HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2024 February 16.

Published in final edited form as:

Cell Rep. 2023 October 31; 42(10): 113178. doi:10.1016/j.celrep.2023.113178.

Variation in the CENP-A sequence association landscape across diverse inbred mouse strains

Uma P. Arora[1,2,*], **Beth A. Sullivan**[3], **Beth L. Dumont**[1,2,4,5,*]

[1]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

[2]Graduate School of Biomedical Sciences, Tufts University, 136 Harrison Avenue, Boston, MA 02111, USA

[3]Department of Molecular Genetics and Microbiology, Duke University Medical Center, 213 Research Drive, Box 3054, Durham, NC 27710, USA

[4]Graduate School of Biomedical Science and Engineering, University of Maine, 5775 Stodder Hall, Room 46, Orono, ME 04469, USA

[5]Lead contact

SUMMARY

Centromeres are crucial for chromosome segregation, but their underlying sequences evolve rapidly, imposing strong selection for compensatory changes in centromere-associated kinetochore proteins to assure the stability of genome transmission. While this co-evolution is well documented between species, it remains unknown whether population-level centromere diversity leads to functional differences in kinetochore protein association. Mice (*Mus musculus*) exhibit remarkable variation in centromere size and sequence, but the amino acid sequence of the kinetochore protein CENP-A is conserved. Here, we apply *k*-mer-based analyses to CENP-A chromatin profiling data from diverse inbred mouse strains to investigate the interplay between centromere variation and kinetochore protein sequence association. We show that centromere sequence diversity is associated with strain-level differences in both CENP-A positioning and sequence preference along the mouse core centromere satellite. Our findings reveal intraspecies sequence-dependent differences in CENP-A/centromere association and open additional perspectives for understanding centromere-mediated variation in genome stability.

In brief

*Correspondence: uma.arora@jax.org (U.P.A.), beth.dumont@jax.org (B.L.D.).

AUTHOR CONTRIBUTIONS

U.P.A. and B.L.D. were involved in conceptualization, methodology, data interpretation, and writing, reviewing, and editing the manuscript. U.P.A. performed the experiments, formal analysis, and visualization. B.A.S. contributed key reagents and reviewed and edited the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

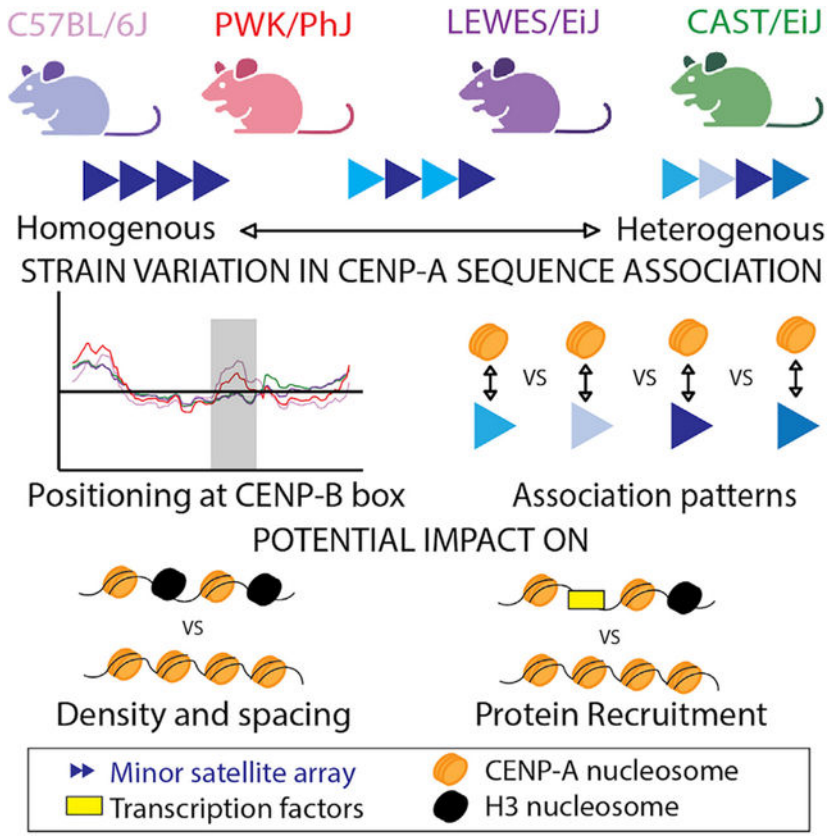
The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113178.

Centromeres are rapidly evolving and highly polymorphic, but the functional consequences of this variation are poorly understood. Arora et al. show that population variation at centromere satellite DNA leads to striking differences in the sequence association landscape of CENP-A, a kinetochore complex protein vital for centromere identity and stability.

Graphical Abstract



INTRODUCTION

Centromeres are satellite-rich chromatin domains that serve as sites for the assembly of the multiprotein kinetochore complex, which facilitates chromosome segregation during mitosis and meiosis.[1–4] Loss of centromere integrity and function can lead to apoptosis, chromosome mis-segregation, and widespread genome instability, phenomena linked to cancer and infertility.[5,6] Despite their essential biological roles, centromeres are highly variable in sequence and structure within and between species.[7–14] Rapid sequence-level evolution of centromeres imposes strong selection for cognate changes to DNA-associated kinetochore proteins. As a result, several kinetochore proteins are adaptively co-evolving with centromere satellite DNA.[15,16] The most well-established case involves CENP-A, an H3 histone variant that delimits the core centromere domain, helps maintain centromere repeat stability, and directs assembly of the kinetochore complex.[4,5,17–19] Although its homolog, histone H3, has been subject to intense purifying selection over evolutionary time and is highly conserved across taxa, CENP-A exhibits clear signals of adaptive evolution

within its DNA-associating NH_2-terminal tail.[7,15,16,20–22] The molecular arms race between CENP-A and centromere satellites is critical for the compatibility of this protein-DNA association and, by extension, the assembly and function of the kinetochore.

Recently, we used *k*-mer-based bioinformatic methods in conjunction with cytogenetic approaches to uncover remarkable variation in the size and sequence composition of centromeres across diverse inbred mouse strains and wild-caught house mice (*Mus musculus*).[10] In house mice, each centromere can be broadly categorized into two chromatin domains. The core functional centromere domain is composed of a single 120-base pair (bp) minor satellite repeat unit that is iterated in tandem over ~1 Mb of sequence and is responsible for facilitating CENP-A association and kinetochore assembly. The minor satellite array is flanked by an ~2-Mb pericentromeric chromatin domain defined by a focal 234-bp major satellite repeat monomer and the heterochromatin mark H3K9me3. While our prior work established that inbred mouse strains differ in the abundance of their resident minor satellite repeats and exhibit variable levels of minor satellite sequence heterogeneity, the potential functional consequences of this sequence diversity for the dynamics of centromeric chromatin and kinetochore assembly remain to be elucidated. Importantly, even subtle perturbations to kinetochore assembly or stability could have downstream impacts on the fidelity of chromosome segregation and propensity for centromere drive.[15,16,20,23–26]

Despite significant sequence variation in centromere repeats across house mice, CENP-A is relatively conserved within *Mus musculus*, with strains derived from the *M. m. domesticus* and *M. m. castaneus* subspecies harboring identical CENP-A amino acid sequences (Table S1). This conservation raises the question of whether CENP-A exhibits sequence-specific binding in *Mus musculus* strains with divergent centromere satellite DNA. Indeed, inbred mouse strains C57BL/6J and ZALENDE/EiJ have identical CENP-A amino acid sequences but distinct positioning along the minor satellite consensus sequence.[27] Similarly, prior work examining human centromeres suggests that CENP-A preferentially associates with centromere satellite repeat arrays with lower repeat diversity.[28] In maize, different alternating centromere satellite monomers position CENH3 (the plant ortholog of CENP-A) with translational and rotational phasing to ensure regular spacing of nucleosomes on each monomer.[29] Furthermore, in fission yeast, certain DNA sequence features have been found to promote CENP-A incorporation.[30] Thus, strain-specific organization of centromere satellite diversity could influence CENP-A density and modulate kinetochore complex density, with potential consequences for genome stability.[31]

The extensive variation in centromere satellites across inbred mouse strains provides a powerful experimental framework to explicitly test how centromere sequence variation influences CENP-A association.[10] To this end, we performed CENP-A chromatin immunoprecipitation followed by sequencing (chromatin immunoprecipitation sequencing [ChIP-seq]) across four diverse inbred mouse strains: C57BL/6J (reference strain, *M. m. domesticus*), LEWES/EiJ (*M. m. domesticus*), CAST/EiJ (*M. m. castaneus*), and PWK/PhJ (*M. m. musculus*). These strains capture the breadth of inbred mouse minor satellite size and heterogeneity and sample across the three principal *Mus musculus* subspecies. Further, three of these strains (C57BL/6J, LEWES/EiJ, and CAST/EiJ) possess identical CENP-A amino acid sequences, allowing us to isolate the influence of centromere DNA

sequence on protein association, independent of protein-level diversity (Table S1). PWK/PhJ, an inbred strain developed from wild-caught *M. m. musculus* mice, harbors three amino acid substitutions relative to the other strains (positions 19, 21, and 46), permitting simultaneous comparisons of how protein-level divergence influences CENP-A association. Using both consensus-guided and reference-independent computational approaches, we uncover divergent CENP-A sequence association landscapes between strains. Overall, our findings point to strain differences in functional centromere size, CENP-A density, and transcription factor association, hinting at potential strain differences in kinetochore protein assembly and chromosome segregation dynamics.

RESULTS

Validation of CENP-A ChIP in diverse inbred mouse strains

Mammalian centromere chromatin is defined by interspersed blocks of histone H3 and CENP-A, the centromere-specific histone H3 variant. To characterize strain differences in CENP-A sequence association, we performed CENP-A ChIP on MNase-digested chromatin from three diverse inbred mouse strains with identical CENP-A coding sequences (C57BL/6J, LEWES/EiJ, CAST/EiJ) and one inbred strain (PWK/PhJ) distinguished by three amino acid substitutions (Table S1; Figure 1A). In parallel, we chromatin immunoprecipitated DNA with antibodies targeting H3K4me3 (positive control for euchromatic genomic regions) and IgG (negative control) for each strain. After confirming the efficacy of immunoprecipitation in each inbred strain via qPCR (Figure S1A), we sequenced CENP-A ChIP DNA (three technical replicates for CAST/EiJ, LEWES/EiJ, and PWK/PhJ; two technical replicates for C57BL/6J; >30 million reads per replicate) along with corresponding MNase-digested input chromatin samples (Figure S1B). To evaluate the concordance between replicates, we quantified the frequency of all unique *31*-mers in each replicate CENP-A ChIP-seq dataset relative to the corresponding input sample and performed a principal component analysis. Overall, CENP-A-enriched *k*-mers are more similar within strains than between strains, attesting to the high experimental reproducibility across ChIP replicates (Figures S1C and S1D).

We next mapped ChIP-seq reads to the C57BL/6J-derived consensus minor satellite sequence allowing for up to 32 bp of mismatches, insertions, and deletions (Figure 1A). Prior work has showed that CENP-A primarily associates with the minor satellite, with negligible CENP-A association with the pericentromere-enriched major satellite.[1] As expected, a higher proportion of sequencing reads map to the minor satellite consensus sequence, as opposed to the major satellite consensus sequence (Figure 1B), in CENP-A ChIP normalized to input samples (Wilcox rank sum exact test $p = 2.83 \times 10^{-6}$). The percentage of input reads mapping to the minor satellite consensus sequence is a proxy for the size of the minor satellite array in each strain and, expectedly, aligns with earlier genomic and cytogenetic estimates of relative centromere array size in these mouse strains (PWK/PhJ mean = 5.16% of input reads > LEWES/EiJ mean = 0.58% > CAST/EiJ mean = 0.57% > C57BL/6J mean = 0.5%; Table S2).[10,32] These observed strain differences in the percentage of input reads that map to the minor satellite consensus are statistically significant (Kruskal-Wallis one-way ANOVA $p < 0.05$; Table S2). In contrast, the percentage

of CENP-A ChIP reads that mapped to the minor satellite consensus is not significantly different across strains (Kruskal-Wallis one-way ANOVA $p = 0.95$; Table S2). The apparent uniform enrichment of minor satellite-like sequences in CENP-A ChIP samples across strains suggests that the abundance of CENP-A-associated DNA does not necessarily scale with minor satellite DNA abundance. However, we acknowledge that strain differences in CENP-A ChIP efficiency could confound this interpretation.

Despite the gapped status of centromeres on the current mouse reference genome (mm39), we observed that all CENP-A ChIP-seq reads that mapped to the minor and major satellite consensus sequences also mapped to the reference genome (Table S2). Reads mapping to the minor satellite consensus sequence localize to discrete regions on chromosomes 2, 6, and X (Figure S2A), whereas reads that mapped to the major satellite consensus aggregate at loci on chromosomes 2 and 9 (Figure S2B). These unexpected findings suggest that multiple non-centromeric genomic regions contain sequences that resemble the minor and major satellite consensus sequences. With the exception of the region on chromosome 9 (which potentially lies within the pericentromeric region), none of these loci are included in current sets of blacklisted loci in the mouse genome.[33] Whether these regions represent assembly errors, non-functional occurrences of major and minor satellite DNA, or regions of the mouse genome that may have the capacity to form neocentromeres remains to be determined. In contrast, CENP-A ChIP reads that do not map to the minor or major satellite consensus sequences map with near uniformity across the reference genome and likely represent experimental noise (Figure S2C).

We next visualized the distribution of mismatches to the minor satellite consensus sequence in CENP-A ChIP relative to input samples across strains. Reads with no mismatches to the consensus minor satellite sequence are the most differentially enriched in CENP-A ChIP across strains, with the C57BL/6J reference strain exhibiting the highest enrichment of sequences that represent a perfect match to the consensus (F-statistic, $p < 0.05$). Nonetheless, the majority of CENP-A-enriched reads harbor multiple mismatches to the minor satellite consensus sequence (Figure S3A). To decipher the locations of polymorphic sites along the minor satellite consensus sequence, we computed the enrichment of each consensus nucleotide in CENP-A ChIP relative to input samples. Interestingly, strains vary in consensus nucleotide enrichment at the CENP-B box, with strains CAST/EiJ and LEWES/EiJ harboring an increased representation of non-consensus nucleotides in CENP-A ChIP relative to input (Figure 1C). CAST/EiJ and PWK/PhJ exhibit increased representation of consensus nucleotides in the 3′ region adjacent to the CENP-B box (Figure 1C).

Taken together, these consensus- and reference-based analyses reveal comparable efficacy of CENP-A ChIP in each of the four profiled strains, demonstrate high experimental reproducibility across ChIP replicates, and expose patterns of centromere satellite sequence enrichment in CENP-A-associated DNA.

Strain variation in the relative enrichment of CENP-A at pericentromeric chromatin

As with the minor satellite, the proportion of CENP-A ChIP reads mapping to the major satellite consensus sequence differs between strains (Figure 1B). In strains C57BL/6J and PWK/PhJ, there is no enrichment of reads mapping to the major satellite in CENP-A

ChIP compared with input samples (Figure 1B). This pattern accords with expectations, as CENP-A is thought to primarily associate with the minor satellite.[1] However, strains CAST/EiJ and LEWES/EiJ show an opposite pattern, with a slight enrichment of reads mapping to the major satellite relative to input samples (Kruskal-Wallis one-way ANOVA $p = 0.07$; Figure 1B). These intriguing findings add to a similar, prior observation in the inbred strain ZALENDE/EiJ[27] and suggest that CENP-A may be associated with the major satellite in some mouse strains. Thus, minor satellite array size might not be the only genetic factor influencing functional centromere size across mouse strains.

Strain differences in CENP-A positioning along the minor satellite consensus sequence

To investigate strain variation in CENP-A positioning, we profiled relative read mapping coverage along the minor satellite consensus sequence in CENP-A ChIP compared with input samples (Figure 2A). All strains share a common peak at minor satellite positions 1–25 bp, identifying this region as a universal CENP-A association sequence across strains. However, despite this commonality, strains exhibit distinct spatial patterns of CENP-A positioning (Figure 2B; Kruskal-Wallis; degrees of freedom [df] = 3, $p = 0.0019$). C57BL/6J and PWK/PhJ share a peak of association at positions 65–75. These latter sites overlap the CENP-B box, a 17-bp motif that confers sequence-specific binding of the kinetochore protein CENP-B.[34,35] CENP-B is a non-essential component of the kinetochore,[36] but it is thought to enhance stability of CENP-A/DNA association and play important roles in *de novo* centromere formation.[35,37]

CENP-A ChIP and input DNA were sequenced using 76-bp single-end reads, extending beyond the expected midpoint of a single ~130-bp CENP-A nucleosome. Thus, we expected to recover an artificial doubling of sequencing coverage centered on any site of preferential CENP-A association within the minor satellite (Figure 2A). To confirm this expected trend, we trimmed sequence reads to 60 bp (i.e., less than half the predicted length of a CENP-A nucleosome) and assessed the resulting impact on inferred CENP-A positioning at the CENP-B box. Trimming reads in this fashion resulted in the disappearance of the C57BL/6J and PWK/PhJ association peaks over the CENP-B box (Figure 2B), implying that CENP-A is indeed centrally positioned at the CENP-B box in those strains. In contrast, read trimming had less impact on CENP-A positioning in CAST/EiJ and LEWES/EiJ (Figure 2B).

We next asked whether strains with preferential CENP-A positioning at the CENP-B box (C57BL/6J and PWK/PhJ) simply had more CENP-B boxes in their CENP-A ChIP-seq reads. While the CENP-B box motif is more abundant in C57BL/6J than in other strains, PWK/PhJ has similar numbers of CENP-B box motifs as LEWES/EiJ and CAST/EiJ, the two strains with no peak of CENP-A association at this locus (Figure 2C; Dunn test pairwise comparison $p > 0.05$). Thus, variation in absolute CENP-B box frequency is not a unifying explanation for these strain patterns. Sequence variants at the CENP-B box could also modulate CENP-A association dynamics in a strain-specific manner. Consistent with this interpretation, we note that several consensus nucleotides within the CENP-B box are under-represented among mapped CENP-A ChIP reads from LEWES/EiJ and CAST/EiJ (Figure 1C). Our findings indicate that the CENP-B box does not universally centralize positioning of CENP-A in all inbred mouse strains.

Pairwise strain correlations among minor satellite CENP-A enrichment profiles indicate that strain pair CAST/EiJ and LEWES/EiJ and strain pair C57BL/6J and PWK/PhJ have more similar CENP-A positioning patterns (Figure 2D; C57BL/6J-PWK/PhJ correlation $r^2 = 0.870$, $p < 2.2 \times 10^{-16}$; CAST/EiJ-LEWES/EiJ correlation $r^2 = 0.962$, $p < 2.2 \times 10^{-16}$). This trend counters expectations based on overall strain relatedness. LEWES/EiJ and C57BL/6J share a common principal subspecies designation (*M. m. domesticus*), and *a priori* might be expected to exhibit a more highly conserved CENP-A association landscape than inter-subspecies comparisons. The absence of such a trend implies the action of rapid positional changes in CENP-A binding at centromeres. Taken together, our findings suggest that CENP-A associates with distinct sequence-specific contexts within the minor centromere satellite unit in different mouse strains.

CENP-A positioning along the major satellite consensus sequence shows higher strain similarity than observations with the minor satellite ($df = 3$, $p = 0.02214$; Figures 3B and 3C). However, we acknowledge that with a limited number of reads mapping to the major satellite, we may be underpowered to detect strain differences. Interestingly, the two strains that exhibit slight enrichment of major satellite sequences in CENP-A ChIP compared with input (CAST/EiJ and LEWES/EiJ; Figure 1B) also harbor discrete association peaks at positions 60–70 and 145–155 of the major satellite consensus sequence (Figure S3B). These patterns were not impacted by read trimming (to 60 bp) as the length of the major satellite sequence (234 bp) is more than twice the length of our sequenced reads (76 bp).

Most CENP-A-enriched *k*-mers are strain specific

The read mapping approach employed above offers insights into strain-level CENP-A-associated sequence variation in the context of the C57BL/6J-derived consensus sequence. This approach represents a facile and sensible analysis strategy in the absence of contiguous centromere assemblies for the mouse reference genome but is potentially biased by its reliance on a consensus sequence derived from a single inbred strain. To address this limitation, we employed a complementary *k*-mer-based approach to investigate strain differences in CENP-A sequence association in a manner agnostic to the consensus centromere satellite sequence.[38] First, we counted the frequency of all unique *31*-mers observed in each replicate CENP-A ChIP and input sample pair. Next, for each replicate, we quantified the enrichment of each *31*-mer as the ratio of read count normalized *31*-mer frequency in CENP-A ChIP relative to input (“enrichment score”; Figure S4A). Lastly, for each *31*-mer, we averaged the CENP-A ChIP/input enrichment score across replicates and extracted the 0.1% most enriched *31*-mers for each strain. This approach enables unbiased selection of the most CENP-A-enriched *31*-mers in each strain.

We compared sets of CENP-A-enriched *31*-mers across strains and observed a small, but significantly more than expected, number of shared *k*-mers ($p < 1 \times 10^{-6}$; Figure 3A). This finding reveals some underlying conservation of CENP-A-associated sequences between diverse strains. However, despite this backdrop of conservation, the vast majority of enriched *31*-mers are unique to a single strain (78.8% for C57BL/6J; 76.1% for CAST/EiJ; 82.6% for LEWES/EiJ; 72.4% for PWK/PhJ). Qualitatively similar results are obtained when considering *15*-mers, with a minority of *15*-mers shared across all strains (27.7% of *15*-mers

shared across all strains, with 22.9%–26.6% of *15*-mers unique to a specific strain; Figure S4C). Thus, diverse mouse strains harbor libraries of CENP-A-associated sequences that feature many strain-specific sequences.

We next compared the set of CENP-A-enriched *31*-mers for each strain with the consensus minor satellite sequence. Remarkably, none of these enriched *31*-mers present a perfect match to the minor satellite consensus sequence, with most harboring two or more sequence mismatches from the consensus (Figure 3B). Additionally, these *31*-mers align to various parts of the minor satellite consensus sequence in a strain-specific pattern (Figure 3C), although our single-end sequencing approach may lead to the underestimation of *31*-mer diversity in the middle of the minor satellite sequence. We conclude that, while the house mouse consensus sequence serves as a useful tool, it does not comprehensively represent sequences that preferentially associate with CENP-A in any inbred mouse strain, including the C57BL/6J reference strain.

Reference-independent identification and analysis of strain-specific CENP-A-associated sequence landscapes

To analyze strain-specific CENP-A-associated sequences in a consensus-free manner, we first identified the most sequence-abundant and *31*-mer-enriched CENP-A-associated sequences in each strain (Figure S4A). Briefly, we assigned each CENP-A ChIP-seq read a score based on the number of enriched *31*-mers within the read (top 0.1% enriched *31*-mers; read score) and the normalized abundance of the read in the CENP-A ChIP data (read count). We then jointly sorted reads by these two criteria and focused on the 1,000 top-scoring reads (Figure S4B).

We analyzed the relationship among the 1,000 most strongly CENP-A-associated sequences identified in each strain using a neighbor joining tree (Figure 4). CENP-A-associated sequences cluster into seven broad clades, with multiple sequences from each strain clustering in each clade (Figure 4; Table S3). In some clades, there is near-equal representation among strains (clades 2, 5, 6), whereas the remaining clades (1, 3, 4, 7) exhibit disproportionate representation of CENP-A sequences from specific strains (Figure 4; Chi-square test, $p < 0.05$). Most clades are distinguished from their sister clades by long branches, implying significant divergence between sequences in distinct clades. Overall, our analyses identify seven distinct classes of CENP-A-associated sequences that are present in all strains, although the functional distinctions between these sequence groups, if any, remains to be determined. Additional work is also required to determine whether sequences from specific clades are disproportionately represented on individual chromosomes.

We derived clade-specific consensus sequences to capture the sequence diversity represented in these seven groups (Table S3). We then used these consensus sequences, along with the canonical centromere major and minor satellite consensus sequences, as scaffolds for mapping CENP-A ChIP and input reads. This expanded consensus approach increases the number of mapped CENP-A ChIP-seq and input reads relative to isolated mapping against the reference-based consensus (Table S4). The proportion of additionally mapped CENP-A ChIP reads for each sample exceeds that of the corresponding input (Table S4), demonstrating that our reference-independent strategy for identifying CENP-A-enriched

reads provides a more comprehensive representation of the centromere sequence diversity across strains.

CENP-A-associated sequences exhibit phylogenetic similarities across strains

We next employed a computational stylistics approach to evaluate the text-based similarity among each strain's set of 1,000 most CENP-A-enriched sequences. Results from this analysis are summarized in a two-dimensional principal components analysis (PCA) plot (Figure 5). C57BL/6J and LEWES/EiJ, two *M. musculus domesticus* strains, share high similarity in their CENP-A-associated sequences. CAST/EiJ and PWK/PhJ have distinct CENP-A-associated sequence libraries that partition these strains along unique coordinates in PC space (Figure 5). These strain differences are not observed in random subsamples of CENP-A ChIP and input reads (see STAR Methods; Figure S5), implying that observed relationships are not due to chance. While our strain sample size is small, our findings are consistent with subspecies level divergence in CENP-A-associated centromere satellite sequences. This patterning stands in contrast to the absence of an evolutionary signal in measures of overall centromere sequence and architecture,[10] and it could suggest that the CENP-A-associated sequence landscape is more slowly evolving than the underlying sequence of the centromere itself.

Strain differences in transcription factor motif presence in CENP-A-associated sequences

Centromere transcription is essential for differentiation and development, and it is tightly regulated in both timing and rate.[39] For example, centromere-derived RNAs are crucial for localizing HJURP, a CENP-A chaperone that guides CENP-A incorporation into nucleosomes to specify centromere identity.[40] Thus, centromere transcription plays critical roles in early kinetochore assembly, including CENP-A recruitment.[39] Centromere transcription is carried out through RNA polymerase II and is initiated through DNA sequence-specific binding of a transcription factor.[41,42] However, the specific transcription factors (TFs) that act at centromeres to regulate their transcriptional activity remain largely unknown.[43] We reasoned that strain differences in CENP-A sequence association could lead to differences in the TF binding motifs found at the functional centromere core and potentially strain differences in centromere transcription dynamics.

To explore this possibility, we used MEME-ChIP to identify TF-binding motifs in strain-specific CENP-A-enriched sequences. We initially uncovered 105 TF motifs present in at least one strain. To rule out potentially spurious TF motifs, we ran MEME-ChIP on 10 sets of randomly sampled input reads and eliminated seven TFs that were significantly enriched in any set. Of the remaining 98 TF motifs, only 14 were present among the CENP-A-enriched sequences from all strains (ATF3, BATF, CUX1, DLX5, FOSB, HNF1A, HNF1B, HSF2, JUN, JUNB, LHX3, LHX6, PRGR, and SOX2), although strains differ in the frequency of these TF motifs (Figure S6A). Overall, the frequency and presence of TF motifs is highly variable across the suite of CENP-A-associated sequences in each strain.

We next sought to determine whether the TFs of the enriched motifs identified above associate with centromere DNA in experimental data. To this end, we identified 14 TFs with publicly available ChIP-seq datasets (selection criteria described in STAR Methods) and

mapped reads from both ChIP and input samples to the canonical minor satellite consensus sequence. Putative centromere-associated proteins were defined as those exhibiting an enrichment of mapped reads localizing to the minor satellite consensus in TF ChIP compared with the corresponding input sample (i.e., ChIP/Input >1). As expected, we found evidence of histone H3 and histone modification H3K9me3 at minor centromere satellite DNA (Figure S6B), validating our methodology.[44] Four TFs are generally more abundant at centromere satellite DNA than corresponding input samples: HNF1A, SMAD3, VDR, and SOX2 (Figure 6A). However, this enrichment is not significant (Sign test; $p > 0.05$), and there is considerable variability among experiments that is likely due to differences in profiled cell types and experimental conditions. These limitations aside, we note that SMAD3 motifs are differentially prevalent in the top 1,000 CENP-A-associated sequences across our four strains, with the motif altogether absent in CAST/EiJ CENP-A-associated sequences (Figure 6B). Although speculative, the absence and variability of different TF motifs in the CENP-A-associated sequences of these strains could lead to strain differences in centromere transcription dynamics.

DISCUSSION

Rapid centromere sequence evolution is theorized to impose complementary selection pressures on centromere-associated kinetochore proteins, leading to species-level centromere satellite and kinetochore protein co-evolution. However, the related possibility that within-species genetic diversity at the centromere impacts kinetochore protein association remains largely unaddressed. Our recent work exposed substantial polymorphism in centromere satellite sequences between inbred house mouse strains.[10] In this study, we quantitatively assess the effect of this centromere satellite diversity on the sequence association of one key kinetochore protein—the centromere-specific histone variant, CENP-A—using CENP-A ChIP-seq in four diverse inbred mouse strains.

CENP-A is adaptively co-evolving with centromere satellite DNA in some taxa, but three of our four surveyed mouse strains share an identical CENP-A amino acid sequence (Table S1). Thus, our experimental framework directly addresses how centromere DNA diversity impacts the association of CENP-A, independent of protein-level variation. The fourth strain we profiled, PWK/PhJ, harbors three amino acid substitutions relative to other strains. Two of these substitutions (positions 19 and 21) fall in a region with polar residue composition bias and replace non-polar amino acids with polar amino acids, thereby increasing the number of polar residues in this domain. The third amino acid substitution in PWK/PhJ is a lysine to arginine substitution at position 46 and resides in a region of CENP-A that is important for flexible DNA ends on the nucleosome. The functional consequences, if any, of these amino acid substitutions are unclear. Our findings do not expose PWK/PhJ as a broad outlier for CENP-A association or localization, suggesting that CENP-A divergence has limited influence on protein/DNA association in this system.

We uncover pronounced strain differences in the positioning of CENP-A along the minor satellite sequence that comprises the functional centromere core (Figure 2B). CENP-A positioning along the minor satellite trended with minor satellite repeat diversity,[10] with strains harboring reduced minor satellite repeat heterogeneity (C57BL/6J and PWK/PhJ)

exhibiting more similar association profiles than strains with higher minor satellite repeat diversity (CAST/EiJ and LEWES/EiJ). Prior research has suggested that centromere satellites composed of different repeat structures have variable stability and competence for centromere function.[28] Our findings reinforce the possibility that centromere repeat diversity influences CENP-A spacing and density, properties that could, in turn, have consequences for the architecture and stability of the kinetochore complex.[31]

Canonically, CENP-A is thought to primarily associate with the minor satellite array located in the functional centromere core. The minor satellite array is therefore predicted to be the primary determinant of centromere size in house mice. In contrast to this prevailing model, we find enrichment of CENP-A in the pericentromeric major satellite in inbred strains CAST/EiJ and LEWES/EiJ (Figure 1B). Our findings add to a prior report of CENP-A enrichment at the major satellite in ZALENDE/EiJ, a wild-derived inbred strain harboring multiple Robertsonian fusions and unusually short minor satellite arrays.[27] Together, these observations reveal a possible role of the major satellite array in the regulation of centromere size and imply that functional centromere size may not always be strictly proportional to the length of the minor satellite array. The major satellite array is known to recruit microtubule destabilizers, which enable stronger centromeres to reorient toward the oocyte during female meiosis.[23] Evidence of CENP-A association with the major satellite brings into question its influence on microtubule destabilizer recruitment and the propensity for centromere drive. Given the established relationship between centromere size and centromere drive potential,[27,45] our results raise the possibility that strain differences in CENP-A association with pericentromeric chromatin could lead to strain-dependent differences in chromosome transmission dynamics.

CENP-A association was previously shown to center on the 17-bp CENP-B motif in the C57BL/6J and ZALENDE/EiJ inbred mouse strains.[27] We replicate this earlier finding for C57BL/6J and further show that CENP-A binding localizes to the CENP-B box in PWK/PhJ (Figure 2A). However, two other profiled strains in our study—LEWES/EiJ and CAST/EiJ—show no evidence of preferential association at this locus. CENP-B is a constitutively expressed and conserved DNA binding protein but, paradoxically, is not required for chromosome segregation.[36] Despite its non-essentiality, CENP-B has multiple centromere-associated functions, including the recruitment and stabilization of CENP-A, regulation of centromeric heterochromatin, and *de novo* centromere formation.[46] We show that these strain differences in CENP-A localization to the CENP-B box are not strictly due to variation in the relative abundance of CENP-B boxes across strains (Figure 2C), suggesting that they reflect genuine underlying strain differences in CENP-A affinity for the CENP-B motif. Whether this differential affinity is, in turn, mediated by strain differences in CENP-B binding at the CENP-B motif remains unknown. Regardless, our findings indicate that the CENP-B box does not centralize the positioning of CENP-A in all inbred mouse strains.

CENP-A localization and integration into centromeric histones is dependent on RNA polymerase II-mediated centromere transcription.[40,42] We report evidence for strain differences in TF motif presence at CENP-A-associated sequences, suggesting strain variation in centromere transcriptional response. As a notable example, we find SMAD3

binding sites in C57BL/6J, LEWES/EiJ, and PWK/PhJ, but not CAST/EiJ, CENP-A-associated sequences. Through re-analysis of published SMAD3 ChIP-seq data from C57BL/6J, we show that SMAD3 binding enrichment is limited to the minor satellite (to the exclusion of the major satellite) across multiple cell types. SMAD3 is a TF that regulates cellular proliferation in response to extracellular cues,[47] particularly in the context of wound healing and cancer. Intriguingly, prior studies have established that centromere transcription plays key roles in cellular proliferation and differentiation,[39] processes that are integral to wound healing and often dysregulated in cancer. The variability of SMAD3 motif abundance at CENP-A-associated sequences across strains (including its absence in CAST/EiJ) could contribute to strain differences in regenerative potential and cancer progression. Thus, strain variation in centromere-associated TF motif abundance may provide a mechanism for individual differences in genome stability, a prospect that merits further investigation.

Our work also pioneers the application of powerful reference-independent strategies for probing the chromatin and TF landscape of repetitive sequences refractory to analysis with conventional methods. Our *k*-mer-based, consensus-independent method for identifying CENP-A-associated sequences provides a more comprehensive snapshot of strain similarities and differences in the CENP-A sequence association landscape than analyses based on a single consensus sequence (Table S4). Application of the methods presented here to other ChIP-seq datasets will enhance understanding of the chromatin environment of centromeres, as well as other regions of the genome that are currently missing from the mouse reference assembly. Such analyses will yield a more holistic picture of chromatin regulation across the genome and offer functional insights into the most intractable genomic regions. Additionally, emerging applied applications of long-read sequencing technologies, such as DiMeLo-seq, stand to provide comprehensive maps of long single DNA molecule-protein interactions, including the dynamics of CENP-A-centromere DNA associations.[48]

The centromere paradox asserts that centromere-associated kinetochore proteins, like CENP-A, are locked in a co-evolutionary arms race with rapidly evolving centromere satellite DNA. However, a key underlying requirement for the centromere paradox to hold is that kinetochore proteins like CENP-A must exhibit variable sequence affinity for distinct satellite sequence variants. Whether centromere DNA sequence impacts function has been the focus of rigorous prior debate.[49,50] Three of our four profiled strains share a common CENP-A protein sequence but harbor highly variable centromere sequences and architectures. We observe clear differences in the localization of CENP-A across these strains, demonstrating that, at least in house mice, CENP-A appears to harbor some sequence promiscuity, with unique satellite sequence variants favored in different strain backgrounds. Our investigation demonstrates sequence-dependent differences in kinetochore protein binding and provides direct support for an important assumption of the centromere paradox. However, the extent to which the distinct strain-level CENP-A association pattern documented here provides a molecular readout for other measures of centromere function, including genome stability, remains an open question with crucial relevance for the genetic etiology of cancer and infertility.

Limitations of the study

Our work carries several technical and analytical limitations. First, alternative chromatin profiling methods like CUT&Tag may increase the signal-to-noise ratio and serve as more powerful alternatives to the ChIP-seq assays used in our investigations. These trade-offs could be especially advantageous for CENP-A, a protein that is present at very low levels across the genome. Second, our use of single-end sequencing reads, rather than paired-end reads, resulted in an artificial doubling of coverage across portions of CENP-A nucleosomes. While this phenomenon offers a signal of CENP-A positioning along the minor satellite, we recognize that paired-end reads would have allowed us to define full, contiguous minor satellite sequence repeats associated with CENP-A. Third, the absence of centromere sequences from the mouse reference genome assembly required the development and application of reference-independent analysis methods. While our approach is adaptable to other repetitive loci, we acknowledge that our methodology is ad hoc and not governed by established guidelines for data analysis. Finally, genetic variation is not restricted to the centromere in our profiled mouse strains, and we cannot rule out the possibility that factors independent of centromere DNA influence differences in CENP-A sequence association across strains. Future studies could explicitly test centromere DNA sequence changes across a controlled genetic background using directed engineering of centromere satellite DNA or by harnessing differences between chromosomes within a mouse strain.

STAR★METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Beth L. Dumont (Beth.Dumont@jax.org).

Materials availability—This study did not generate new unique reagents.

Data and code availability

- Raw data files for the ChIP-sequencing experiment have been deposited in the NCBI Sequence Read Archive (SRA) under project accession number PRJNA838487.
- Custom scripts used for the analysis are available on Zenodo: https://doi.org/10.5281/zenodo.8303213.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Mice—C57BL/6J (stock no 000664), LEWES/EiJ (stock no 002798), CAST/EiJ (stock no 000928), and PWK/PhJ (stock no 003715) mice were obtained from The Jackson Laboratory. Sex breakdown for samples were as follows – 1 male and 1 female C57BL/6J, 1 male and 2 female CAST/EiJ, 1 male and 2 female LEWES/EiJ, and 1 male and 2 female PWK/PhJ.

Mice were housed in a low barrier room and provided food and water *ad libitum*. Mice were euthanized by CO_2 asphyxiation or cervical dislocation in accordance with recommendations from the American Veterinary Medical Association. All animal experiments were approved by the Institutional Animal Care and Use Committee at The Jackson Laboratory under Animal Use Summary #17021 and were consistent with the National Institute of Health guidelines.

METHOD DETAILS

Strain sequence variation in CENP-A—We used publicly available whole genome sequences from the Sanger Mouse Genomes Project and ENSEMBL SNP effect information to ascertain amino acid altering variants within the dominant *Cenpa* mRNA transcript (NR_126074). Variant call format (VCF) files produced against mm39 were accessed from https://www.mousegenomes.org/snps-indels/and subset to include only variants within the *Cenpa* locus. The resulting file was then manually scanned for coding sequence variants in CAST/EiJ, LEWES/EiJ, and PWK/PhJ. The fourth strain profiled in our study, C57BL/6J, is the reference genome strain.

Chromatin extraction—Chromatin was extracted from two or three flash-frozen mouse livers per CENP-A ChIP as previously described.[27] Livers were homogenized in 4 mL ice-cold Buffer I (0.32 M Sucrose, 60 mM KCl, 15 mM NaCl, 15 mM Tris-Cl pH 7.5, 5 mM $MgCl_2$, 0.1 mM EGTA, 0.5 mM DTT, 0.1 mM PMSF, 50 μL PIC (Sigma P8340)) per g of tissue. The homogenate was filtered through a 100 mm cell strainer (Fisher) and centrifuged at 6000g for 10 min at 4°C. The pellet was first resuspended in the same volume of Buffer I, followed by an equivalent volume of ice-cold Buffer I supplemented with 0.2% NP-40 alternative (Sigma). Samples were incubated on ice for 10 min to release nuclei. 4mL of nuclei were gently laid upon 16 mL of ice-cold Buffer III (1.2 M Sucrose, 60 mM KCl, 15 mM NaCl, 15 mM Tris-Cl pH 7.5, 5 mM $MgCl_2$, 0.1 mM EGTA, 0.5 mM DTT, 0.1 mM PMSF, 50 μL PIC (Sigma P8340)) in a 50 mL conical tube. Samples were centrifuged at 10,000g for 20 min at 4°C. The supernatant was discarded, and the pellet was resuspended in 26 μL MNase buffer (50 mM Tris, 1 mM $CaCl_2$, 4 mM $MgCl_2$, 4% NP-40) supplemented with 1 mM PMSF and 1:100 dilution of PIC (Sigma P8340) per 10×10^6 cells. Cell number was calculated invoking the assumption that 1 g of mouse liver tissue is equivalent to 125 million cells.[51] MNase (100U/uL Thermo Fisher Scientific PI88216) was added at a concentration of $1uL/33.3\times10^6$ cells and samples were incubated at 37°C for 12 min. MNase digestion was stopped by adding 0.5M EDTA to a final concentration of 10 mM. Samples were incubated on ice for 5 min and then spun at 15000 rpm for 10 min. The supernatant was transferred to a new tube and spun again at 15000 rpm for 10 min. Finally, the supernatant was transferred to a new tube and either stored at −20°C or carried forward into chromatin immunoprecipitation.

Chromatin immunoprecipitation—Chromatin immunoprecipitation was performed using an antibody against CENP-A (D601AP; rabbit anti-mouse targeting synthetic peptide KPQTPRRRPSSPAPGPSRQSSSVGSC found from amino acid position 7 to 32 in the N terminus of mouse CENP-A developed for ChIP by Dr. Beth Sullivan, Duke University), IgG (Millipore Cat. # 12–370), and H3K4me3 (Millipore Cat. # 07–473).[27,52] 25 μL

Protein G Dynabeads (Invitrogen Cat. #10003D) were used per ChIP reaction. Beads were washed two times with RBD (RIPA buffer (Sigma Cat. #R0278), 50 mg/mL Bovine Serum Albumin (BSA), and 0.5 mg/mL Herring Sperm DNA). Beads were resuspended in RBD, combined with antibody, and then allowed to conjugate at room temperature for 20 min with gentle rotation. Again, beads were washed two times in RBD and resuspended in RBD supplemented with 1 mM PMSF and 1:100 dilution of PIC (Sigma P8340). Chromatin was added to tubes and allowed to incubate at 4°C with rotation overnight. A 10x chromatin volume was used for CENP-A ChIP and a 2x chromatin volume was used for H3K4me3 and IgG ChIP compared to input chromatin. Beads were then washed three times with RIPA buffer and 50 mg/mL BSA. Beads were subsequently washed three times with TE pH 8.0 and transferred to a new tube. Elution Buffer (1% SDS, 20 mM Tris-HCl pH 8.0, 200 mM NaCl, 5 mM EDTA) supplemented with Proteinase K (New England Biolabs Cat. #P8107S) was added to the beads for both ChIP and input chromatin. Samples were incubated at 68°C overnight with vigorous shaking in a thermomixer. Chromatin from ChIP samples was recovered from beads using a magnet and placed in a clean tube. Samples were processed using the GeneJET PCR purification kit and eluted in 50 μL of 10 mM Tris-HCl pH 8.0. DNA concentration was measured using a Qubit dsDNA High Sensitivity kit (Thermo Fisher Scientific Cat. #Q32854) according to the manufacturer's instructions.

qPCR validation of ChIP—Quantitative PCR (qPCR) was performed on all replicate samples from CENP-A, H3K4me3, IgG ChIP, and input experiments. Reactions were run with PowerUp SYBR Green 2x master mix (Thermo Fisher Scientific Cat no A25742) and the following primers: ActB Promoter F primer 5′-GCCATAAAAGGCAACTTTCG-3′, ActB Promoter R primer 5′-TTTCAAAAGGAGGGGAGAGG-3′, Minor Satellite qPCR F primer 5′-CATGGAAAATGATA AAAACC-3′, Minor Satellite qPCR R primer 5' -CATCTAATATGTTCTACAGTGTGG-3'. PCR was carried out for 40 cycles on a ViiA 7 real-time PCR system (Thermo Fisher Scientific). We measured the relative cycle number (determined by automated threshold analysis by the machine) for ChIP compared to input samples for each ChIP reaction and primer pair.

Library preparation and sequencing—ChIP libraries were constructed using the KAPA HyperPrep Kit (Roche Sequencing and Life Science) according to the manufacturer's protocols. Briefly, the protocol entails ligating Illumina specific barcoded adapters, size selection, and PCR amplification. The quality and concentration of the libraries were assessed using the High Sensitivity D5000 ScreenTape (Agilent Technologies) and KAPA Library Quantification Kit (Roche Sequencing and Life Science), respectively, according to the manufacturers' instructions. Libraries were sequenced using 75 bp single-end reads on an Illumina NextSeq 500 using the High Output Reagent Kit v2.5. Each CENP-A ChIP and input sample was sequenced to a minimum of 30 million reads, in excess of the 20 million minimum reads recommended by ENCODE.[53]

Centromere consensus and mm39 reference genome read mapping analysis —We used fastp (version 0.23.1) for preprocessing of fastq files to filter low quality reads and trim adapter sequences.[54] Reads were then mapped to three tandem copies of the major and minor satellite centromere consensus sequences derived from C57BL/6J[55] and the

mouse reference genome (mm39) using bwa version 0.7.9.[56] Mapping was performed with the default bwa mem parameters (k = 19, w = 100, d = 100, r = 1.5, c = 10000, A = 1, B = 4, O = 6, E = 1, L = 5, U = 9, T = 30, v = 3), which allows up to 32 bp of mismatches, insertions, and deletions. The percentage of reads that mapped to each consensus sequence or the reference genome was calculated from output of the idxstats command in samtools (version 1.8) and the total number of reads in the fastq file.[57] Enrichment at centromeres was quantified relative to input and normalized to the number of sequenced reads. To visualize mapping of CENP-A ChIP and input reads across the genome, we plotted per base pair read depth along the genome in R (version 4.0.5). Uniformity of CENP-A ChIP reads that map in the genome was assessed with a Kolmogorov-Smirnov test using the cumulative mean coverage across positions as the cumulative distribution function (Table S6). All CENP-A ChIP-seq replicates had a D value <0.5, and a p value $<2.2\times10^{-16}$ (Table S6). Blacklisted regions of the mm10 reference genome were retrieved from https://github.com/Boyle-Lab/Blacklist/tree/master/lists and converted to mm39 coordinates using liftover.

To summarize the sequence polymorphism landscape across the minor satellite repeats, we used the sequencing reads, in conjunction with their mapped positions across the minor satellite consensus sequences, to derive a vector of relative nucleotide probabilities for each position in the satellite consensus sequence. For reads that mapped in the forward orientation, for a given position, we computed the total frequency of reads with an “A”, “C”, “G”, or “T” at the focal position. These per-nucleotide frequencies were then converted to relative probabilities summing to one and used to populate a 4xN“polymorphism matrix” for each analyzed sample, where N = 120 for the minor satellite sequence. We then compared the percentage of non-consensus nucleotides for each strain across the minor consensus satellite sequence in CENP-A ChIP compared to input. A heatmap was constructed with package pheatmap (version 1.0.12) in R (version 4.0.5).

***K*-mer tables and strain-specific enriched *k*-mers—**Fastp-filtered reads were processed through clumpify (v37.44; https://sourceforge.net/projects/bbmap) to remove optical duplicates. This step ensured that *k*-mer counts were not skewed by PCR-related artifacts. Each sequenced read in a sample’s fastq file was then decomposed into its constituent nucleotide words of length *k*, or *k*-mers, using a custom Python script (KmerComposition.py). We set k = 31 to balance the competing demands of computational resource usage and sequence specificity. Several investigations were repeated with k = 15 to demonstrate robustness of conclusions to an arbitrarily selected value of *k*.

k-mer frequencies were normalized to the number of reads. We then calculated an “enrichment score” for each *k*-mer as the ratio of the normalized *k*-mer frequency in CENP-A ChIP compared to input samples. The enrichment scores of read count normalized *k*-mers were used to assess the distance between replicate and strain samples using the prcomp function in R (version 4.0.5). For each strain, *k*-mers were then ranked by the enrichment score to identify the top 0.1% most enriched *k*-mers. We used 0.1% as the cutoff to ensure selection of an optimal number of *k*-mers that capture both strain-specific features and *k*-mers shared across strains.

To identify centromere consensus-derived *k*-mers, we mapped each strain's top 0.1% CENP-A ChIP enriched *k*-mers to the consensus minor satellite sequence using bwa version 0.7.9 and allowing up to 4 mismatches.

Scoring CENP-A ChIP-seq reads for enrichment of *k*-mers and abundance— To prioritize a list of highly enriched CENP-A associated sequences in each strain, we assigned two numerical scores to each CENP-A ChIP read. First, we tallied the number of 0.1% CENP-A ChIP/input enriched *k*-mers observed in a given sequencing read from that strain ('read score'). We then counted the frequency of each 76-bp read in each library, normalizing to the total number of sequencing reads ('read count'). We then jointly ranked sequences by their read score and read count and focused on the top 1000 ranking reads.

Analyzing strain-specific sequence groups—To visualize the relationship between sets of CENP-A enriched sequences across strains, we constructed a neighbor joining tree using MEGA11 and analyzed subclusters of sequences with the package ape (version 5.6–2) in R (version 4.0.5).

To compare different sets of sequences across strains, we performed a computer-assisted text analysis using the package stylo (version 0.7.4) in R (version 4.0.5). Relatedness among groups of words (in our case, strain-specific top 1000 ranked 76-bp reads) was assessed via principal component analysis. To assess significance and robustness of observed strain-level relationships, we randomly sampled 1000 sequences from each sample's input and CENP-A ChIP library and assessed the classic delta distance (as developed by Burrows) between replicate pairs. The delta distances were measured from a matrix of sequence frequencies. We then identified where the distance between samples of the CENP-A enriched reads fell in comparison to the 1000 randomly sampled delta distances of CENP-A ChIP and input replicates. The distance of the samples in the observed CENP-A enriched sequences are clear outliers compared to the distributions of randomly sampled CENP-A ChIP and input replicates (Figure S5), indicating that our observations are biological and not due to chance.

Clade consensus read mapping analysis—We derived clade-specific consensus sequences from the neighbor joining tree of CENP-A enriched sequences using Clustal Omega[58] (Table S3). Reads were mapped to the clade consensus sequences and three tandem copies of the major and minor satellite centromere consensus sequences derived from C57BL/6J[55] using bwa version 0.7.9.[56] Mapping was performed with the default bwa mem parameters (k = 19, w = 100, d = 100, r = 1.5, c = 10000, A = 1, B = 4, O = 6, E = 1, L = 5, U = 9, T = 30, v = 3). The percentage of reads that mapped to each consensus sequence was calculated from output of the idxstats command in samtools (version 1.8) and the total number of reads in the fastq file.[57]

MEME-ChIP motif enrichment analysis—We used the MEME-ChIP tool in the MEME suite (v5.4.1) to identify transcription factor motifs enriched in CENP-A associated sequences. MEME-ChIP performs comprehensive motif analysis on a set of sequences, assuming that motifs are centrally located within the input sequence. MEME-ChIP was run on the top 1000 most CENP-A enriched sequences in each strain with the following command: "meme-chip -oc. -time 240 -ccut 100 -fdesc description -dna -order 2 -minw

6 -maxw 15 -db db/MOUSE/HOCOMOCOv11_core_MOUSE_mono_meme_format.meme -meme-mod zoops -meme-nmotifs 3 -meme-searchsize 100000 -streme-pvt 0.05 -streme-totallength 4000000 -centrimo-score 5.0 -centrimo-ethresh 10.0 STRAIN_top1000.fa"

Transcription factor association analysis—BioProjects containing ChIP-seq data from 64 transcription factors (TFs) with motif occurrence in C57BL/6J CENP-A enriched reads were selected for analysis (Table S5). These BioProjects contained ChIP-seq data for TFs and various histone modifications. We restricted our focus to datasets that: (1) had both ChIP and input sequencing samples to enable comparisons of ChIP/input relative TF enrichment for each independent experiment and (2) featured ChIP-seq data from more than one experiment (segregated by BioProject, cell type, treatment, genotype, and strain). Sequencing data were obtained in fastq format from the NCBI Sequence Read Archive (SRA). Reads from each sample were mapped to the major and minor satellite consensus sequences using bwa (default parameters, version 0.7.9[56]). The percentage of reads that mapped to each consensus sequence was calculated from the output of the idxstats command in samtools (version 1.8) and the total number of reads in the fastq file.[57] The data for each TF and histone protein was represented by the ratio of the percentage of reads mapped in ChIP/input for each experiment. Sample information was annotated using the metadata from SRA for each BioProject. Significance of ChIP/Input enrichment of a TF was determined with a Sign test in package BSDA (version 1.2.1) in R (version 4.0.5).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests and methods were conducted as described in the STAR Methods detail sections. Analyses were performed in R (version 4.0.5) using $p < 0.05$ as the cutoff for declaring statistical significance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We gratefully acknowledge the contribution of the Genome Technologies Service at The Jackson Laboratory for library preparation and sequencing. We thank Dr. Christopher Baker and Catrina Spruce for protocols and guidance on chromatin immunoprecipitation and data analysis. We also thank members of the Dumont Lab, and we thank Drs. Mary Ann Handel and Christopher Baker for comments on the manuscript.

This work was funded by a Maximizing Investigators' Research Award from the National Institute of General Medical Sciences to B.L.D. (R35 GM133415). U.P.A. was supported by a Ruth L. Kirschtein Predoctoral Individual Fellowship from the National Cancer Institute (F31CA268727). CENP-A antibody production and affinity purification was supported through NIH grants R01 GM124041 and R01 GM129263 awarded to B.A.S. The content of this manuscript is the sole responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as a gender minority in their field of research.

REFERENCES

1. McKinley KL, and Cheeseman IM (2016). The molecular basis for centromere identity and function. Nat. Rev. Mol. Cell Biol. 17, 16–29. 10.1038/nrm.2015.5. [PubMed: 26601620]
2. Bakhoum SF, Thompson SL, Manning AL, and Compton DA (2009). Genome stability is ensured by temporal control of kinetochore-microtubule dynamics. Nat. Cell Biol. 11, 27–35. 10.1038/ncb1809. [PubMed: 19060894]
3. Schalch T, and Steiner FA (2017). Structure of centromere chromatin: from nucleosome to chromosomal architecture. Chromosoma 126, 443–455. 10.1007/s00412-016-0620-7. [PubMed: 27858158]
4. Fukagawa T, and Earnshaw WC (2014). The centromere: chromatin foundation for the kinetochore machinery. Dev. Cell 30, 496–508. 10.1016/j.devcel.2014.08.016. [PubMed: 25203206]
5. Giunta S, and Funabiki H (2017). Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. Proc. Natl. Acad. Sci. USA 114, 1928–1933. 10.1073/pnas.1615133114. [PubMed: 28167779]
6. Shrestha RL, Rossi A, Wangsa D, Hogan AK, Zaldana KS, Suva E, Chung YJ, Sanders CL, Difilippantonio S, Karpova TS, et al. (2021). CENP-A overexpression promotes aneuploidy with karyotypic heterogeneity. J. Cell Biol. 220, e202007195. 10.1083/jcb.202007195. [PubMed: 33620383]
7. Malik HS, and Henikoff S (2009). Major evolutionary transitions in centromere complexity. Cell 138, 1067–1082. 10.1016/j.cell.2009.08.036. [PubMed: 19766562]
8. Ventura M, Antonacci F, Cardone MF, Stanyon R, D'Addabbo P, Cellamare A, Sprague LJ, Eichler EE, Archidiacono N, and Rocchi M (2007). Evolutionary formation of new centromeres in macaque. Science 316, 243–246. 10.1126/science.1140615. [PubMed: 17431171]
9. Rocchi M, Archidiacono N, Schempp W, Capozzi O, and Stanyon R (2012). Centromere repositioning in mammals. Heredity 108, 59–67. 10.1038/hdy.2011.101. [PubMed: 22045381]
10. Arora UP, Charlebois C, Lawal RA, and Dumont BL (2021). Population and subspecies diversity at mouse centromere satellites. BMC Genom. 22, 279. 10.1186/s12864-021-07591-5.
11. Cacheux L, Ponger L, Gerbault-Seureau M, Richard FA, and Escudé C (2016). Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: Cercopithecus solatus. BMC Genom. 17, 916. 10.1186/s12864-016-3246-5.
12. Musich PR, Brown FL, and Maio JJ (1980). Highly repetitive component alpha and related alphoid DNAs in man and monkeys. Chromosoma 80, 331–348. 10.1007/BF00292688. [PubMed: 7438883]
13. Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, and Yurov Y (2001). Alpha-satellite DNA of primates: old and new families. Chromosoma 110, 253–266. 10.1007/s004120100146. [PubMed: 11534817]
14. Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, Ryder OA, Purgato S, Zoli M, Della Valle G, Eichler EE, and Ventura M (2011). Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 21, 137–145. 10.1101/gr.111278.110. [PubMed: 21081712]
15. Henikoff S, Ahmad K, and Malik HS (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293, 1098–1102. 10.1126/science.1062939. [PubMed: 11498581]
16. Talbert PB, Bryson TD, and Henikoff S (2004). Adaptive evolution of centromere proteins in plants and animals. J. Biol. 3, 18. 10.1186/jbiol11. [PubMed: 15345035]
17. Foltz DR, Jansen LET, Black BE, Bailey AO, Yates JR, and Cleveland DW (2006). The human CENP-A centromeric nucleosome-associated complex. Nat. Cell Biol. 8, 458–469. 10.1038/ncb1397. [PubMed: 16622419]
18. Okada M, Cheeseman IM, Hori T, Okawa K, McLeod IX, Yates JR, Desai A, and Fukagawa T (2006). The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. Nat. Cell Biol. 8, 446–457. 10.1038/ncb1396. [PubMed: 16622420]
19. Perpelescu M, and Fukagawa T (2011). The ABCs of CENPs. Chromosoma 120, 425–446. 10.1007/s00412-011-0330-0. [PubMed: 21751032]

20. Malik HS, and Henikoff S (2001). Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics 157, 1293–1298. [PubMed: 11238413]
21. Malik HS (2009). The centromere-drive hypothesis: a simple basis for centromere complexity. Prog. Mol. Subcell. Biol. 48, 33–52. 10.1007/978-3-642-00182-6_2. [PubMed: 19521811]
22. Drinnenberg IA, Henikoff S, and Malik HS (2016). Evolutionary Turnover of Kinetochore Proteins: A Ship of Theseus? Trends Cell Biol. 26, 498–510. 10.1016/j.tcb.2016.01.005. [PubMed: 26877204]
23. Kumon T, Ma J, Akins RB, Stefanik D, Nordgren CE, Kim J, Levine MT, and Lampson MA (2021). Parallel pathways for recruiting effector proteins determine centromere drive and suppression. Cell 184, 4904–4918.e11. 10.1016/j.cell.2021.07.037. [PubMed: 34433012]
24. Pardo-Manuel de Villena F, and Sapienza C (2001). Female meiosis drives karyotypic evolution in mammals. Genetics 159, 1179–1189. 10.1093/genetics/159.3.1179. [PubMed: 11729161]
25. Kursel LE, and Malik HS (2018). The cellular mechanisms and consequences of centromere drive. Curr. Opin. Cell Biol. 52, 58–65. 10.1016/j.ceb.2018.01.011. [PubMed: 29454259]
26. Kruger AN, and Mueller JL (2021). Mechanisms of meiotic drive in symmetric and asymmetric meiosis. Cell. Mol. Life Sci. 78, 3205–3218. 10.1007/s00018-020-03735-0. [PubMed: 33449147]
27. Iwata-Otsubo A, Dawicki-McKenna JM, Akera T, Falk SJ, Chmátal L, Yang K, Sullivan BA, Schultz RM, Lampson MA, and Black BE (2017). Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes that Drive in Female Meiosis. Curr. Biol. 27, 2365–2373.e8. 10.1016/j.cub.2017.06.069. [PubMed: 28756949]
28. Aldrup-MacDonald ME, Kuo ME, Sullivan LL, Chew K, and Sullivan BA (2016). Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. Genome Res. 26, 1301–1311. 10.1101/gr.206706.116. [PubMed: 27510565]
29. Zhang T, Talbert PB, Zhang W, Wu Y, Yang Z, Henikoff JG, Henikoff S, and Jiang J (2013). The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. Proc. Natl. Acad. Sci. USA 110, E4875–E4883. 10.1073/pnas.1319548110. [PubMed: 24191062]
30. Catania S, Pidoux AL, and Allshire RC (2015). Sequence features and transcriptional stalling within centromere DNA promote establishment of CENP-A chromatin. PLoS Genet. 11, e1004986. 10.1371/journal.pgen.1004986. [PubMed: 25738810]
31. Walstein K, Petrovic A, Pan D, Hagemeier B, Vogt D, Vetter IR, and Musacchio A (2021). Assembly principles and stoichiometry of a complete human kinetochore module. Sci. Adv. 7, eabg1037. 10.1126/sciadv.abg1037. [PubMed: 34193424]
32. Komissarov AS, Gavrilova EV, Demin SJ, Ishov AM, and Podgornaya OI (2011). Tandemly repeated DNA families in the mouse genome. BMC Genom. 12, 531. 10.1186/1471-2164-12-531.
33. Amemiya HM, Kundaje A, and Boyle AP (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. 9, 9354. 10.1038/s41598-019-45839-z. [PubMed: 31249361]
34. Masumoto H, Masukata H, Muro Y, Nozaki N, and Okazaki T (1989). A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J. Cell Biol. 109, 1963–1973. 10.1083/jcb.109.5.1963. [PubMed: 2808515]
35. Ohzeki J.i., Nakano M, Okada T, and Masumoto H (2002). CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. J. Cell Biol. 159, 765–775. 10.1083/jcb.200207112. [PubMed: 12460987]
36. Hudson DF, Fowler KJ, Earle E, Saffery R, Kalitsis P, Trowell H, Hill J, Wreford NG, de Kretser DM, Cancilla MR, et al. (1998). Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. J. Cell Biol. 141, 309–319. [PubMed: 9548711]
37. Fachinetti D, Han JS, McMahon MA, Ly P, Abdullah A, Wong AJ, and Cleveland DW (2015). DNA Sequence-Specific Binding of CENP-B Enhances the Fidelity of Human Centromere Function. Dev. Cell 33, 314–327. 10.1016/j.devcel.2015.03.020. [PubMed: 25942623]
38. Smith OK, Limouse C, Fryer KA, Teran NA, Sundararajan K, Heald R, and Straight AF (2021). Identification and characterization of centromeric sequences in *Xenopus laevis*. Genome Res. 31, 958–967. 10.1101/gr.267781.120. [PubMed: 33875480]

39. Smurova K, and De Wulf P (2018). Centromere and Pericentromere Transcription: Roles and Regulation, in Sickness and in Health. Front. Genet. 9, 674. 10.3389/fgene.2018.00674. [PubMed: 30627137]
40. Quénet D, and Dalal Y (2014). A long non-coding RNA is required for targeting centromeric protein A to the human centromere. Elife 3, e03254. 10.7554/eLife.03254. [PubMed: 25117489]
41. Koster MJE, Snel B, and Timmers HTM (2015). Genesis of chromatin and transcription dynamics in the origin of species. Cell 161, 724–736. 10.1016/j.cell.2015.04.033. [PubMed: 25957681]
42. Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, Choo KHA, and Wong LH (2012). Active transcription and essential role of RNA polymerase II at the centromere during mitosis. Proc. Natl. Acad. Sci. USA 109, 1979–1984. 10.1073/pnas.1108705109. [PubMed: 22308327]
43. Ferri F, Bouzinba-Segard H, Velasco G, Hubé F, and Francastel C (2009). Non-coding murine centromeric transcripts associate with and potentiate Aurora B kinase. Nucleic Acids Res. 37, 5071–5080. 10.1093/nar/gkp529. [PubMed: 19542185]
44. Sullivan BA, and Karpen GH (2004). Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nat. Struct. Mol. Biol. 11, 1076–1083. 10.1038/nsmb845. [PubMed: 15475964]
45. Chmátal L, Gabriel SI, Mitsainas GP, Martínez-Vargas J, Ventura J, Searle JB, Schultz RM, and Lampson MA (2014). Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. Curr. Biol. 24, 2295–2300. 10.1016/j.cub.2014.08.017. [PubMed: 25242031]
46. Gamba R, and Fachinetti D (2020). From evolution to function: Two sides of the same CENP-B coin? Exp. Cell Res. 390, 111959. 10.1016/j.yexcr.2020.111959. [PubMed: 32173469]
47. Roberts AB, Russo A, Felici A, and Flanders KC (2003). Smad3: a key player in pathogenetic mechanisms dependent on TGF-beta. Ann. N. Y. Acad. Sci. 995, 1–10. 10.1111/j.1749-6632.2003.tb03205.x.
48. Altemose N, Maslan A, Smith OK, Sundararajan K, Brown RR, Mishra R, Detweiler AM, Neff N, Miga KH, Straight AF, and Streets A (2022). DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome wide. Nat. Methods 19, 711–723. 10.1038/s41592-022-01475-6. [PubMed: 35396487]
49. Lamb JC, and Birchler JA (2003). The role of DNA sequence in centromere formation. Genome Biol. 4, 214. 10.1186/gb-2003-4-5-214. [PubMed: 12734002]
50. Dumont M, and Fachinetti D (2017). DNA Sequences in Centromere Formation and Function. Prog. Mol. Subcell. Biol. 56, 305–336. 10.1007/978-3-319-58592-5_13. [PubMed: 28840243]
51. Sohlenius-Sternbeck AK (2006). Determination of the hepatocellularity number for human, dog, rabbit, rat and mouse livers from protein concentration measurements. Toxicol. Vitro 20, 1582–1586. 10.1016/j.tiv.2006.06.003.
52. Maloney KA, Sullivan LL, Matheny JE, Strome ED, Merrett SL, Ferris A, and Sullivan BA (2012). Functional epialleles at an endogenous human centromere. Proc. Natl. Acad. Sci. USA 109, 13704–13709. 10.1073/pnas.1203126109. [PubMed: 22847449]
53. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. (2012). ChIP-seq Guidelines and Practices of the ENCODE and modENCODE Consortia (Cold Spring Harbor Laboratory Press).
54. Chen S, Zhou Y, Chen Y, and Gu J (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890. 10.1093/bioinformatics/bty560. [PubMed: 30423086]
55. Wong AK, and Rattner JB (1988). Sequence organization and cytological localization of the minor satellite of mouse. Nucleic Acids Res. 16, 11645–11661. [PubMed: 3211746]
56. Li H, and Durbin R (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595. 10.1093/bioinformatics/btp698. [PubMed: 20080505]
57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. 10.1093/bioinformatics/btp352. [PubMed: 19505943]
58. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. (2011). Fast, scalable generation of high-quality protein multiple sequence

alignments using Clustal Omega. Mol. Syst. Biol. 7, 539. 10.1038/msb.2011.75. [PubMed: 21988835]

Highlights

- Pioneers reference-independent strategy to query the epigenetic landscape of repetitive DNA
- Inbred mouse strains differ in CENP-A positioning along the minor satellite sequence
- CENP-A associates with unique minor satellite variants in different strain backgrounds
- CENP-A associates with pericentromeric major satellite DNA in some inbred strains

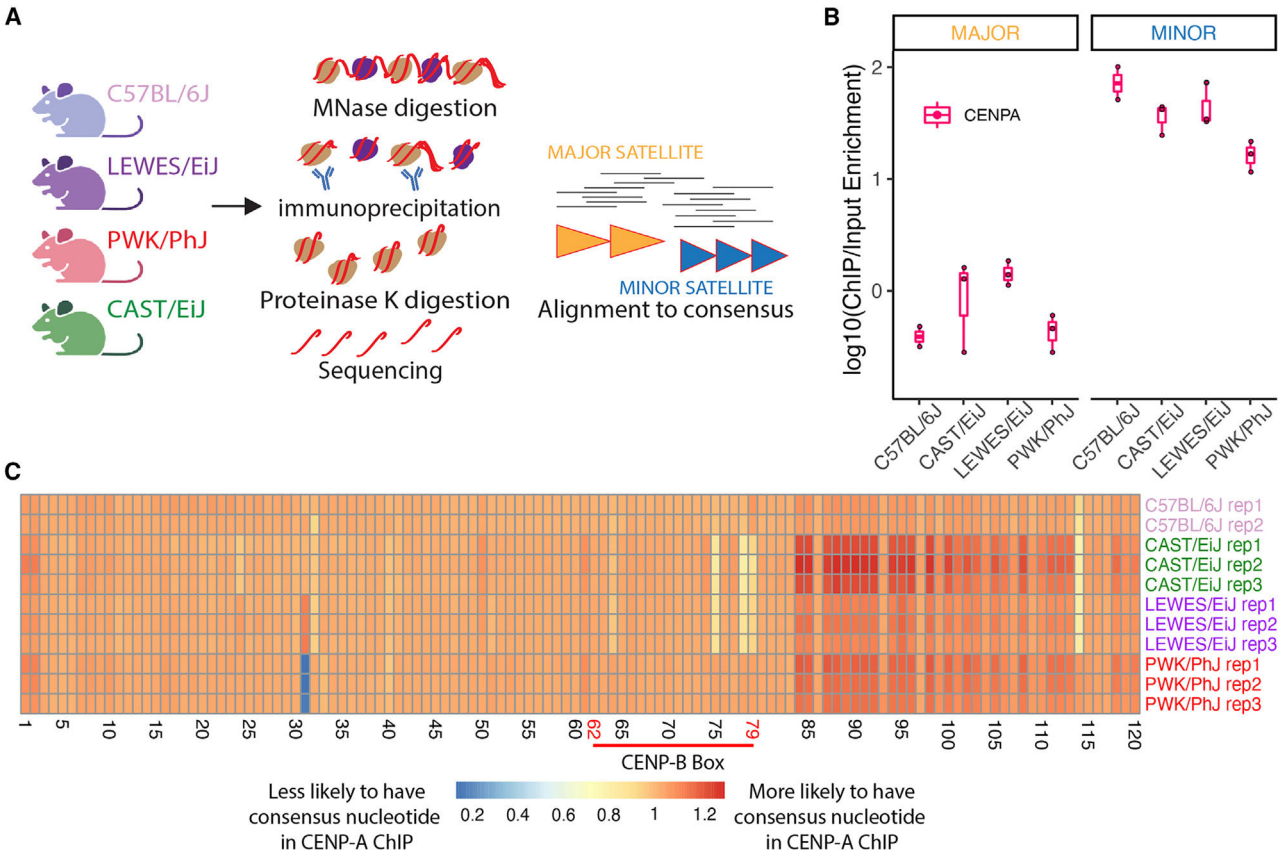


Figure 1. Consensus-based alignment analysis of strain differences in CENP-A enrichment at centromeres

(A) Experimental design of ChIP-seq and the consensus-guided genomic analyses for CENP-A ChIP sequence enrichment.

(B) Boxplots representing the enrichment of reads that mapped to either the major satellite (left) or minor satellite (right) consensus sequences in CENP-A ChIP relative to input samples. Color represents sample identity (AB, antibody). For each boxplot, the horizontal line represents the median, and the vertical line represents the range of values across replicates.

(C) Heatmap showing the enrichment of the consensus nucleotide along the minor satellite consensus sequence in CENP-A ChIP compared with input samples from each strain.

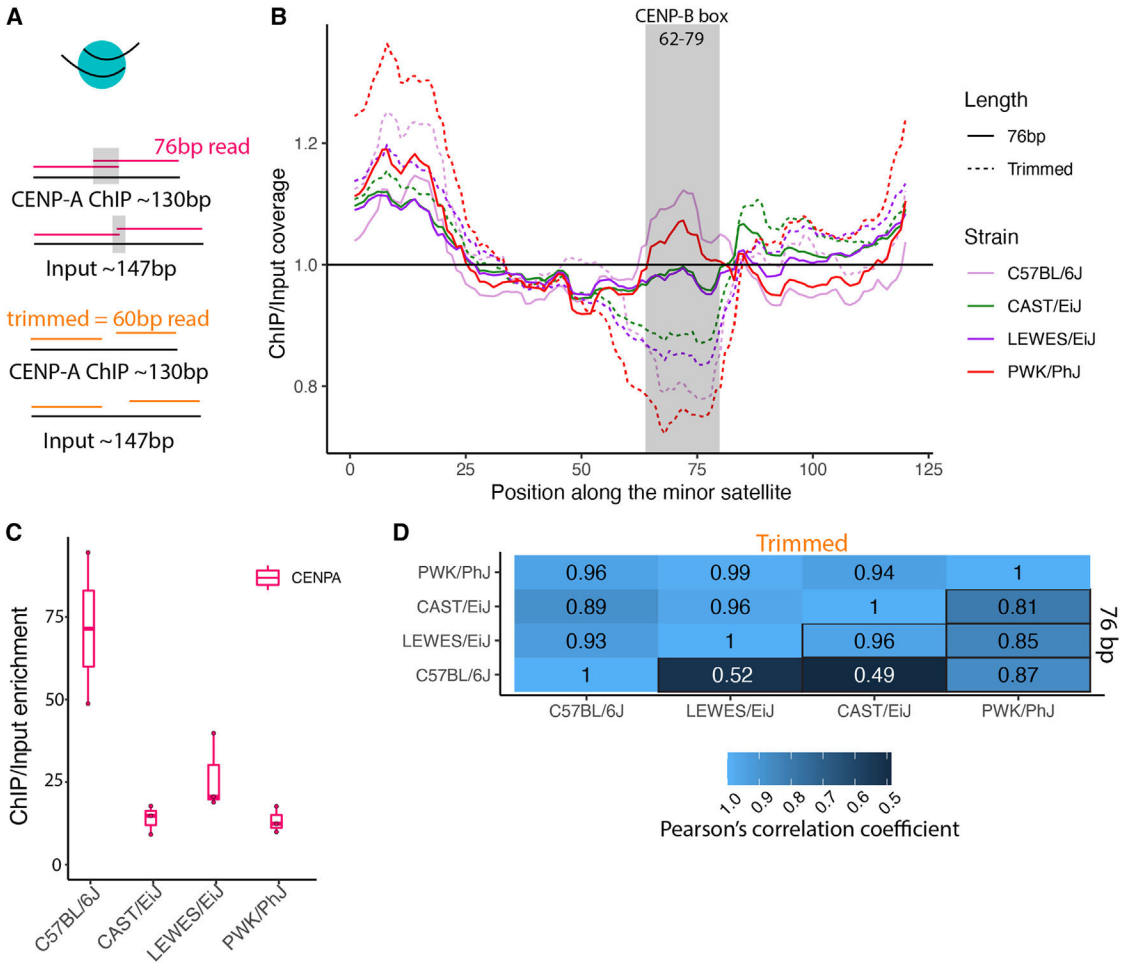


Figure 2. Strain differences in CENP-A positioning along the minor satellite consensus

(A) Schematic depicting the difference in CENP-A and H3 nucleosome coverage between 76-bp single-end reads and trimmed reads.

(B) Line plot representing the CENP-A ChIP/input enrichment of read coverage (y axis) at each position along the minor satellite consensus sequence (x axis). Solid (dashed) lines correspond to coverage values calculated using untrimmed (trimmed) reads. The 17-bp CENP-B box motif is marked in gray.

(C) Boxplots representing the percent enrichment of CENP-B box motif frequency in ChIP/input samples. For each boxplot, the horizontal line represents the median, and the vertical line represents the range of values.

(D) Heatmap presenting pairwise strain Pearson correlation coefficients for the average CENP-A ChIP/input enrichment pattern along the minor satellite consensus sequence. All Pearson correlation coefficients are significant ($p < 0.05$).

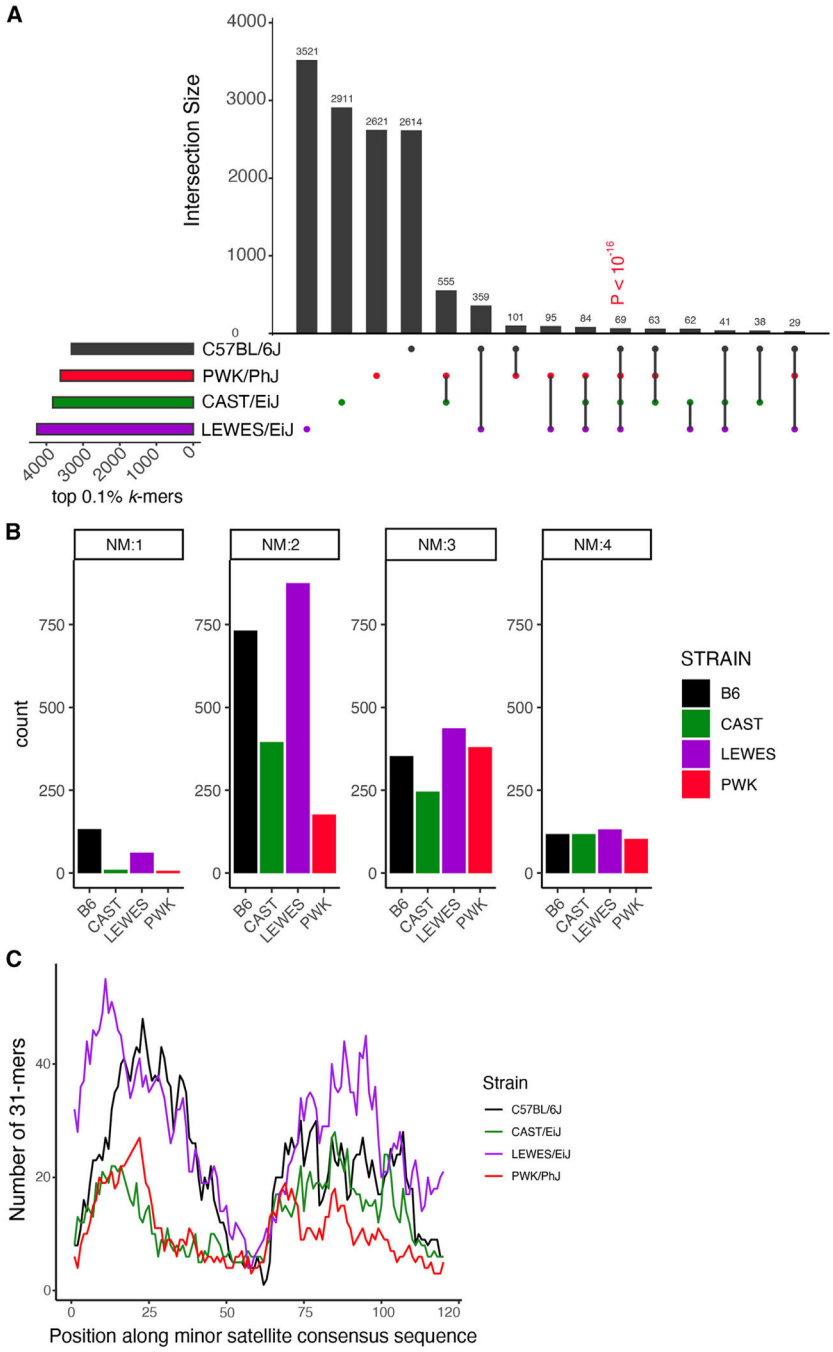


Figure 3. Shared and unique CENP-A-enriched *k*-mers across inbred mouse strains

(A) Upset plot representing the extent of *k*-mer sharing among the top 0.1% most enriched CENP-A-associated *k*-mers in each strain. Bar height corresponds to the total number of *k*-mers in each strain set. Each strain set of CENP-A-enriched *k*-mers is represented by the horizontal bars. The vertical bars represent the number of *k*-mers that belong to one or more strains, indicated by the dots below. p value was calculated by comparing the observed number of shared *k*-mers with the distribution of 100,000 bootstrap samples of the data.

(B) Bar plots showing the number of CENP-A-enriched *31*-mers from each strain with a given number of mismatches (NM) relative to the minor satellite consensus sequence. (C) Line plot depicting where CENP-A-enriched *31*-mers map along the minor satellite consensus sequence. The y axis represents the number of *31*-mers at a particular position along the minor satellite consensus sequence (x axis). Strains are indicated by line color.

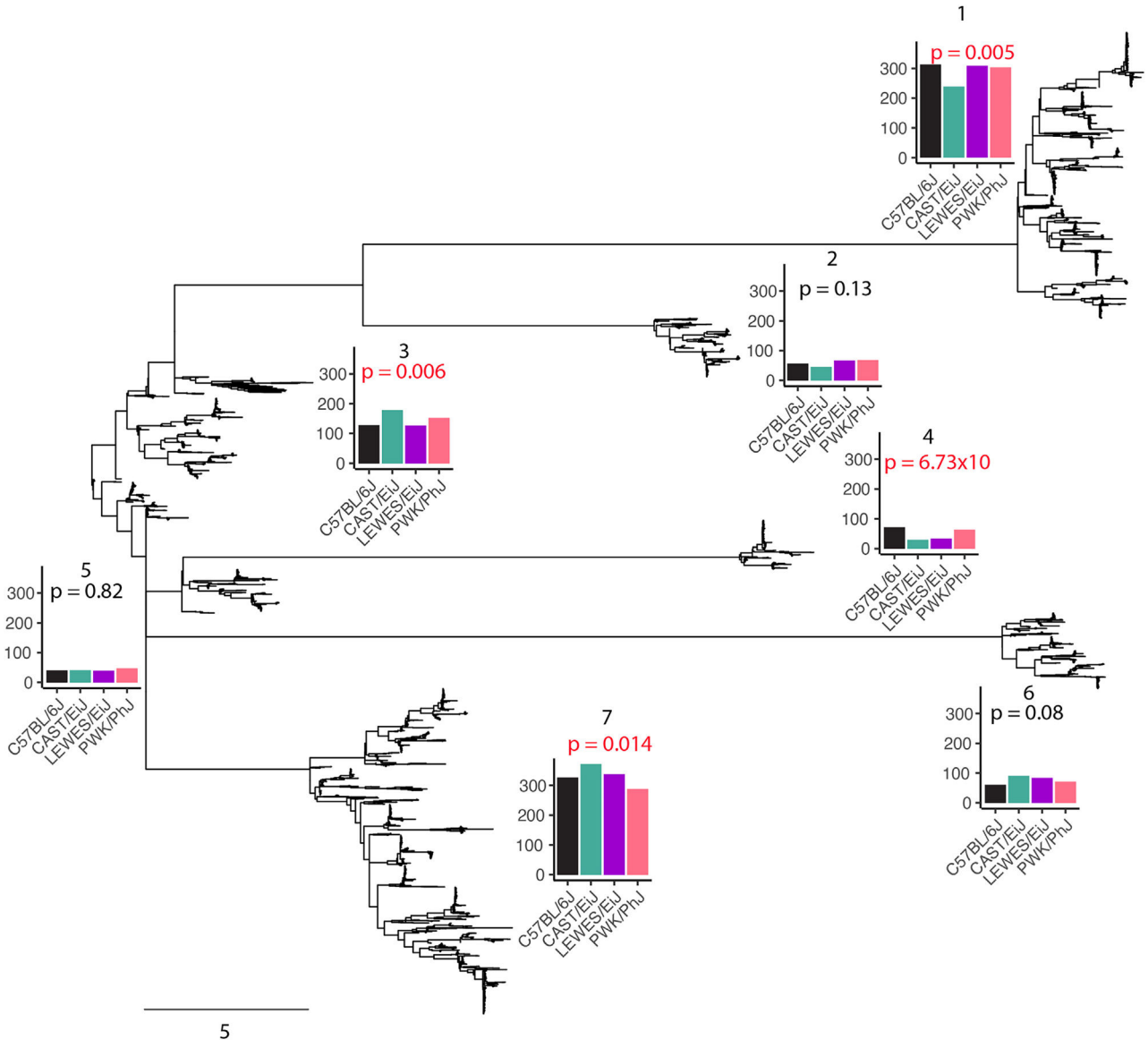


Figure 4. CENP-A-associated sequences identified by a reference-agnostic, *k*-mer-based approach form distinct subgroups

Neighbor joining tree constructed from each strain's top 1,000 CENP-A-associated sequences. Sequences cluster into seven groups. Strain-level contributions to each group are depicted by bar plots. Groups with skewed strain representation are indicated by red Chi-square p values (groups 1, 3, 4, 7). For clades with biased strain representation, Bonferroni post hoc tests were used to identify strains driving the significant signal.

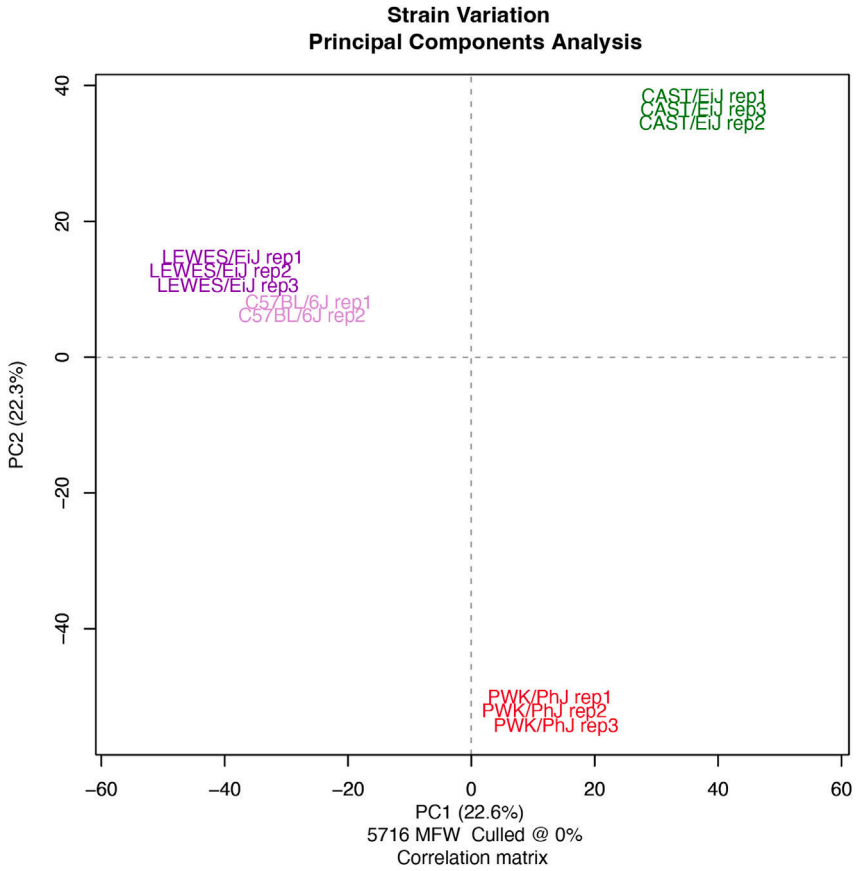


Figure 5. Relationship between CENP-A-associated sequences across strains

Principal component analysis based on the stylistic similarity of CENP-A-associated sequences in each strain.

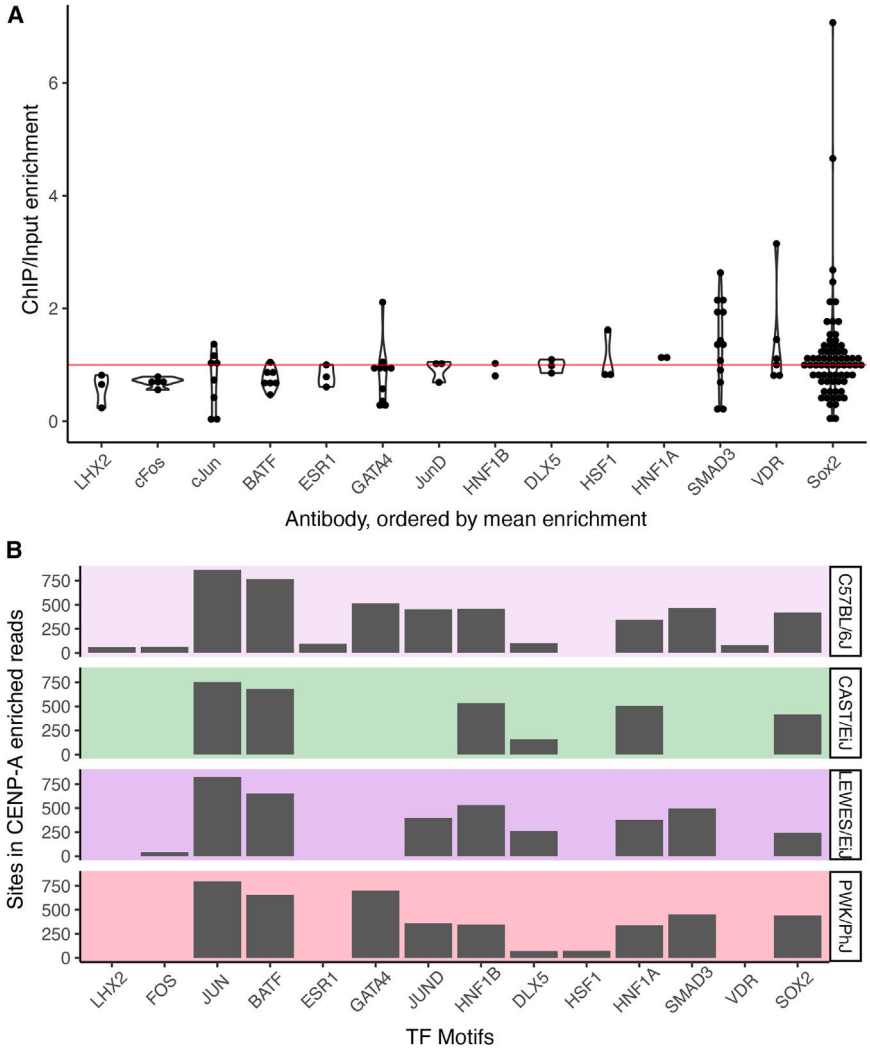


Figure 6. Transcription factor occupancy at centromere satellite DNA

(A) Association of transcription factors at centromere satellite DNA determined from publicly available ChIP-seq datasets. The y axis represents the enrichment of reads that map to the minor satellite consensus sequence in ChIP/input samples for each transcription factor (x axis). Each dot represents the average value of an experiment. The red horizontal line corresponds to ChIP/input value of 1.

(B) The frequency of multiple TF motifs among CENP-A-enriched sequences across diverse mouse strains.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit α-mouse-specific CENP-A	Beth Sullivan, Duke University	D601AP
Anti-rabbit IgG	Millipore	cat. # 12-370; RRID:AB_145841
Anti-histone H3K4me3	Millipore	Cat. # 07-473; RRID:AB_1977252
Chemicals, peptides, and recombinant proteins		
RIPA buffer	Sigma	Cat. #R0278
Bovine Serum Albumin	Sigma	CAS 9048-46-8
Herring Sperm DNA	Promega	D1811
Protease Inhibitor Cocktail	Sigma	P8340
MNase 100 U/μL	Sigma	PI88216
Proteinase K	New England Biolabs	Cat. #P8107S
Nonidet P-40 alternative	Sigma	492016
Protein G Dynabeads	Invitrogen	Cat. #10003D
PowerUp SYBR Green 2× Master Mix	Thermo Fisher Scientific	A25742
Critical commercial assays		
Qubit dsDNA High Sensitivity kit	Thermo Fisher Scientific	Cat. #Q32854
GeneJet PCR purification kit	Thermo Fisher Scientific	K0721
KAPA HyperPrep kit	Roche	KK8505
KAPA Library quantification kit	Roche	N/A
High Sensitivity TapeStation D5000	Agilent Technologies	5067-5592/5593/5594
High Output Reagents Kit	Illumina	N/A
Deposited data		
Custom Scripts	Zenodo	https://doi.org/10.5281/zenodo.8303213
Raw ChIP-sequecing data	SRA	PRJNA838487
Experimental models: Organisms/strains		
Mouse: C57BL/6J	The Jackson Laboratory	Stock no. 000664; RRID:IMSR_JAX:000664
Mouse: CAST/EiJ	The Jackson Laboratory	Stock no. 000928; RRID:IMSR_JAX:000928
Mouse: LEWES/EiJ	The Jackson Laboratory	Stock no. 002798; RRID:MGI:2164295
Mouse: PWK/PhJ	The Jackson Laboratory	Stock no. 003715; RRID:IMSR_JAX:003715
Oligonucleotides		
ActB Promoter F primer: 5'- GCCATAAAAGGCAACTTTCG-3'	This paper	N/A
ActB Promoter R primer 5'- TTTCAAAAGGAGGGGAGAGG-3'	This paper	N/A

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Minor Satellite F primer 5'- CATGGAAAATGATA AAAACC-3'	Iwata-Otsubo et al.[27]	N/A
Minor Satellite qPCR R primer 5' - CATCTAATATGTTCTACAGTGTGG-3'	Iwata-Otsubo et al.[27]	N/A
Software and algorithms		
fastp (v0.23.1)		RRID:SCR_016962
bwa (v0.7.9)		RRID:SCR_010910
samtools (v1.8)		RRID:SCR_002105
clumpify (v37.44)		N/A
liftover		RRID:SCR_018160
ape (v5.6–2)		N/A
BSDA (v. 1.2.1)		N/A
stylo (v0.7.4)		N/A
pheatmap (v1.0.12)		RRID:SCR_016418
R v4.0.5		RRID:SCR_001905
Clustal Omega		RRID:SCR_001591
MEME (v5.4.1)		RRID:SCR_001783
MEGA 11 (v5.6–2)		RRID:SCR_023017
Other		
ENSEMBL SNP effect	Sanger Mouse Genomes Project	https://www.mousegenomes.org/snps-indels/
Blacklisted genomic regions	Boyle Lab	https://github.com/Boyle-Lab/Blacklist/tree/master/lists
NCBI SRA		RRID:SCR_004891