel IIe instrument with PacBio Sequel II Sequencing kit 2.0.

Raw PacBio HiFi reads for CG44, CG119, CG108, Tx777, and Tx779 are available under NCBI Bioproject PRJEB59044, and were sequenced as part of the Genomes to Fields Project. Raw PacBio HiFi reads for K64 and CML442 are available under NCBI Bioproject PRJEB66502 and were sequenced as part of an effort to sample diverse maize haplotypes containing sources of unique alleles available at the USDA-ARS maize Germplasm Resources Information Network (GRIN).

*Repeat Consensus Sequences*

To capture repeat content and variation, de novo repeat identification was performed with RepeatExplorer2 (v3.6.4) (-p -put ILLUMINA -c 20 --max_memory 500000000 -tax VIRIDIPLANTAE3.0) using PE150 Illumina data for NC350 from the maize pangenome [10,56]. RepeatExplorer uses TAREAN under-the-hood to reconstruct repetitive DNA through de Bruijn graphs for high-copy k-mers. The output provides both a single primary consensus, based on the best-supported graph path through the most frequent k-mers, and a list of common variants

in the data, based on the other common k-mers. The primary consensus sequence and consensus variants for the satellites of interest (CentC, Cent4, knob180, and TR1) were identified in the RepeatExplorer output by comparing sequences to previously described consensus monomers using Blast+ [9,57].

Using only a single primary consensus sequence can be an issue in tandem repeat analysis, as blast prioritizes maximum similarity to a single sequence at a time, resulting in many overlaps in the output data and partial copies [58]. In decomposing a repetitive region into individual monomers, finding distinct monomers is key for downstream structural analysis [28]. As an alternative to using only a single consensus and blast for finding full-length monomers, nhmmer was used. Consensus variants identified by RepeatExplorer were aligned with a multiple sequence alignment in MUSCLE (v3.8.1551) and used to generate phmm's (profile hidden markov models) with the makehmmerdb command in HMMER (v3.3.2) using default parameters [59]. Rather than reporting all hits, including overlapping and partial hits like BLAST, HMMER utilizes a collection of sequences within one phmm to find the best match for each region. Only the best, most complete similarity hit for each monomer is reported when repeat monomers are identified using the nhmmer command.

*Repeat Content*

To calculate estimated total repeat content, satellite consensus sequences were compared to raw reads of multiple classes and genome assemblies using Blast+. For TE enrichment analysis in raw reads, the analysis was repeated using Shojun Ou's TE library [60]. Repeat hits were converted to a bed file and then merged using bedtools (v2.3) [61,62]. Total length of reads and assemblies was counted with Bioawk (v1.0) [63]. Total genomic proportion of repeats was then calculated by dividing total repeat length, summed from merged blast hits, by total read (or genome) length. Alternatively, estimated genomic content was determined by dividing the total repeat length by the estimated read depth. Estimated repeat depth for HiFi reads was reported by hifiasm [64]. For other read types, estimated read depth was gathered from their source papers [10,24].

*Read Depth Over Repeats in Old Assemblies*

Raw PacBio CLR reads from the NAM pangenome project [10] were aligned to the B73-Ab10 assembly and two maize pangenome assemblies (B73, NC350) to check read depth over centromeric regions [10,24]. Reads were aligned to both full assemblies (with unscaffolded contigs included) and only the pseudomolecules (representing only assembled chromosomes).

For the B73-Ab10, ONT and HiFi reads were also used. In each case, reads were aligned with minimap2 (v.2.24) with appropriate default settings for the data type [65,66]. Read alignments were filtered using the 2308 filter with samtools (v1.6) [67] to remove multi-mapping. Alignment bam file was then converted to a bed file using bedtools (v2.3) bamtobed with default parameters, with a fourth column of read length added. Average read depth over 10 kb was calculated with bedtools map (v2.3) (-c 4 -o sum) for each alignment [61,62] (Quinlan and Hall 2010). For each bin, total read length was summed and divided by 10 kb. Bin content of satellites was also identified, using bedtools intersect to combine satellite blast hits with 10 kb bins. Satellite hits for each bin were summed and divided by 10 kb to identify repeat proportion.

*HiFi Assembly*

HiFi contigs were generated using hifiiasm (v 0.19.6) with homozygous settings with end-joining disabled (--write-ec --write-paf -u0 -l0) [64]. Contigs were checked for repeat content using Blast+ (v2.10.1) (blastn -outfmt 6 -num_threads 10 -max_target_seqs 5000000) and anchored contigs were assessed [57]. In this study, an anchored contig is defined as a contig that starts and/or ends with non-satellite DNA. Contigs with at least one anchored end can be confidently placed during scaffolding. Anchoring was assessed by checking for repeat hits within 100 bp of either end of the contig using bedtools intersect.

The resulting contigs were variable in quality– for many genomes, the contig-level assembly was enriched for short, low support contigs that were dense with satellite content (Figure S3b). The presence of these contigs was unrelated to read depth, length, or quality. These low-confidence contigs inflated the total relative repeat content, with contig repeat content outpacing genome estimates from raw reads. To enrich for accurate array assemblies, low-confidence contigs were removed, leaving only contigs with greater than half the expected read support (i.e. for an inbred with HiFi read depth of 20, only contigs with >10 read support were used). The high-confidence contigs are longer, have repeat content consistent with raw reads, and are largely anchored, meaning they have at least one edge that is not satellite DNA and can be confidently scaffolded into a final assembly (Figure S3b-d). High confidence contigs were scaffolded using the Mo17 assembly using RagTag (v2.0.1) [18,68].

In the initial assembly for CG119 scaffolded with Mo17, chromosome 3 had two large CentC arrays, and chromosome 7 had none. The CG119 assembly scaffolded with B73 v5 did not have this same anomaly. To manually correct this issue, the misplaced contig was manually added to a bed file in the correct spot for chromosome 7 and removed from chromosome 3 in another bed file. 100N gaps were placed between contigs.

*Annotation of Core Genes*

Scaffolded assemblies were annotated with Liftoff (v1.6.3), using gene annotations from Mo17 as the reference [69]. Annotated genes were subset down to a highly conserved set, defined as genes that are single-copy in all new assemblies, mapping to the same chromosome, and were core genes in the maize pangenome [70]. To identify core genes, gene coordinates were collected from the annotation GFF file and converted to a bed file. Then, gene annotations for B73 v5 were similarity converted, subset to the list of known core genes, and extracted as a fasta file using bedtools getfasta (v.2.29.2). The B73-derived core genes were then aligned to Mo17 with minimap2 (v2.22), filtered to their best hit using samtools view (-F 2308) (v.0.1.2), and converted to a bed file using bedtools bamtobed [65–67]. B73 core gene hits were then intersected with Mo17 annotations to find equivalent genes with bedtools intersect.

*Repeat Array Positions*

Repeat monomers previously identified via blast+ were filtered to a minimum size of 30bp and merged within 10kb with bedtools to define arrays. To define array positions, array coordinates were compared to the core genes using bedtools to identify the nearest upstream gene (-iu -D a -a) and downstream gene (-id -D a -a), which were used as the array coordinates for comparison. Array positions were clustered among lines using the graph_from_edgelist function in igraph (v1.2.6), with each node representing an array, and edges among arrays indicating that the two arrays share either or both core genes on either side [71], [72]. Repeat arrays in the same position relative to conserved genes were referred to as homologous arrays.

*LDA Model Training*

For model building, test data from centromeres 2, 7, and 10 from the maize NC350 reference genome were used [10]. 229 non-overlapping 20 kb bins were manually labeled as HOR, order, or disorder based on repeat similarity dot plots and network topology. The data was split into training and test sets (n = 183 and 46, respectively). Extracted characteristics were then used as predictor variables. Utilizing graph structure (rather than periodicity of monomer subtypes determined by cluster identity) to identify optimal clustering thresholds and predict HOR structure is novel to this study.

The first model was an LDA (Linear Discriminant Analysis) model built with MASS (v7.3) [73]. The model uses 3 linear discriminant functions (LD1, LD2, LD3), the first two of which

explain 91.33% of the data variation (76.48% and 14.85%, respectively). LD1 utilizes the proportion of monomers collapsed into the most prevalent subtype, proportion of monomers in the second largest cluster, number of unconnected clusters, and proportion of monomers in the largest cluster. LD2 utilizes the proportion of monomers collapsed into the most prevalent subtype, proportion of monomers in the largest cluster, number of unconnected clusters, and the average pairwise Jaccard Index. The LDA model was 89% accurate with the test data.

A second model was also built to compare performance. The second model was a decision tree, built with rpart (v4.1) [74,75]. The decision tree utilized the number of unconnected clusters, modularity, and proportion of monomers in the largest cluster to categorize the bins. This model was slightly less accurate (87%) with the test data.

The LDA model was selected for use due to its better performance in test data. For each bin, erroneous classifications were removed (i.e., if all but one threshold level predicted HOR, the one threshold prediction was removed). Then, the classification prediction for each bin with the highest posterior probability was selected, and tandem bins with consistent classifications were merged.

*HOR Pattern Validation and Purity*

To validate identified HOR patterns, monomer patterns were converted to character strings. First, bins classified as HOR were extracted and monomers were re-clustered based on their optimal clustering thresholds, but without collapsing identical monomers. Monomers were then labeled by their cluster identity. For example, all monomers in the largest cluster were given the label “A”, all the monomers in the second largest cluster were labeled “B”. Characters A-Y, 0-9, and a-z were used as labels. “Z” was used to identify all monomers that exist in a private cluster, meaning the monomer was clustered by itself with no similar sequences.

The character strings were then decomposed into k-mers– starting with k=3 up until k where all k-mers occurred only once. The k-mer list was then filtered based on the initial criteria: all k-mers containing “Z” were removed, k-mers that are majority one letter (like CCCCC or CCAC), smaller k-mers that are fully contained in a larger k-mer that occurs equally often (AB with frequency of 4 removed in favor of ABAB with frequency of 2), and larger k-mers with subsets that occur more often (ABCD with frequency of 3 removed for ABC with frequency of 4 and BCD with frequency of 3). Final k-mer lists contained overlapping patterns to capture both largest HOR sizes and smaller variants and partial patterns. The strength of the relationships between total HOR content in the genome assemblies and HOR pattern lengths and frequency were calculated using the lm function in R [76] .

For each HOR, purity was then assessed. Purity was calculated as the number of monomers in a string in an identifiable HOR pattern, represented as a k-mer that passed filtering, divided by the total number of monomers in the string. For this value, monomers in overlapping patterns were only counted once. The small number of bins (~2%) that had no HORs (a purity of 0) were relabeled as order.

*Shared HOR Analysis*

To compare HOR patterns among non-adjacent bins or homologous repeat arrays, consensus representatives of HOR patterns were compared. First, consensus monomers for all monomer subtypes in identified HORs were generated. ClustalOmega (v.1.2.4) was used to make a multiple sequence alignment, and EMBOSS (v6.6) was used to make the consensus [39,77–80]. For example, if the pattern in bin 1 was ABCABCABC, consensus monomers for subtypes A, B, and C were generated.

Then, consensus sequences from bins of the same optimal clustering thresholds were compared using BLAT, as described previously. Bins compared include non-adjacent bins from the same array and bins from homologous arrays in other inbreds. For example, consensus subtypes A, B, and C from a 3-mer HOR in B73's centromere 5 may be compared to consensus subtypes J, K, L, and M from a 5-mer in a homologous array in Mo17, if they had the same optimal clustering threshold of .98. Comparing consensus sequences made it possible to "translate" patterns to identify conserved patterns. For example, in comparing ABC from B73 and JKLM from Mo17, we may find that A and K are at least .98 similar, B and L are at least .98 similar, and C and M are at least .98 similar. From there, we know that ABC and JKLM contain a shared 3-mer, which may indicate recent shared history between the sequences (Figure 4).

Similar to the HOR pattern validation process, similarity matrices of consensus monomers were converted to a network using graph_from_adjaceny_matrix in igraph (v1.2.6) [71]. Edges below the shared optimal clustering threshold were removed and monomers were labeled by cluster identity. Each group was assigned a character as described above. Here, a private monomer, labeled as "Z", indicates a monomer that does not share homology with another monomer equal to or greater than the similarity threshold. Monomers not included in the shared HOR analysis were also labeled as "Z". Then, character strings were decomposed into k- mers as described above, but they were not filtered, to allow for identification of shared partial patterns.

Shared HOR analysis was repeated among all knob180 arrays to assess repeating HOR

patterns. High-frequency patterns, shared among multiple non-homologous arrays, and consensus monomers within these patterns were generated.

*Shared HOR Purity*

Purity for all HOR bins in Mo17 and CG108 was recalculated using shared HOR patterns at multiple levels— first considering all shared patterns (shared across multiple bins in the same array and/or with at least one other inbred), shared only within maize (present in at least one other maize inbred), and shared with teosinte (present in at least one teosinte).

*Whole Array Comparisons*

For whole array pairwise comparisons, non-repetitive sequences within arrays were masked using bedtools (v2.29.2) maskfasta [61,62]. Then, homologous arrays were compared using blastn (v2.2.31) ( -db {homologs.db} -query {homologs.fasta} -outfmt "6 qseqid sseqid qlen slen length nident pident qstart qend sstart send") [57]. Pairwise blast hits were merged using bedtools merge, and total base pairs identical were summed. Then, pairwise similarity was calculated in both directions as a the total number of identical base pairs, divided by total length of repetitive sequence in the array.

*Monomer Similarity-to-consensus Visualization*

To visualize repeat structure among homologous arrays, monomers were compared to their consensus. Repeat arrays were extracted from their assemblies using bedtools getfasta. Then, monomers were identified using HMMER nhmmer. The output file (.out) was converted to a bed file. For hits on the opposite strand, start and end coordinates were flipped. Hits were then extracted from the array file using bedtools getfasta. Finally, monomers were compared to the appropriate primary consensus sequence using BLAT (v3.7) [30].

For this comparison, BLAT was utilized because the output format contains convenient match information for conversion into a Jaccard Index. For monomer comparisons, a Jaccard Index is ideal to measure similarity, penalizing for both length and sequence differences. If multiple scores were provided for a hit, only the highest-scoring similarity hit was used. Repeat array structure was represented by plotting each monomer as a dot with the X axis as the genomic coordinate and the Y axis as the Jaccard score to reference in ggplot2 [81].

*Defining Similarity Blocks Within Large Knobs*

In the similarity-to-consensus dot plots for some of the largest knobs, a repeating pattern was visible (Figure 6,7). In these arrays, there are regions of monomers with low similarity to consensus, visible as a vertical stripe in the dot plot, followed by regions of close similarity to consensus, visible as a horizontal line at the the top of the plot. Each unit of this pattern is roughly ~1 Mb on average. This same striped pattern was observed in the large knob of chromosome 7 in 8 inbreds where the assembled knob was at least 1 Mb (TIL25, K64, CML442, Tx779, Tx777, CG44, CG119 and CG108).

To directly compare the similarity blocks, four 1 Mb windows were selected, three in CG108 K7L and one in CG108 K8L. These windows represent sample similarity blocks. Within each block, monomers were extracted using bedtools (v2.3) getfasta and compared all-to-all with BLAT (v3.7) [28,30], [61,62]. Monomers with at least .98 pairwise Jaccard similarity were represented in a dot plot.

*CENH3 Enrichment*

CENH3 illumina CHIP-seq reads and their corresponding input runs from Mo17 were downloaded from NCBI BioProject PRJNA751841 [18]. These reads were generated in the same study as the Mo17 PacBio HiFi reads [18].

For repeat-sensitive mapping, the process from Logsdon et al [4] was used. Briefly, reads were trimmed and dedupped using fastp (--dedup --detect_adapter_for_pe --cut_front --cut_tail) (v.0.23.2) [4,82]. Prepped reads were then aligned to the generated Mo17 assembly using BWA-MEM (v.0.7.17) (-k 50 -c 1000000) [83]. Hits were filtered with samtools (v0.1.19) (view -b -S -F 2308 ) [67]. To find unique k-mers for sensitive mapping, unique 51-mers were found with meryl (v1.4.1) (meryl count k-51 | meryl equal-to 1 | meryl-lookup -bed-runs) [84]. Unique k-mers were then used to filter alignment bam files, where read alignments that fully overlapped with a unique k-mer were extracted using bedtools (v.2.29) intersect (-b1 {chip.bam} -b2 {unique.bed} -ubam -wa -F 1) [61,62,84]. Uniquely mapping hits were then normalized and 1 kb bins compared using deepTools bamcompare (-of bedgraph --operation ratio --binSize 1000 --scaleFactorsMethod None --normalizeUsing RPKM) [85]. For plotting, CenH3 enrichment was averaged over 100kb bins.

## Availability of data and materials

Raw PacBio HiFi reads for CG44, CG119, CG108, Tx777, and Tx779 are available under NCBI Bioproject PRJEB59044. Raw PacBio HiFi reads for K64 and CML442 are available under NCBI Bioproject PRJEB66502. The genome assemblies used in this work are available at Zenodo, https://zenodo.org/records/14537663. All code is available on Github at https://github.com/dawelab/MaizeSatelliteEvolution.

## Acknowledgements

We thank Arun Seetharam and Matthew Hufford for providing early access to the raw PacBio reads from the TIL lines, and Jonathan Gent, Meghan Brady and Yibing Zeng for critically reading the manuscript. This study was supported by resources and technical expertise from the Georgia Advanced Computing Resource Center, funding from the USDA-ARS to MCR and ESB, as well as a grant from the National Science Foundation (IOS-2040218) to RKD.

## References

1. Charlesworth B, Langley CH, Stephan W. The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics. 1986;112:947–62.

2. Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, et al. Epigenetic patterns in a complete human genome. Science. 2022;376:eabj5089.

3. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, et al. Complete genomic and epigenetic maps of human centromeres. Science. 2022;376:eabl4178.

4. Logsdon GA, Rozanski AN, Ryabov F, Potapova T, Shepelev VA, Catacchio CR, et al. The variation and evolution of complete human centromeres. Nature. 2024;629:136–45.

5. Miga KH, Alexandrov IA. Variation and evolution of human centromeres: A field guide and perspective. Annu Rev Genet. 2021;55:583–602.

6. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, et al. The genetic and epigenetic landscape of the centromeres. Science. 2021;374:eabi7489.

7. Ananiev EV, Phillips RL, Rines HW. Chromosome-specific molecular organization of maize (Zea mays L.) centromeric regions. Proc Natl Acad Sci U S A. 1998;95:13073–8.

8. Schneider KL, Xie Z, Wolfgruber TK, Presting GG. Inbreeding drives maize centromere evolution. Proc Natl Acad Sci U S A. 2016;113:E987–96.

9. Gent JI, Wang N, Dawe RK. Stable centromere positioning in diverse sequence contexts of complex and satellite centromeres of maize and wild relatives. Genome Biol. 2017;18:121.

706 10. Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, et al. De novo
707 assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science.
708 2021;373:655–62.

709 11. Peacock WJ, Dennis ES, Rhoades MM, Pryor AJ. Highly repeated DNA sequence limited to
710 knob heterochromatin in maize. Proc Natl Acad Sci U S A. 1981;78:4490–4.

711 12. Ananiev EV, Phillips RL, Rines HW. A knob-associated tandem repeat in maize capable of
712 forming fold-back DNA segments: are chromosome knobs megatransposons? Proc Natl Acad
713 Sci U S A. 1998;95:10785–90.

714 13. Dawe RK, Lowry EG, Gent JI, Stitzer MC, Swentowsky KW, Higgins DM, et al. A Kinesin-14
715 Motor Activates Neocentromeres to Promote Meiotic Drive in Maize. Cell. 2018;173:839–
716 50.e18.

717 14. Swentowsky KW, Gent JI, Lowry EG, Schubert V, Ran X, Tseng K-F, et al. Distinct kinesin
718 motors drive two types of maize neocentromeres. Genes Dev. 2020;34:1239–51.

719 15. Rhoades MM. Preferential Segregation in Maize. Genetics. 1942;27:395–407.

720 16. Dawe RK. The maize abnormal chromosome 10 meiotic drive haplotype: a review.
721 Chromosome Res. 2022;30:205–16.

722 17. Buckler ES 4th, Phelps-Durr TL, Buckler CS, Dawe RK, Doebley JF, Holtsford TP. Meiotic
723 drive of chromosomal knobs reshaped the maize genome. Genetics. 1999;153:415–26.

724 18. Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, et al. A complete telomere-to-telomere
725 assembly of the maize genome. Nat Genet. 2023;55:1221–31.

726 19. Sevim V, Bashir A, Chin C-S, Miga KH. Alpha-CENTAURI: assessing novel centromeric
727 repeat sequence variation with long read sequencing. Bioinformatics. 2016;32:1921–4.

728 20. Kunyavskaya O, Dvorkina T, Bzikadze AV, Alexandrov IA, Pevzner PA. Automated
729 annotation of human centromeres with HORmon. Genome Res. 2022;32:1137–51.

730 21. Dover GA, Tautz D. Conservation and divergence in multigene families: alternatives to
731 selection and drift. Philos Trans R Soc Lond B Biol Sci. 1986;312:275–89.

732 22. Plohl M, Meštrović N, Mravinac B. Centromere identity from the DNA point of view.
733 Chromosoma. 2014;123:313–25.

734 23. Zhang Y, Chu J, Cheng H, Li H. De novo reconstruction of satellite repeat units from
735 sequence data. Genome Res. 2023;33:1994–2001.

736 24. Liu J, Seetharam AS, Chougule K, Ou S, Swentowsky KW, Gent JI, et al. Gapless assembly
737 of maize chromosomes using long-read technologies. Genome Biol. 2020;21:121.

738 25. Cannon EK, Portwood JL 2nd, Hayford RK, Haley OC, Gardiner JM, Andorf CM, et al.
739 Enhanced pan-genomic resources at the maize genetics and genomics database. Genetics.
740 2024;227:iyae036.

741 26. Page BT, Wanous MK, Birchler JA. Characterization of a maize chromosome 4 centromeric
742 sequence: evidence for an evolutionary relationship with the B chromosome centromere.

743 Genetics. 2001;159:291–302.

744 27. Wlodzimierz P, Hong M, Henderson IR. TRASH: Tandem Repeat Annotation and Structural
745 Hierarchy. Bioinformatics [Internet]. 2023;39. Available from:
746 http://dx.doi.org/10.1093/bioinformatics/btad308

747 28. Dvorkina T, Bzikadze AV, Pevzner PA. The string decomposition problem and its
748 applications to centromere analysis and assembly. Bioinformatics. 2020;36:i93–101.

749 29. Gao S, Yang X, Guo H, Zhao X, Wang B, Ye K. HiCAT: a tool for automatic annotation of
750 centromere structure. Genome Biol. 2023;24:58.

751 30. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656–64.

752 31. Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS. Selection versus demography: a multilocus
753 investigation of the domestication process in maize. Mol Biol Evol. 2004;21:1214–25.

754 32. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, et al. The effects
755 of artificial selection on the maize genome. Science. 2005;308:1310–4.

756 33. Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J. Recent
757 demography drives changes in linked selection across the maize genome. Nat Plants.
758 2016;2:16084.

759 34. Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS. Investigation of the bottleneck
760 leading to the domestication of maize. Proc Natl Acad Sci U S A. 1998;95:4441–6.

761 35. Rice WR. A Game of Thrones at Human Centromeres II. A new molecular/evolutionary
762 model [Internet]. bioRxiv. 2019 [cited 2024 Aug 5]. p. 731471. Available from:
763 https://www.biorxiv.org/content/biorxiv/early/2019/08/10/731471

764 36. Rice W. Why do centromeres evolve so fast: BIR replication, hypermutation, transposition,
765 and molecular-drive. 2020; Available from: https://www.preprints.org/manuscript/202012.0669

766 37. Liu L, Malkova A. Break-induced replication: unraveling each step. Trends Genet.
767 2022;38:752–65.

768 38. Showman S, Talbert PB, Xu Y, Adeyemi RO, Henikoff S. Expansion of human centromeric
769 arrays in cells undergoing break-induced replication. Cell Rep. 2024;43:113851.

770 39. Saayman X, Graham E, Nathan WJ, Nussenzweig A, Esashi F. Centromeres as universal
771 hotspots of DNA breakage, driving RAD51-mediated recombination during quiescence. Mol
772 Cell. 2023;83:523–38.e7.

773 40. Martínez-A C, van Wely KHM. Centromere fission, not telomere erosion, triggers
774 chromosomal instability in human carcinomas. Carcinogenesis. 2011;32:796–803.

775 41. Barra V, Fachinetti D. The dark side of centromeres: types, causes and consequences of
776 structural abnormalities implicating centromeric DNA. Nat Commun. 2018;9:4340.

777 42. Pryor A, Faulkner K, Rhoades MM, Peacock WJ. Asynchronous replication of
778 heterochromatin in maize. Proc Natl Acad Sci U S A. 1980;77:6705–9.

779 43. Ghaffari R, Cannon EKS, Kanizay LB, Lawrence CJ, Dawe RK. Maize chromosomal knobs
780 are located in gene-dense areas and suppress local recombination. Chromosoma.
781 2013;122:67–75.

782 44. Smith GP. Evolution of repeated DNA sequences by unequal crossover. Science.
783 1976;191:528–35.

784 45. Sturtevant AH. The effects of unequal crossing over at the bar locus in Drosophila.
785 Genetics. 1925;10:117–47.

786 46. Bridges CB. The bar “gene” a duplication. Science. 1936;83:210–1.

787 47. Wang Y, Copenhaver GP. Meiotic recombination: Mixing it up in plants. Annu Rev Plant
788 Biol. 2018;69:577–609.

789 48. Stack SM, Shearer LA, Lohmiller L, Anderson LK. Meiotic Crossing Over in Maize Knob
790 Heterochromatin. Genetics. 2017;205:1101–12.

791 49. Kikudome GY. Studies on the Phenomenon of Preferential Segregation in Maize. Genetics.
792 1959;44:815–31.

793 50. Willard HF, Smith KD, Sutherland J. Isolation and characterization of a major tandem repeat
794 family from the human X chromosome. Nucleic Acids Res. 1983;11:2017–33.

795 51. Alkan C, Cardone MF, Catacchio CR, Antonacci F, O’Brien SJ, Ryder OA, et al. Genome-
796 wide characterization of centromeric satellites from multiple mammalian genomes. Genome
797 Res. 2011;21:137–45.

798 52. Waye JS, Durfy SJ, Pinkel D, Kenwrick S, Patterson M, Davies KE, et al. Chromosome-
799 specific alpha satellite DNA from human chromosome 1: hierarchical structure and genomic
800 organization of a polymorphic domain spanning several hundred kilobase pairs of centromeric
801 DNA. Genomics. 1987;1:43–51.

802 53. Vlahovic I, Gluncic M, Rosandic M, Ugarkovic Đ, Paar V. Regular Higher Order Repeat
803 Structures in Beetle Tribolium castaneum Genome. Genome Biol Evol. 2017;9:2668–80.

804 54. Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, et al. Highly accurate long-
805 read HiFi sequencing data for five complex genomes. Sci Data. 2020;7:399.

806 55. Doyle JJ, Doyle JL. A rapid DNA isolation procedure from small quantities of fresh leaf
807 tissues. Phytochem Bull. 1987;19:11–5.

808 56. Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a Galaxy-based web
809 server for genome-wide characterization of eukaryotic repetitive elements from next-generation
810 sequence reads. Bioinformatics. 2013;29:792–3.

811 57. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
812 architecture and applications. BMC Bioinformatics. 2009;10:421.

813 58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J
814 Mol Biol. 1990;215:403–10.

815 59. HMMER: Profile Hidden Markov Models For Biological Sequence Analysis.

60. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275.

61. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014;47:11.12.1–34.

62. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

63. Li H. bioawk: BWK awk modified for biological data [Internet]. Github; 2015 [cited 2024 Mar 28]. Available from: https://github.com/lh3/bioawk

64. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.

65. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37:4572–4.

66. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

67. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021;10. Available from: http://dx.doi.org/10.1093/gigascience/giab008

68. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. 2022;23:258.

69. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021;37:1639–43.

70. Zeng Y, Dawe RK, Gent JI. Natural methylation epialleles correlate with gene expression in maize. Genetics [Internet]. 2023;225. Available from: http://dx.doi.org/10.1093/genetics/iyad146

71. Csardi G, Nepusz T. The igraph software package for complex network research [Internet]. InterJournal. 2006. p. 1695. Available from: https://igraph.org

72. Kassambara A. Network Analysis and Visualization in R: Quick Start Guide. STHDA; 2017.

73. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. Fourth. New York: Springer; 2002. Available from: https://www.stats.ox.ac.uk/pub/MASS4/

74. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines. 1997; Available from: http://stat.ethz.ch/R-manual/R-patched/library/rpart/doc/longintro.pdf

75. Therneau TM. rpart: recursive partitioning. R package version 3. http://www mayo edu/hsr/Sfunc html. 2005;1–23.

76. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/

853 77. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein
854 sequences. Protein Sci. 2018;27:135–45.

855 78. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of
856 high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol.
857 2011;7:539.

858 79. Sievers F, Barton GJ, Higgins DG. Multiple sequence alignments. Bioinformatics [Internet].
859 2020; Available from:
860 https://books.google.com/books?hl=en&lr=&id=hwbQDwAAQBAJ&oi=fnd&pg=PA227&dq=multi
861 ple+sequence+alignments&ots=SztFKaikuS&sig=1LpH74cdocP-U_pf9oRB6DjMgd0

862 80. Rice PM, Bleasby AJ, Ison JC. EMBOSS User's Guide: Practical Bioinformatics. Cambridge
863 University Press; 2011.

864 81. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business
865 Media; 2009.

866 82. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
867 Bioinformatics. 2018;34:i884–90.

868 83. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM
869 [Internet]. arXiv [q-bio.GN]. 2013. Available from: http://github.com/lh3/bwa

870 84. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness,
871 and phasing assessment for genome assemblies. Genome Biol. 2020;21:245.

872 85. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for
873 exploring deep-sequencing data. Nucleic Acids Res. 2014;42:W187–91.

874

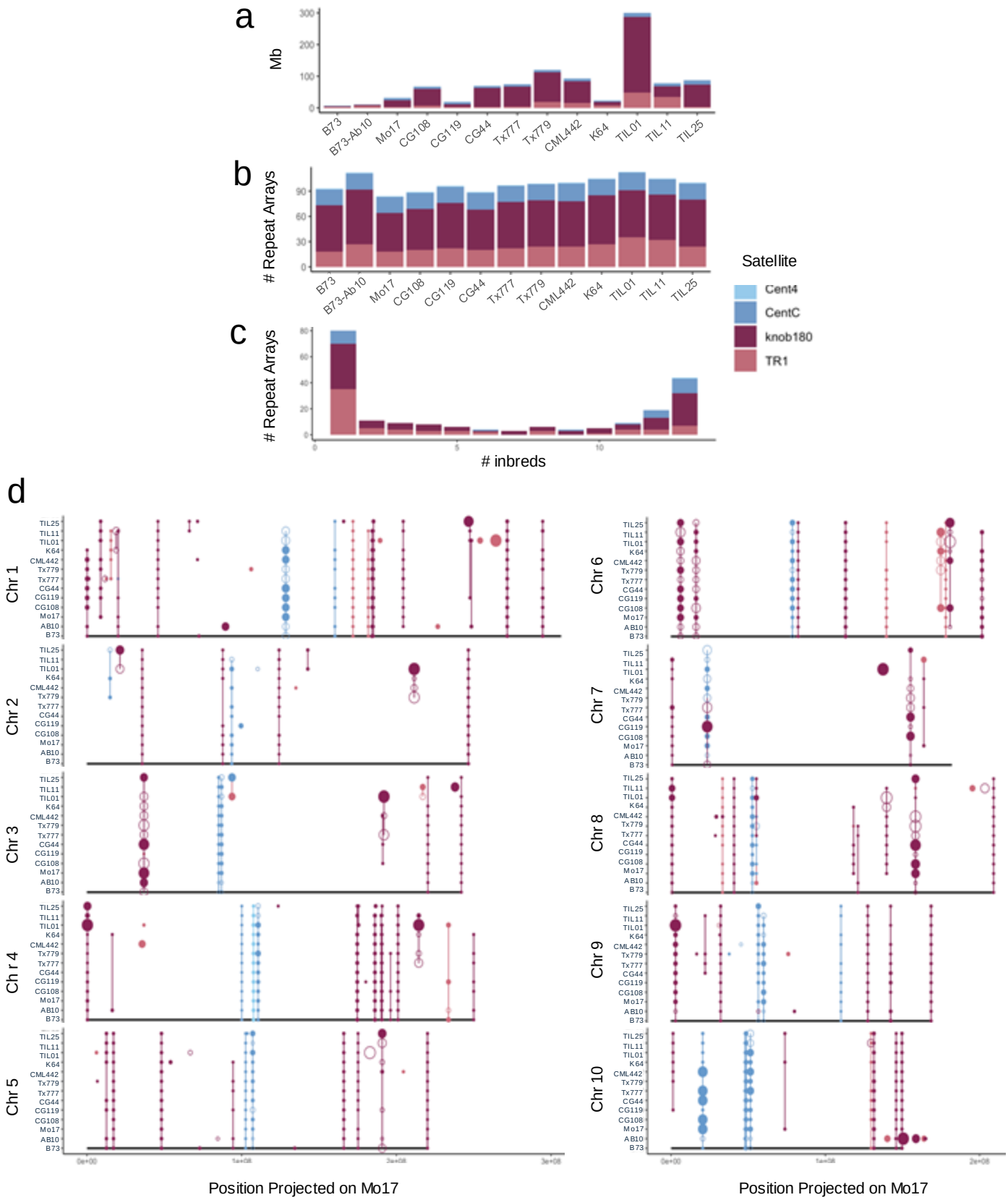



**Figure 1.** Assembly repeat content. **a**) Total repeat content by satellite. **b**) Repeat array counts. **c**) Frequency spectrum of arrays. **d**) Satellite arrays projected onto the Mo17 assembly. Open circles represent arrays that are not fully assembled (contain an N gap). Filled circles represent fully assembled arrays (do not contain an N gap). Lines connecting arrays represent homology based on synteny with conserved genes.

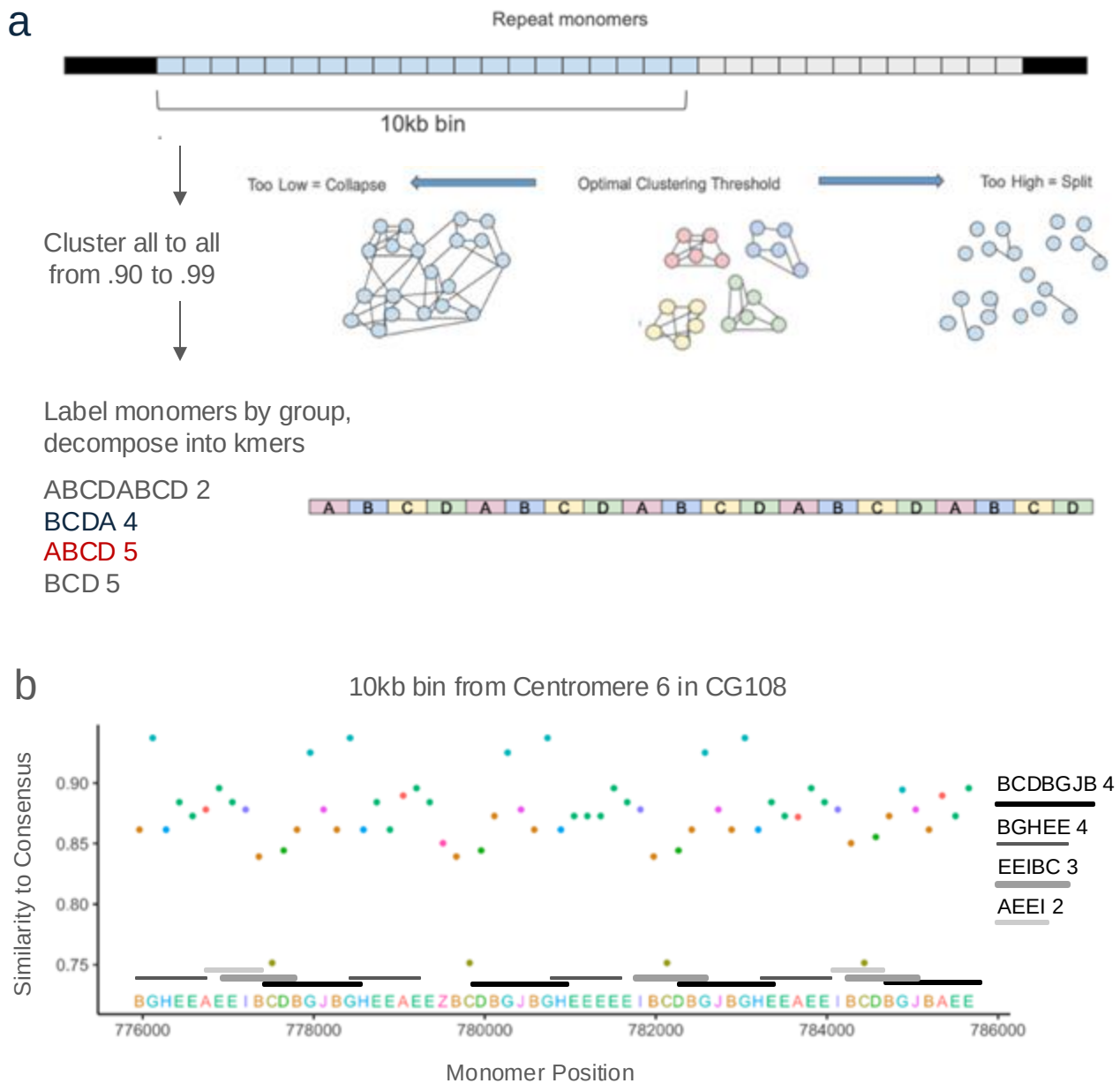



**Figure 2.** Local HOR identification. **a**) Monomers in 10kb bins are extracted and compared all-to-all using BLAT. Jaccard similarity scores are calculated. Then, similarity networks are generated for thresholds from .90 to .99, in increments of .01. Each node represents a monomer sequence. If two monomer sequences are at least as similar as the threshold, monomers are connected with an edge. Each distinct cluster in the network is labeled with a letter. Private clusters, or clusters that only contain a single monomer, are labeled with "Z". Monomers, with labels of their cluster identity, are then put back in their original genomic order, resulting in a character string. The character string is decomposed into k-mers, starting with k=3 until all k-mers have a frequency of 1. K-mers are then filtered to remove larger k-mers that occur less frequently than its subset k-mers, smaller k-mers that occur as frequently as larger k-mers containing it, and k-mers that contain the private monomer "Z". In the example given, ABCD is the largest most abundant HOR. **b**) Sample HOR region, where monomers are labeled and colored by cluster identity. HOR k-mers and their frequencies are listed on the side, and each repeated pattern in the arrays is marked with a line corresponding to its k-mer.

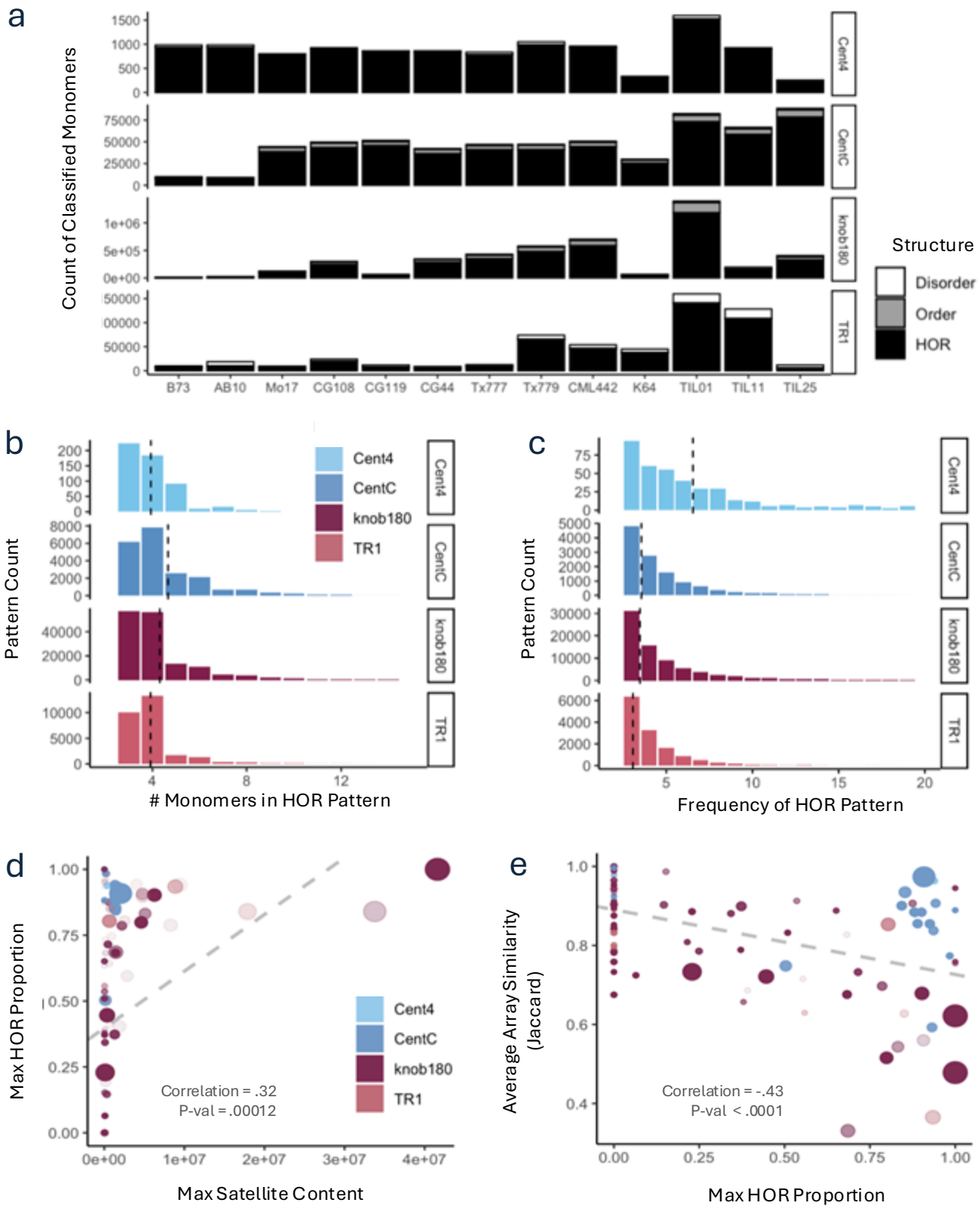
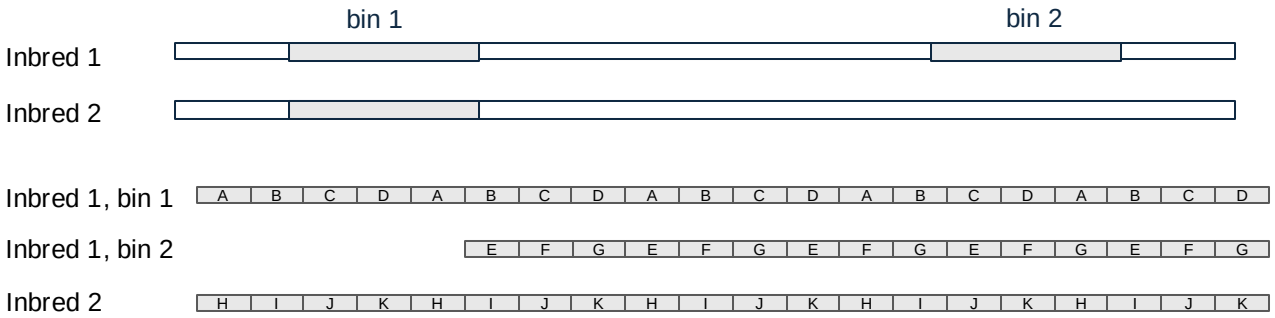
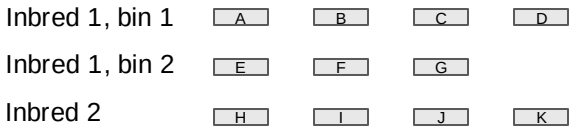



**Figure 3.** HOR Pattern Frequency and Length. a) Monomer count by structure classification for each line. b) HOR pattern length distribution. c) HOR pattern frequency distribution. Dashed line marks mean d) The maximum satellite content is positively correlated the maximum HOR proportion of the arrays. e) The maximum HOR proportion is negatively correlated with average Jaccard similarity of arrays within homolog groups.

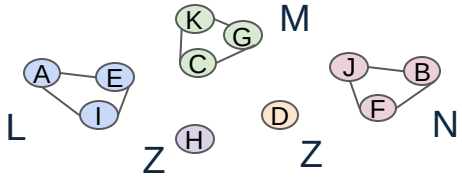
1. Collect HOR regions with the same optimal clustering threshold




2. Generate consensus sequences for all monomers in local HOR patterns




3. Cluster consensus monomers and relabel, private monomers labeled “Z”





**LNM is a shared HOR pattern**

4. Recalculate purity

| Region | Local Purity | Shared Purity |
|---|---|---|
| Inbred 1, bin 1 | 1.00 | .75 |
| Inbred 1, bin 2 | 1.00 | 1.00 |
| Inbred 2 | 1.00 | .75 |

**Figure 4.** Shared HOR Identification**.** 1. HORs with the same optimal clustering thresholds are extracted. This can include multiple distinct regions from the same array. 2. Consensus sequences for each monomer subtype in an original HOR pattern. 3. Consensus sequences are then compared all-to-all with a cutoff equal to the original shared optimal clustering threshold, and a network is generated. Monomers are labeled by their cluster identity and returned to their original orders. Kmers are then generated. Only kmers that exist in one or more regions are shared HORs. 4. Array purity is recalculated as (N monomers in identifiable arrays) /(Total Monomers).

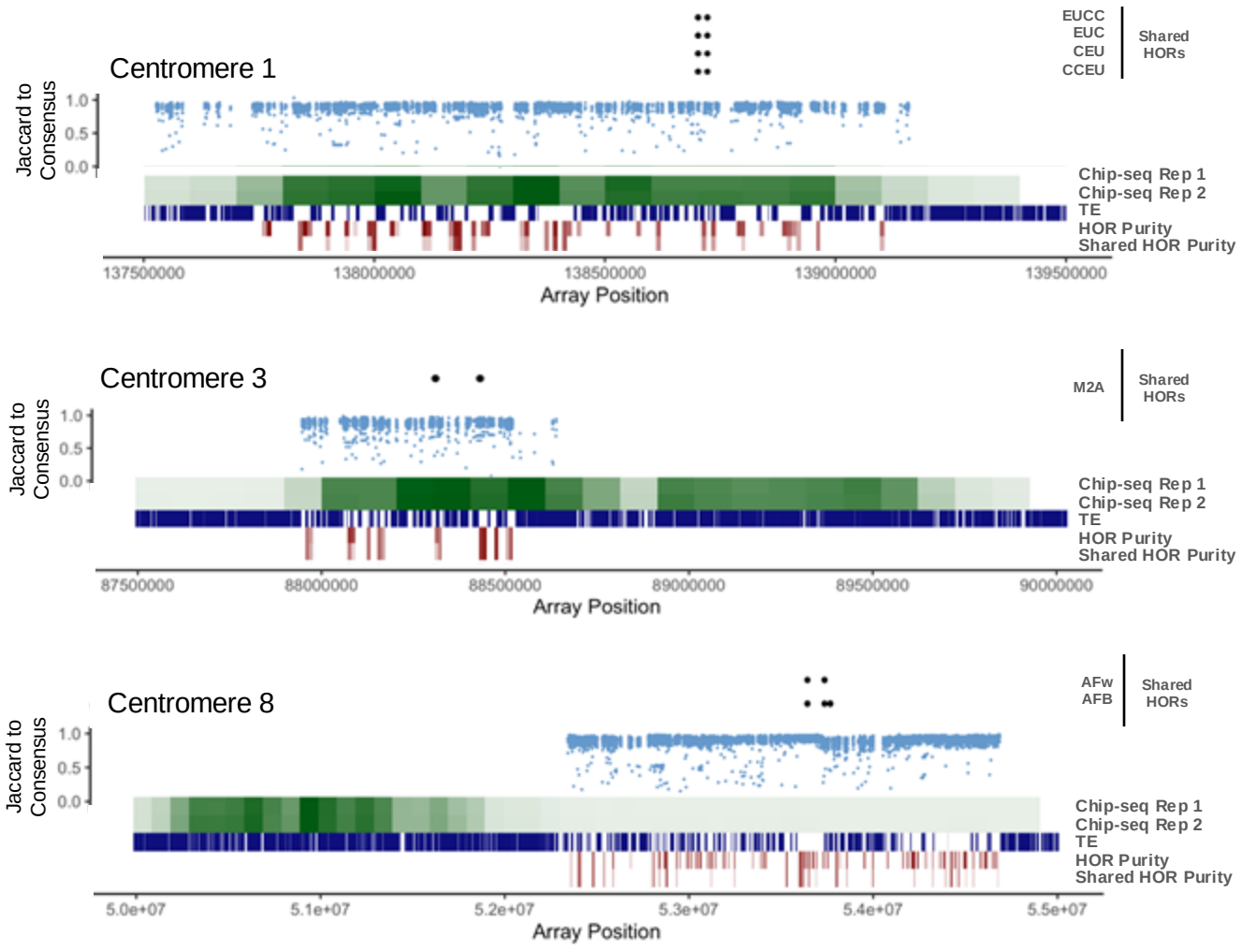



**Figure 5.** CENH3 Chip-seq on Mo17. For each centromere, shared HOR patterns that occur in at least two distinct HOR regions are indicated in the upper panels, where dots indicate presence in array. The Jaccard similarity of each monomer to the consensus is shown as a dotplot. The relative CENH3 ChIP-seq density in 100kb bins (of two independent replicates) is shown in shades of green. TEs are represented by blue boxes. HOR purity and shared HOR purity in 10kb bins is shown in red, where darker shades represent higher purity values.

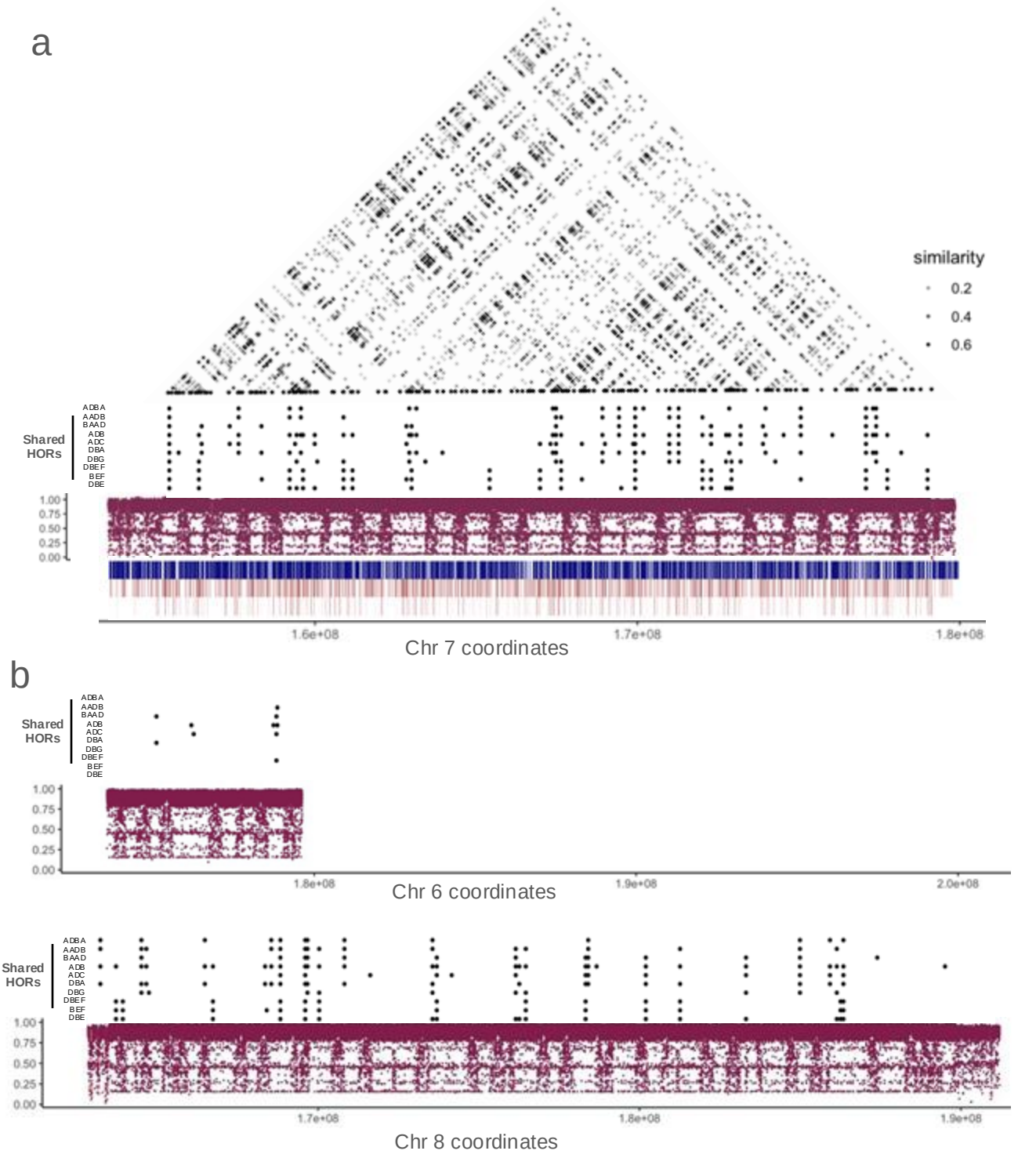



**Figure 6.** Shared HORs among knob180 arrays. **a**) The large knob on chromosome 7L in CG108. The upper triangle shows HOR pattern similarity of all HOR regions of the array. Darker shades indicate a greater proportion of shared patterns, up to 1. Below the triangle are high frequency shared HORs, where dots indicate presence in array, and the x axis is genomic position. Below the shared HORs, in maroon, is a dot plot of monomers displayed according to similarity to the consensus as a Jaccard index on the y axis. Each maroon dot represents a monomer. The heat maps, from upper to lower, show TEs (blue), HOR purity (red), and shared HOR purity (red) (same labeling as Figure 5). **b**) The knobs on chromosome 6L and 8L in CG108. Each panel shows high frequency shared HORs, and dot plots of monomers displayed according to similarity to the consensus as a Jaccard index.

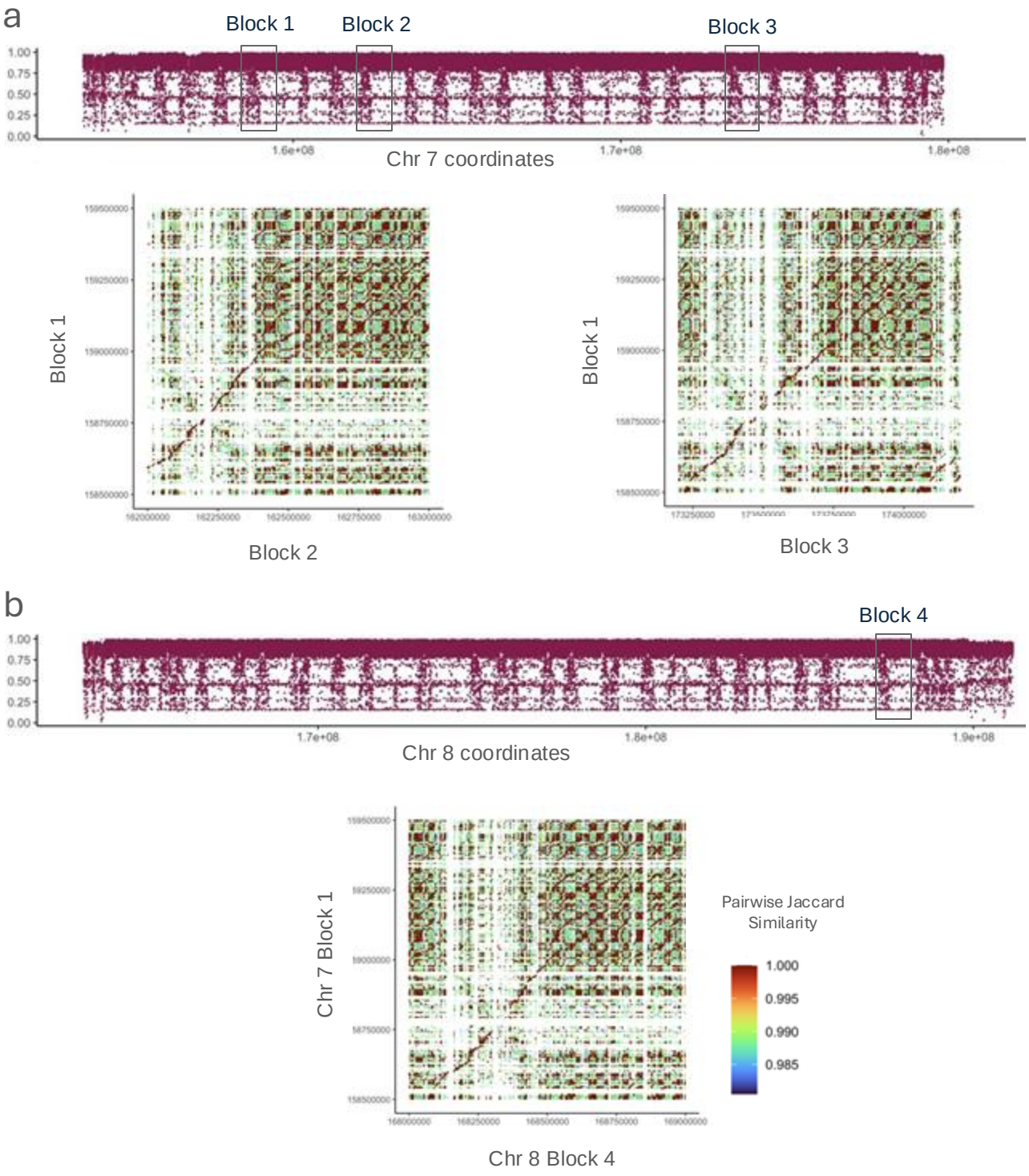



**Figure 7.** Similarity Blocks in knobs. **a**) Similarity blocks in the large knob on chromosome 7L in CG108. The maroon similarity to consensus plot (same plot as Figure 6a) shows the locations of three similarity blocks identified by their characteristic features of having a region of low similarity to the consensus followed by a region of high similarity to the consensus on a megabase scale. Below, there are two traditional dot plots comparing blocks 1 and 2 and blocks 1 and 3. Note the diagonal lines of high similarity. **b**) A similarity block in the large knob on chromosome 8L in CG108 (labeled block 4). Below is a dot plot comparing chromosome 7 block 1 to chromosome 8 block 4. Note again the diagonal line of high similarity.

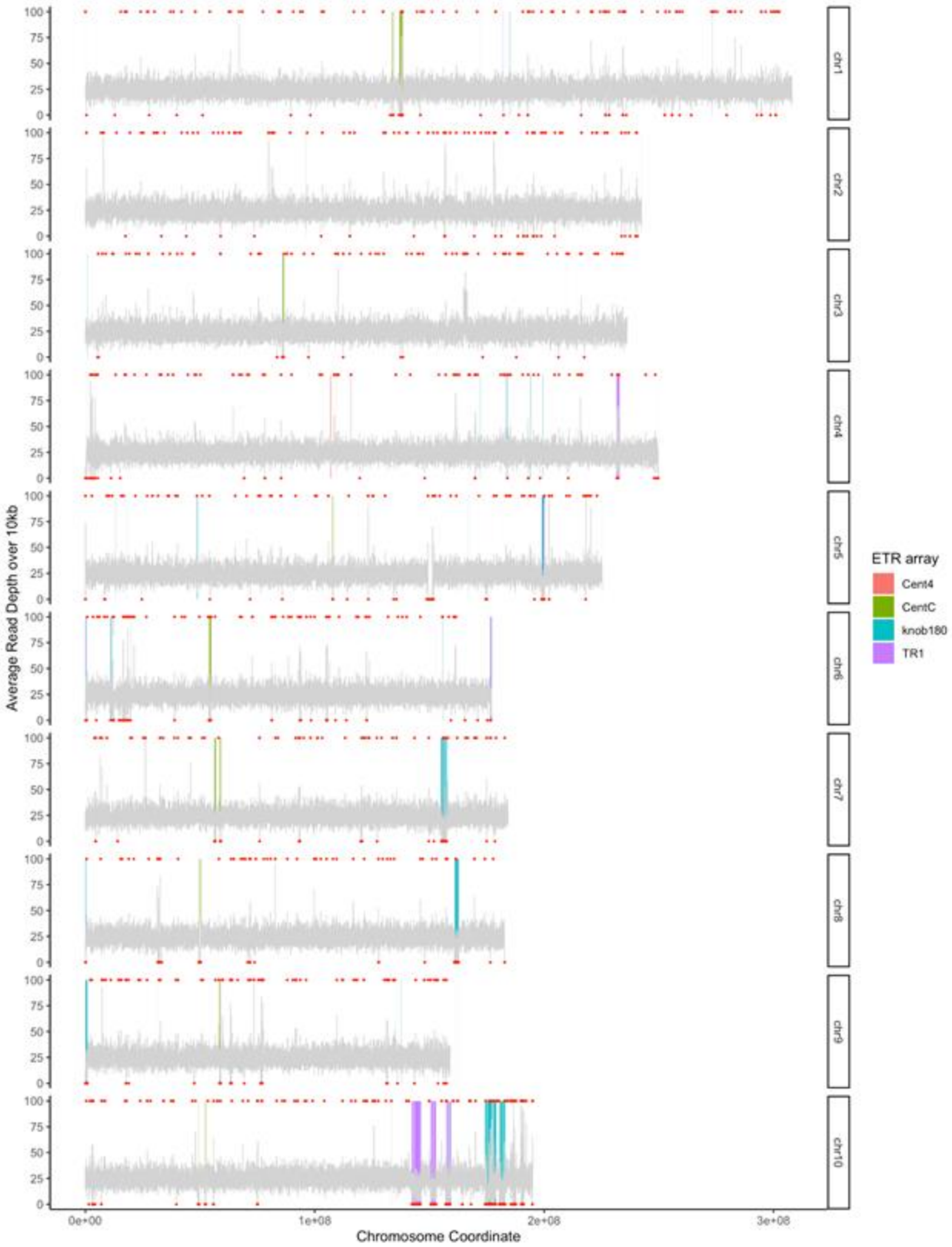



**Figure S1.** PacBio HiFi reads aligned to Gapless AB10 Pseudomolecules. Grey lines represent 10kb average read depth. Red dots above and below make 10kb bins where the average read depth falls outside two standard deviations of the mean depth. Satellite arrays are represented by colored boxes.

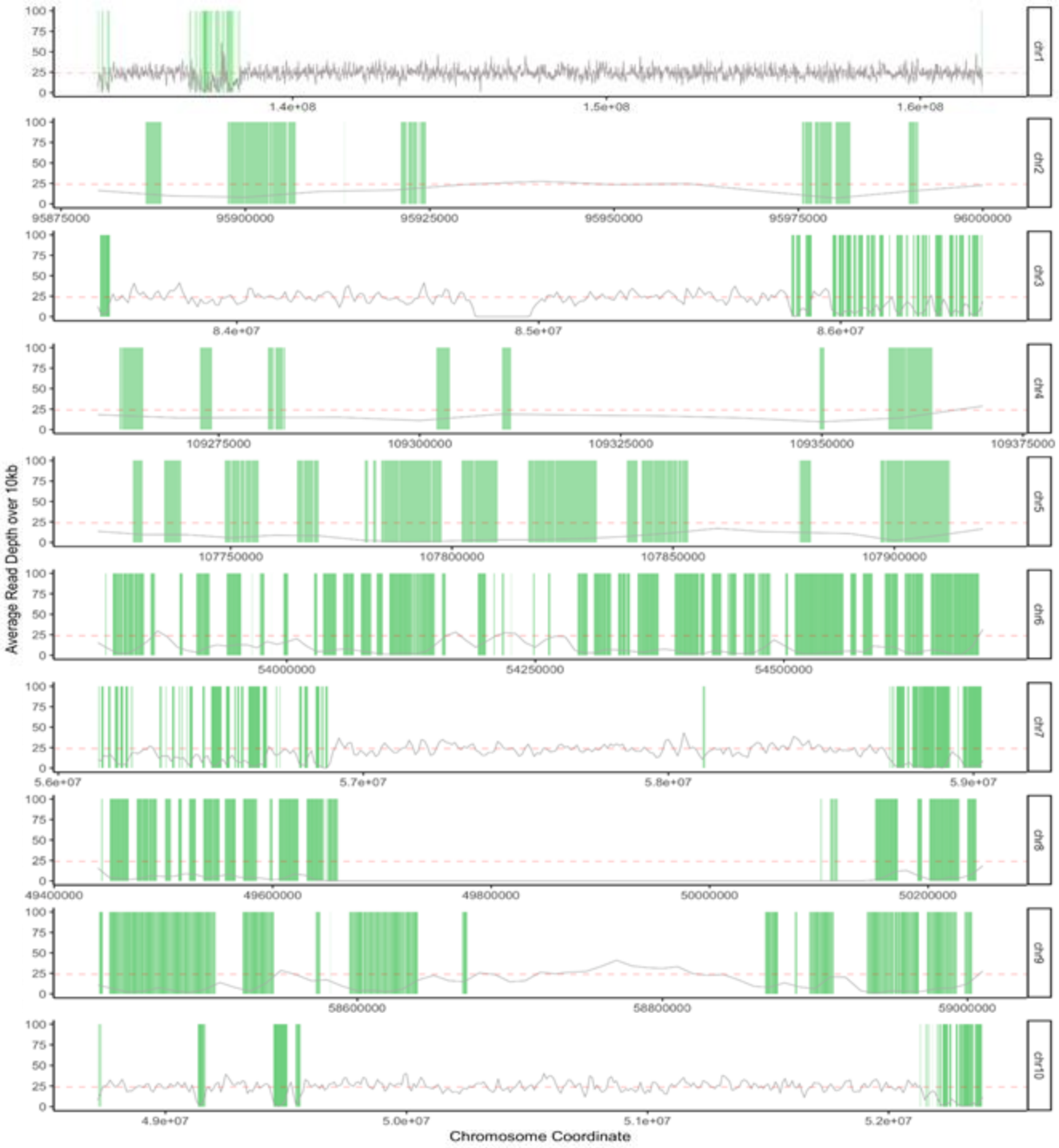



**Figure S2.** PacBio HiFi reads aligned to Gapless AB10 Centromeres. Grey lines represent 10kb average read depth. Red dotted line represents the average read depth. CentC satellites are represented by green boxes.

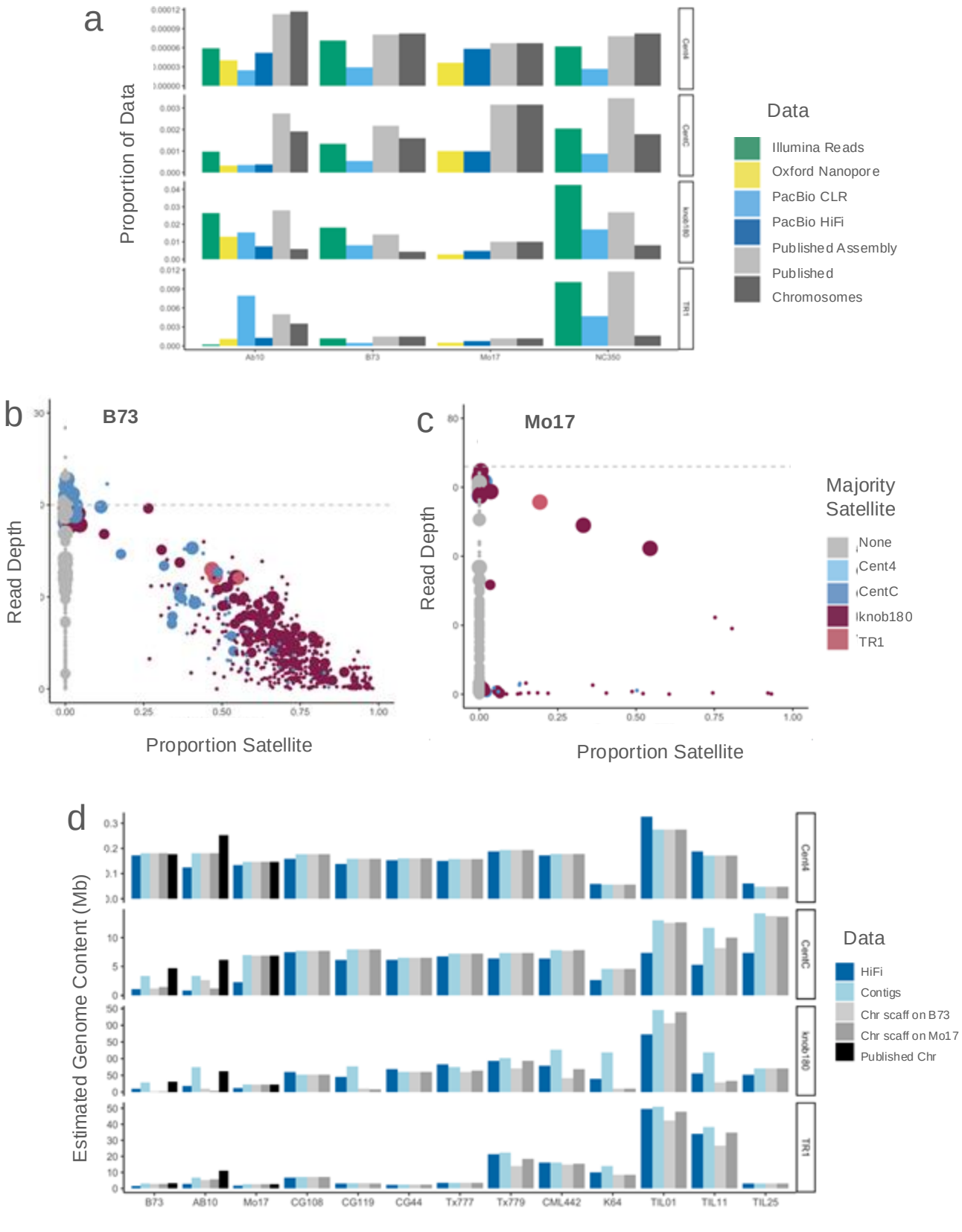



**Figure S3.** Assembly Repeat Biases. a) Satellite content of raw reads and assembled genomes from 4 previously-assembled inbred lines. b) Hifiasm-generated contig read support and satellite content for B73. Size represent contig size. c) Hifiasm-generated contig read support and satellite content for Mo17. d) Satellite content of PacBio HiFi reads and hifiasm-generated contigs for 13 inbreds, compared to previously-published genome assembly and pseudo-molecules generated from the high-confidence set of hifiasm-generated contigs, scaffolding on B73 and Mo17.

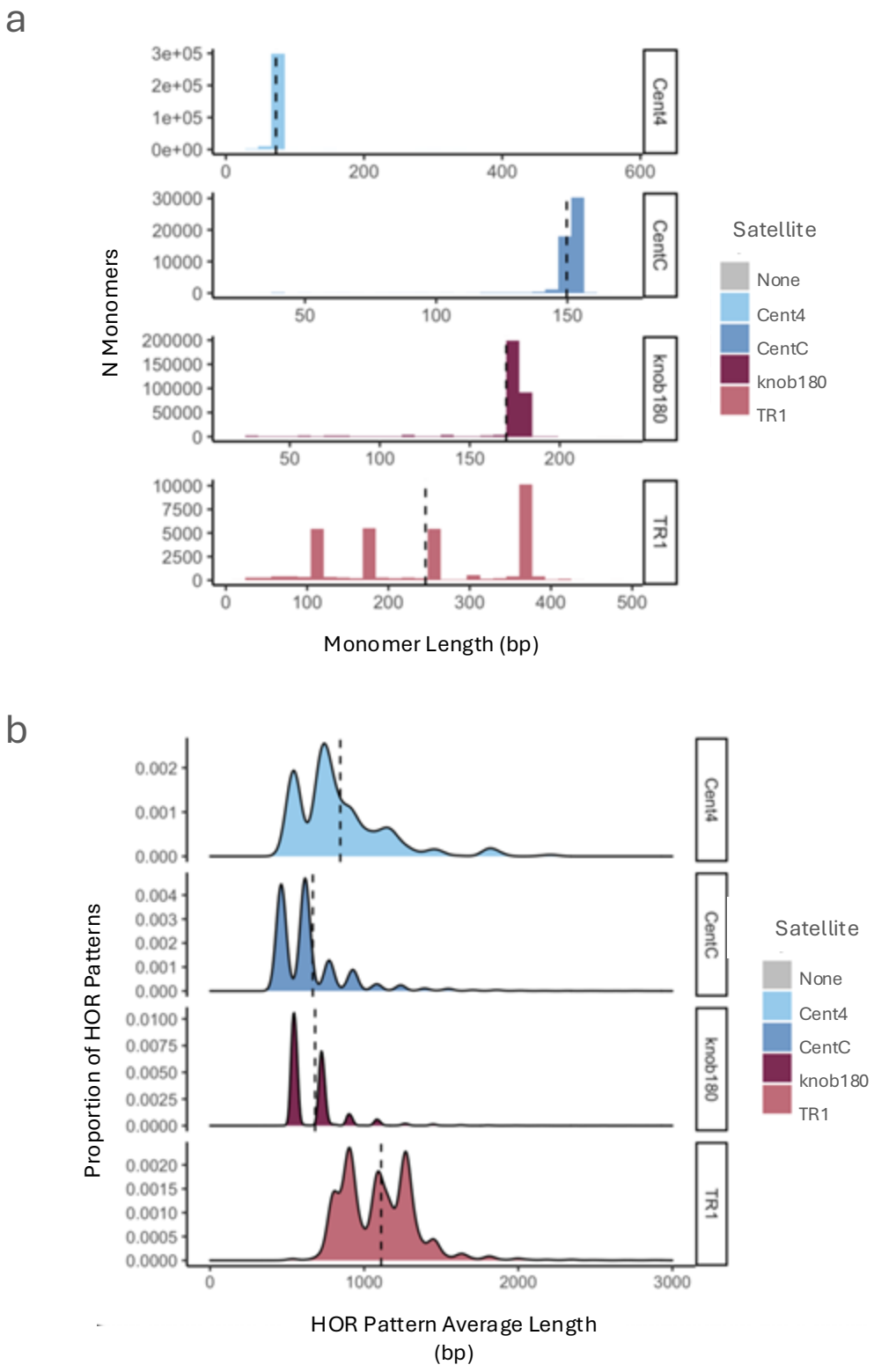



**Figure S4.** Monomer and Pattern Lengths in bp. a) All monomer lengths in CG108 in bp. b) Average Lengths of all HOR Patterns of >=3x occurrences in all 13 lines.

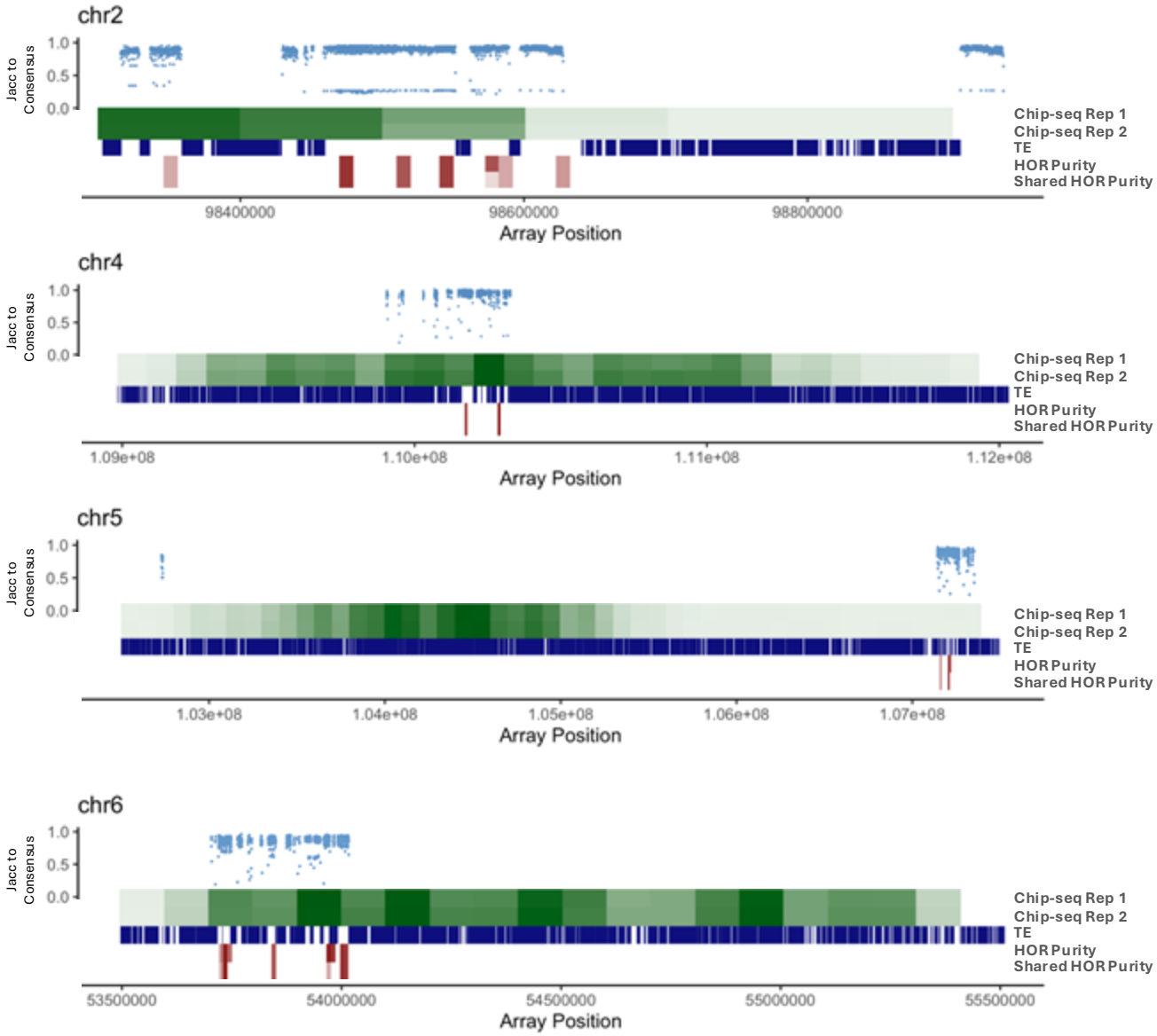



**Figure S5.** CENH3 Chip-seq on Mo17**.** Centromere structure of chromosomes 2, 4, 5, and 6. Relative Chip-seq density in 100kb bins represented by shades of green. TE presence is represented by blue boxes. HOR purity and shared HOR purity are represented in red in 10kb bins. Darker shades of red represent higher purity values.

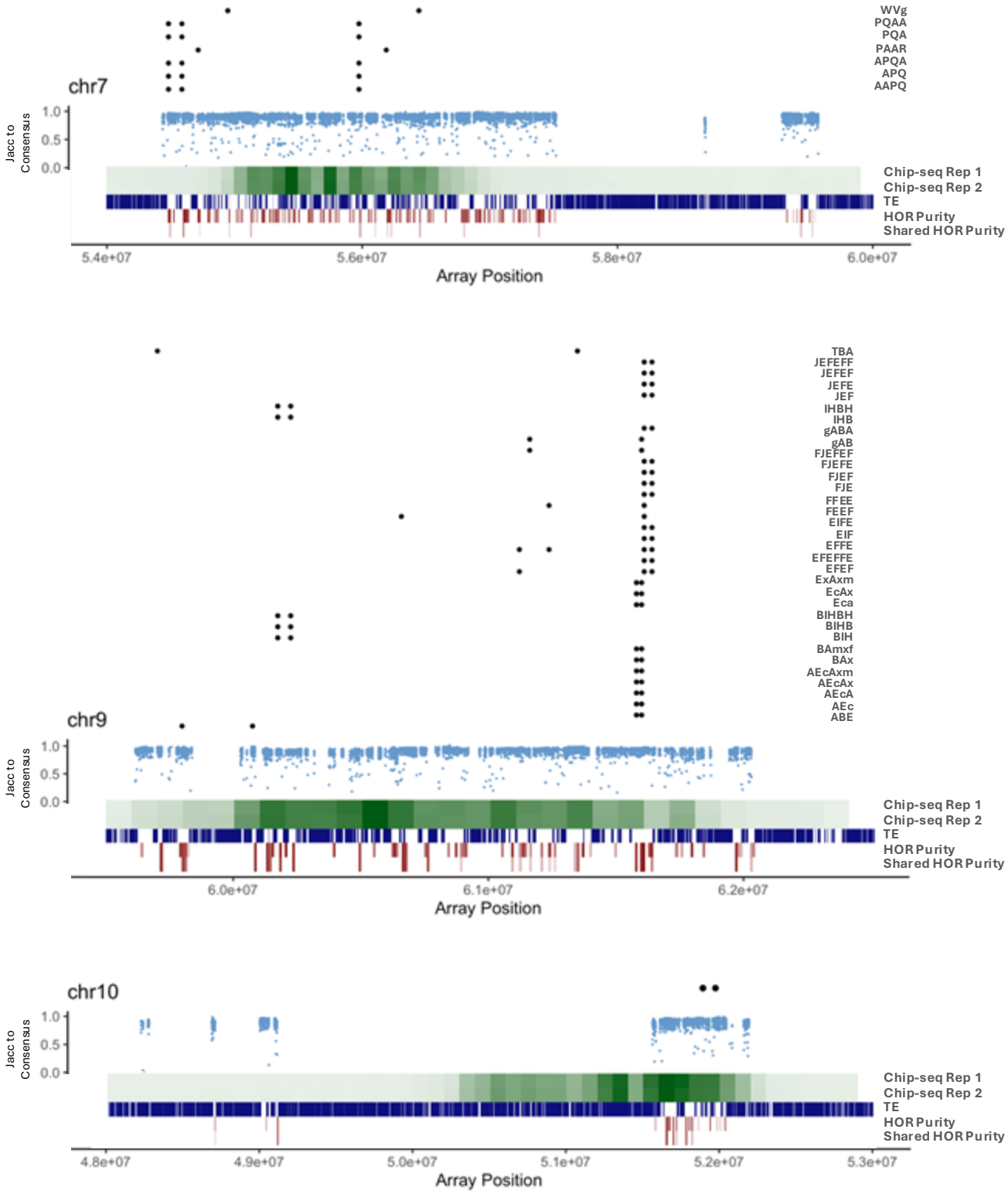



**Figure S6.** CENH3 Chip-seq on Mo17. Centromere structure of chromosomes 7, 9, and 10. Shared HOR patterns that occur in at least two distinct HOR regions. Dot indicates presence in array, x axis is genomic position. Relative Chip-seq density in 100kb bins represented by shades of green. TE presence is represented by blue boxes. HOR purity and shared HOR purity are represented in red in 10kb bins. Darker shades of red represent higher purity values.

**Supplemental Tables**

Table S1. Proportion of satellite repeat content in maize pangenome data.

| Line | Data | Total Length (Mb) | CentC (%) | Cent4 (%) | Knob180 (%) | TR1 (%) |
|---|---|---|---|---|---|---|
| **Ab10** | CLR Assembly | 2,243 | .27% | .011% | 2.8% | .50% |
| | CLR Pseudomolecules | 2,163 | .19% | .012% | **.059%** | 0.35% |
| | Illumina 80x | 184,444 | .097% | .0059% | 2.6% | **.020%** |
| | PacBio CLR 62x | 151,353 | **.035%** | **.0025%** | 1.5% | **.079%** |
| | ONT 50x | 118,869 | **.032%** | .004% | 1.3% | **.11%** |
| **B73** | Assembly | 2,183 | .22% | .0082% | 1.42% | .15% |
| | Pseudomolecules | 2,132 | .16% | .0084% | .44% | .15% |
| | Illumina 21x | 47,778 | .13% | .0071% | 1.8% | .12% |
| | PacBio CLR 79x | 172,780 | **.054%** | .0029% | .80% | .048% |
| **NC350** | Assembly | 2,291 | .34% | .0079% | 2.7% | 1.17% |
| | Pseudomolecules | 21,64 | .18% | .0080% | .80% | .16% |
| | Illumina 69x | 158,744 | .20% | 0.0062% | 4.3% | 1.0% |
| | PacBio CLR 141x | 323,810 | .088% | .0027% | 1.71% | 0.47% |

Red indicates the value is <75% of the assembly content, bold red indicates the value is <25% of the assembly content. Green indicates the values >125% of assembly content, bold green indicates value is >175% of assembly content.

Table S2. Proportion of satellite repeat content in HiFi CCS reads and primary contigs.

| Line | Data | Length (Mb) | CentC | Cent4 | Knob180 | TR1 |
|---|---|---|---|---|---|---|
| **B73** | CCS (20x) | 48,075 | .046% | .0072% | .37% | .069% |
| | Contigs | 2,186 | .15% | .0083% | 1.3% | .14% |
| | Assembly | 2,121 | .067% | .008% | .10% | .13% |
| **Ab10** | CCS (23x) | 53,424 | .038% | .0053% | .74% | .13% |
| | Contigs | 2,285 | **.15%** | .008% | **3.21%** | **.29%** |
| | Assembly | 2,133 | .057% | .008% | **.15%** | **.26%** |
| **Mo17** | CCS (66x) | 151,123 | .10% | .006% | .47% | .076% |
| | Contigs | 2,707 | **.26%** | .005% | **.82%** | .098% |
| | Assembly | 2,151 | **.32%** | .006% | **1.0%** | .12% |
| **CG108** | CCS (27x) | 62,284 | .33% | .007% | 2.6% | .28% |
| | Contigs | 2,258 | .34% | .008% | 2.3% | .32% |
| | Assembly | 2,205 | .35% | .008% | 2.4% | .32% |
| **CG44** | CCS (21x) | 48,968 | .26% | .007% | 2.9% | .10% |
| | Contigs | 2,249 | .29% | .007% | 2.7% | .11% |
| | Assembly | 2,217 | .29% | .007% | 3.0% | .11% |
| **CG119** | CCS (17x) | 38,822 | .26% | .006% | 1.9% | .13% |
| | Contigs | 2,253 | .35% | .007% | **3.4%** | .14% |
| | Assembly | 2,140 | .37% | .007% | **.36%** | .15% |
| **Tx777** | CCS (21x) | 48,209 | .30% | .007% | 3.6% | .15% |
| | Contigs | 2,272 | .32% | .007% | 3.3% | .16% |
| | Assembly | 2,219 | .33% | .007% | 2.9% | .16% |
| **Tx779** | CCS (19x) | 44,744 | .27% | .008% | 4.0% | .91% |
| | Contigs | 2,315 | .32% | .008% | 4.4% | .97% |
| | Assembly | 2,276 | .32% | .008% | 4.1% | .81% |
| **K64** | CCS (23x) | 52,074 | .12% | .003% | 1.73% | .44% |
| | Contigs | 2,333 | .20% | .002% | **5.1%** | .59% |
| | Assembly | 2,145 | **.22%** | .002% | .47% | .39% |
| **CML442** | CCS (19x) | 43,183 | .28% | .008% | 3.43% | .71% |
| | Contigs | 2,413 | .33% | .008% | 5.20% | .67% |
| | Assembly | 2,255 | .35% | .007% | 3.0% | .69% |
| **TIL01** | CCS (28x) | 76,038 | .27% | .012% | 6.4% | 1.8% |
| | Contigs | 2,766 | .47% | .010% | 8.9% | 1.9% |
| | Assembly | 2,540 | **.50%** | .010% | 9.4% | 1.9% |
| **TIL11** | CCS (21x) | 54,395 | .20% | .007% | 2.1% | 1.32% |
| | Contigs | 2,459 | **.48%** | .007% | **4.8%** | 1.55% |
| | Assembly | 2,263 | **.44%** | .007% | 1.5% | 1.5% |
| **TIL25** | CCS (25x) | 55,175 | .33% | .003% | 2.3% | .14% |
| | Contigs | 2,134 | **.67%** | .002% | 3.3% | .14% |
| | Assembly | 2,085 | **.66%** | .002% | 3.4% | .14% |

Red indicates the value is <75% of the CCS content, bold red indicates the value is <25% of the CCS content. Green indicates the values >125% of CCS content, bold green indicates value is >175% of CCS content.

Table S3. Contig Counts and Assembly N Gaps.

| Line | Contigs | Anchored Contigs | Contigs in Assembled Chr | Anchored Contigs in Assembled Chr | N Gaps in Assembly | N Gaps in Arrays |
|---|---|---|---|---|---|---|
| **B73** | 1248 | 1240 | 136 | 134 | 153 | 18 |
| **Ab10** | 2117 | 2117 | 162 | 162 | 148 | 16 |
| **Mo17** | 8795 | 8771 | 56 | 56 | 46 | 0 |
| **CG108** | 774 | 774 | 57 | 57 | 72 | 2 |
| **CG44** | 730 | 730 | 88 | 88 | 101 | 1 |
| **CG119** | 1074 | 1074 | 154 | 154 | 162 | 5 |
| **Tx777** | 910 | 910 | 217 | 217 | 224 | 11 |
| **Tx779** | 777 | 763 | 277 | 275 | 288 | 11 |
| **K64** | 1259 | 1252 | 78 | 77 | 93 | 11 |
| **CML442** | 1580 | 1576 | 124 | 122 | 141 | 9 |
| **TIL01** | 6304 | 6224 | 176 | 164 | 166 | 12 |
| **TIL11** | 2113 | 2009 | 131 | 111 | 121 | 16 |
| **TIL25** | 906 | 894 | 74 | 73 | 64 | 6 |

Table S4. Average Homologous Array Similarity.

| Satellite | # Arrays | Mean Jaccard Similarity | Mean Jaccard Similarity, no N gaps | Mean Jaccard Similarity, Maize | Mean Jaccard Similarity, Teosinte |
|---|---|---|---|---|---|
| **All** | 95 | .82 | .82 | .86 | .88 |
| **Cent4** | 1 | .96 | .96 | .98 | .94 |
| **CentC** | 18 | .86 | .88 | .89 | .84 |
| **knob180** | 64 | .82 | .82 | .87 | .89 |
| **TR1** | 12 | .74 | .74 | .76 | .90 |

Table S5. Shared HOR's for Mo17 and CG108.

| Inbred | Satellite | # HOR regions in analysis | # HOR regions with shared HOR's | % of HOR regions with shared HOR's |
|---|---|---|---|---|
| Mo17 | All | 1470 | 200 | 14% |
| | Cent4 | 8 | 7 | 88% |
| | CentC | 342 | 46 | 13% |
| | knob180 | 1006 | 76 | 8% |
| | TR1 | 114 | 71 | 62% |
| CG108 | All | 3343 | 637 | 19% |
| | Cent4 | 8 | 7 | 88% |
| | CentC | 416 | 53 | 13% |
| | knob180 | 2512 | 335 | 13% |
| | TR1 | 407 | 242 | 59% |

Table S6. Knob High-Frequency HORs.

| Pattern | Total Frequency | N regions | N freq in K6L | N freq in K7L | N freq in K8L | Mean Distance Between Regions | Mean Frequency within each Region |
|---|---|---|---|---|---|---|---|
| **ADBA** | 102 | 33 | 0 | 65 | 37 | 1,163,183 | 4.5 |
| **AADB** | 102 | 32 | 0 | 52 | 50 | 1,473,366 | 5.1 |
| **BAAD** | 85 | 30 | 2 | 48 | 35 | 1,473,366 | 4.1 |
| **ADB** | 348 | 67 | 0 | 181 | 167 | 677,293 | 7.7 |
| **ADC** | 120 | 36 | 8 | 63 | 49 | 1,146,341 | 6.5 |
| **DBA** | 141 | 46 | 4 | 85 | 52 | 953,519 | 5.8 |
| **DBG** | 116 | 30 | 2 | 61 | 53 | 1,227,805 | 4 |
| **DBEF** | 70 | 30 | 0 | 41 | 29 | 1,481,578 | 4.8 |
| **BEF** | 87 | 36 | 2 | 53 | 32 | 1,185,263 | 4.6 |
| **DBE** | 113 | 40 | 0 | 63 | 50 | 1,128,822 | 5.8 |

Table S7. Knob High-Frequency HOR Monomers.

| Monomer Name | Frequency | Clustering Threshold |
|---|---|---|
| **knob180_A** | **2,504** | GGGGTTGTGTGGCCATTGATCGTCGACCAGAGGCTCATACA CCTCACCCCACATATGTTTCCTTGCCATAGATCACATTCTTG GATTTCTGGTGGAGACCATTTCTTGGTCAAAAATCCGTAGGT GTTAGCCTTCGGTATTATTGAAAATGGTCGTTCATGGCTATT TTCGACAAAA |
| **knob180_B** | **1,077** | GGGGTTGTGTGGCCATTTATCATCGACTAGAGGCTCATAAA CCTCACCCCACATATGTTTCCTTGCCATAGATTACATTCTTG GATTTCTGGTGGAAACCATTTCTTGGTTAAAAACTCGTACGT GTTAGCCTTCGGTATTATTGAAAATGGTCATTCATGGCTATT TTCGGCAAAATGG |
| **knob180_C** | **237** | GGGGTTGTGTGGCCATTTATCATCGACTAGAGGCTCATAAA TCTCACCCCACATATGTTTCCTTGCCATAGATCACATTCTTG GATTTTTGGTGGAGACCATTTCTTGGTCAAAAACTCGTACGT GTTAGCCTTCGGTATTATTGAAAATGGTCGTTCATGGCTATT TTCGGCAAAATGG |
| **knob180_D** | **1,462** | GGGGTTGTGTGGCCATTGATCATCGACCAGAGCTCATACAC CTCACCCCACATATGTTTCCTTGCCATAGATCACATTCTTGG ATTTCTGGTGGAGACCATTTCTTGGTCTAAAATCCGTAGGTG TTAGCCTCTAGTATTATTGAAAATGGTCGCTCATGGCTATTT TCAA |
| **knob180_E** | **334** | GGGGTTGTGTGGCCATTGATCATCGACCAGAGGCTCGTACA CCTCACCCCACATATGTTTCCTTGCCATAGATCACATTCTTG GATTTCTGGTGGAGACCATTTCTTGGTCAAAAATCCGTACGT GTTAGCCTTTGGTATTTTTGAAAATGGTCATTCATGGCTATT TTCGACAAAA |
| **knob180_F** | **450** | GGGGTTGTGTGGCCATTGATCTTCGACCAGAGGCTCATACA CCTCACTACACATATGTTTCCTTGCCATAGATCACATTCTTG ATTTATGGTGGAGACCATTTCTTGGTCAAAAATCCGTAGGT GTTAGCCTTCAGTGTCATTGAAAATGTCGTTCATGGCTATTT TCGACAAA |
| **knob180_G** | **204** | GGGGTTGTGTGGCCATTGATCGTCGACCAGAGGCTCATACA CCTCACCCCACATATGTTTCCTTGTCGTAnGATCACATTCTTG GATTTCTGGTGGAGACCATTTCTTGGTCAAAAATCCGTAGGT GTTAGCCTTCGATATTATTGAAAATGGTCATTCATGGCTATT TCGGCAAAATGG |