Title

Conservation of chromatin organization within human and primate centromeres

Authors

Danilo Dubocanin[1], Adriana E. Sedeno Cortes[1], Gabrielle A. Hartley[2], Jane Ranchalis[1], Aman Agarwal[1], Glennis A. Logsdon[3], Katherine M. Munson[3], Taylor Real[1,3], Benjamin J. Mallory[1,3], Evan E. Eichler[3,4], Rachel J. O'Neill[2,5,6], Andrew B. Stergachis[1,3,7,*]

Affiliations

[1]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA
[2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA
[3]Department of Genome Sciences, University of Washington, Seattle, WA
[4]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.
[5]Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA
[6]Department of Genomics and Genome Sciences, UConn Health, Farmington, CT, USA
[7]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA
[*]Correspondence: absterga@uw.edu

Summary

The focal attachment of the kinetochore to the centromere core is essential for genome maintenance, yet the highly repetitive nature of human centromeres limits our understanding of their chromatin organization. We demonstrate that single-molecule chromatin fiber sequencing can uniquely resolve chromatin organization within centromeres at single-molecule and single-nucleotide resolution. We find that the centromere core contains a dichotomous chromatin organization not found elsewhere in the genome, which is characterized by highly accessible chromatin patches heterogeneously punctuated amongst tightly compacted nucleosome arrays. These highly accessible chromatin patches correspond to sites of kinetochore attachment, and clustered CENP-B occupancy within these patches phase nucleosome arrays to the alpha-satellite repeat. This dichotomous chromatin organization is conserved among humans despite the marked divergence of the underlying alpha-satellite organization and is similarly conserved in gibbon centromeres that lack alpha-satellite repeats, indicating that functional conservation within centromeres is mediated at the level of chromatin, not DNA sequence.

Highlights

- Dichotomous accessible and compacted chromatin (dichromatin) marks centromere cores
- Highly accessible chromatin patches punctuate sites of kinetochore attachment
- Dichromatin can form irrespective of CENP-B occupancy
- Conservation within centromeres is mediated at the level of chromatin, not DNA

Keywords

Chromatin, Centromere, Single-molecule, CENP-B, Human, Gibbon, Alpha-satellite, Fiber-seq

Introduction

Human centromeres are comprised of ~171 bp alpha-satellite repeat units organized into higher-order repeats (HORs) that can span several megabases on each chromosome. These alpha-satellite HORs serve as the genetic substrate for kinetochore attachment to the human genome within the centromere core[1,2], and kinetochore interactions within the centromere core are modulated by several DNA-binding proteins, including the sequence-specific DNA binding protein CENP-B[3,4] and the histone H3 variant CENP-A[5]. Imaging and genomic studies have demonstrated that kinetochore attachment is limited to only a small portion of the alpha-satellite HOR array, which is marked by CENP-A and CENP-B occupancy along the autosomes and X chromosome[6,7] and hypo-CpG methylation[8–10] (*i.e.,* the centromere core or 'centromere dip region [CDR]'). However, *in vivo* chromatin compaction and organization within this region remains largely unresolved, with somewhat contradictory chromatin features appearing to localize to the centromere core. For example, the centromere core contains a unique patterning of histones referred to as 'centro-chromatin', which is comprised of both centromere-specific CENP-A nucleosomes, as well as H3 nucleosomes dimethylated on Lys4 (H3K4me2)[11] that are often associated with euchromatic portions of the genome[12]. Despite the presence of these H3K4me2 histone, the centromere core is thought to contain condensed chromatin arrays *in vivo*[13], which are proposed to arise from CENP-A mediated nucleosome condensation[14] as well as other kinetochore proteins bridging and compacting neighboring nucleosomes[15–17].

Resolving exactly how chromatin is organized along individual chromatin fibers within the centromere as well as the relationship between kinetochore binding and chromatin organization has been challenging due to the highly repetitive DNA content of alpha-satellite HORs[13]. Specifically, centromeres are systematically unresolved in most reference genomes, and short-read sequencing-based chromatin profiling methods cannot uniquely resolve chromatin architectures across these repetitive DNA arrays (**Figure S1A**). In addition, short-read sequencing-based chromatin profiling methods are inherently unable to resolve how multiple chromatin features are positioned along a single chromatin fiber, as these methods inherently massively fragment chromatin fibers in the process of studying them.
The completion of the first telomere-to-telomere (T2T) human reference genome[18] in combination with emerging methyltransferase-based single-molecule long-read chromatin profiling methods[19–22] opens the possibility of studying chromatin features across centromeres at near base-pair resolution. However, GpC or CpG methyltransferase-based chromatin profiling methods remain limited in resolving centromere chromatin owing to the uneven distribution of GpC or CpG dinucleotides within alpha-satellite DNA (**Figure S1B and S1C**). In contrast, as the A/T content of centromeric DNA is evenly distributed (**Figure S1B and S1C**), non-specific adenine methyltransferase (m6A-MTase)-based chromatin profiling methods overcome these barriers, enabling nucleotide-precise mappings of chromatin accessibility, protein occupancy, and nucleosome positioning along multi-kilobase chromatin fibers[19]. For example, Fiber-seq utilizes a non-specific m6A-MTase[23] to stencil the chromatin architecture of individual multi-kilobase fibers onto their underlying DNA templates via methylated adenines (**Figure 1A**), which is a non-endogenous DNA modification in humans[24,25]. The genetic and chromatin architecture of each fiber is directly read using highly accurate single-molecule PacBio HiFi long-read DNA sequencing, which is capable of accurately distinguishing m6A-modified bases[19].

Results

Mapping chromatin architectures of human centromeres

To investigate the chromatin architecture of human centromeres, we first applied Fiber-seq to the human hydatidiform mole CHM13 cell line[26], the same line used for constructing the first human telomere-to-telomere (T2T) reference genome[9,18]. We observed that coupling long-read epigenome maps with paired fully sequenced and assembled genome maps from the same individual enabled us to

overcome the substantial diversity between individuals[27] that is enriched within these highly repetitive genomic regions. Application of Fiber-seq to CHM13 cells resulted in 1.7 million individual ~16 kb chromatin fiber architectures (**Figures 1A and 1B**), enabling the nucleotide-precise delineation of over 230 million chromatin features, such as nucleosomes, accessible regulatory elements, and single-molecule transcription factor binding events. Furthermore, the high sequence accuracy of these reads enabled us to uniquely map the chromatin architecture of all centromeres in CHM13 cells (**Figures 1C and 1D**).

Dichotomous accessible chromatin patches mark the centromere core

We first sought to delineate how nucleosomes are organized within centromeres by evaluating the sizes of nucleosome footprints using Fiber-seq. Genome-wide, nucleosomes predominantly form mononucleosome footprints with an average size of ~148 bp (**Figure S1D**). However, within the centromere core, mononucleosome footprints are significantly smaller with a peak size of 125 bp (**Figure S1D**), consistent with a large population of these mononucleosomes containing CENP-A, which is known to wrap only 121 bp of DNA[13,28,29]. Heterochromatic domains are characteristically populated by compacted dinucleosomes[30], which have recently been identified within the centromere core[13]. Consistent with this, we observed that upwards of 30% of all nucleosome footprints within satellite DNA arise from compacted di-nucleosomes, with the centromere core containing the highest enrichment of di-nucleosomes within the genome (**Figures 1E and 1F**).

As satellite DNA traditionally constitutes constitutive heterochromatin and centromeres are known to contain condensed chromatin arrays *in vivo*[13], we predicted that these regions would be depleted of large patches of accessible chromatin, which form the basis of accessible regulatory DNA throughout the euchromatic genome[19,31] (**Figure 1E**). Indeed, satellite DNA was almost entirely devoid of large patches of accessible chromatin. However, centromere cores were markedly enriched for large patches of accessible chromatin, forming the most accessible chromatin domain within the entire human genome (**Figures 1E, 1F and S1E**). Notably, individual chromatin fibers originating from the centromere core contain both tightly compacted and highly accessible chromatin features abutting each other (**Figures 1E, 1G and S1E**), indicating that chromatin compaction within the centromere core does not conform to the standard heterochromatin/euchromatin dichotomy.

Within the centromere core, accessible chromatin patches appear to heterogeneously cluster along individual chromatin fibers (**Figure 1H**). Although hundreds of alpha-satellite repeats within each centromere core appear capable of forming accessible chromatin patches, on average, each chromosome molecule contains only 7-22 accessible chromatin patches within its 80-220 kbp centromere core (**Figure 1I**), a value that mirrors estimates of the number of kinetochore microtubules that form on individual human centromeres[32,33]. Notably, longer chromosomes have significantly more accessible chromatin patches than shorter chromosomes, and after controlling for chromosome length, three of the chromosomes with the lowest amount of accessible chromatin patches also had the highest rate of missegregation in human RPE-1 cells[34] (**Figure 1I**). Together, these findings establish that the centromere core contains a dichotomous chromatin organization not found elsewhere in the genome, which is characterized by highly accessible chromatin patches that likely play a direct role in kinetochore attachment and function within the centromere core.

Centromere core chromatin mirrors the alpha-satellite DNA repeat

The higher-order architecture of chromatin fibers is typically comprised of nucleosomes repeating every 180-200 bp (this includes both the nucleosome footprint, as well as the internucleosomal linker region)[35,36]. However, CENP-A containing nucleosomes within the centromere core are thought to be

preferentially positioned along the 171 bp alpha-satellite repeat[29,37], yet how this impacts higher order chromatin architectures within the centromere core remains unclear. To resolve this, we identified recurrent chromatin repeat units along individual chromatin fibers in CHM13 cells by applying a Fourier Transform to each fiber's raw m6A signal. Consistent with prior reports[35,36], chromatin fibers originating from euchromatic and heterochromatic genomic regions have nucleosomes positioned every ~179-191 bp (**Figures 2A, 2B and S2**), including alpha-satellite regions not involved in kinetochore binding. However, chromatin fibers originating from the centromere core show markedly divergent higher order architectures, with a chromatin repeat unit of 170 bp that mirrors the underlying repeat length of alpha-satellite DNA (**Figure 2B**). These findings demonstrate that the centromere core is marked by a unique chromatin higher order organization, and that the genetic structure of the alpha-satellite DNA repeat is playing a direct role in guiding this organization.

CpG methylation and CENP-B occupancy within the centromere core

Accessible chromatin patches within euchromatin are typically bound by sequence-specific DNA binding proteins that play a role in maintaining nucleosome-free chromatin patches[38,39]. As CENP-B is the only kinetochore component that binds DNA in a sequence-specific manner, we sought to evaluate the relationship between CENP-B occupancy and chromatin architecture within the centromere core. CENP-B is known to occupy a non-palindromic A/T-rich 17 bp CENP-B box[3,4] that is readily resolved *in vitro* using DNaseI footprinting[40]. We similarly found that CENP-B occupancy can be resolved using Fiber-seq, as many CENP-B boxes located within the centromere core demonstrate occupancy with an MTase-footprint that mirrors the *in vitro* CENP-B DNaseI footprint and the protein-DNA contacts within the crystal structure of the CENP-B-DNA complex[3,40,41] (**Figures 3A and 3B**).

We next sought to determine the role of CpG methylation in modulating CENP-B occupancy across the centromere, as CENP-B boxes contain 2 CpG dinucleotides, and CpG methylation is a known modifier of CENP-B affinity and occupancy *in vitro*[42,43]. To accomplish this, we first evaluated whether we could uncover CpG methylation patterns directly from our Fiber-seq reads. Notably, we observed that CpG methylation patterns derived from Fiber-seq reads mirrored those from Nanopore data (**Figures 3A and 3C**), indicating that Fiber-seq enables the accurate measurement of chromatin architectures and CpG methylation patterns along individual chromatin fibers.

Utilizing the paired single-molecule CpG methylation and footprint occupancy information at each CENP-B box within the centromere, we identified that within the centromere core, CENP-B boxes without CpG methylation demonstrate robust CENP-B occupancy. In contrast, CpG methylation at even one of the 2 CpG dinucleotides within the CENP-B box largely abrogates CENP-B occupancy (**Figures 3B, 3D-3F, S3A, and S3B**). Notably, outside of the centromere core, CENP-B boxes were largely unoccupied irrespective of their CpG methylation status (**Figures 3D and S3C**), with weak CENP-B occupancy only observed at sporadic unmethylated CENP-B boxes located within alpha-satellite HORs flanking the centromere core, a pattern consistent with prior imaging studies[7]. Together, these findings demonstrate that accessible chromatin patches within the centromere core are frequently occupied by the kinetochore component CENP-B.

CENP-B occupancy within the centromere core phases dichromatin architectures

We next sought to determine whether CENP-B occupancy is guiding the atypical chromatin patterns present within the centromere core, as CENP-B boxes predominantly punctuate alternating repeats within the alpha-satellite HOR[44] (**Figure S3D**), and CENP-B occupancy is known to induce neighboring nucleosome positioning along these repeats *in vitro*[40,45]. Consistent with these *in vitro* findings, we found that nucleosomes are preferentially positioned immediately adjacent to CENP-B boxes that lack CpG

methylation (**Figures 3E and 3G**). In addition, accessible chromatin patches similarly were preferentially positioned adjacent to CENP-B boxes that lack CpG methylation (**Figure S3E**). Together, these findings suggest a model whereby CENP-B occupancy within the centromere core is organizing chromatin architecture to mirror the underlying alpha-satellite repeat.

To disentangle the role of CENP-B in dichromatin formation, we evaluated chromatin architectures along the Y chromosome centromere, which is formed along alpha-satellite repeats that lack CENP-B boxes[46]. Application of Fiber-seq to 46,XY GM24385 cells enabled the delineation of chromatin architectures along the complete GM24385 Y chromosome reference genome (*i.e.,* HG002)[10,18], demonstrating that the Y chromosome centromere core is marked by two distinct hypo-CpG methylated regions (**Figures 4A and S4**). Notably, these hypo-CpG methylated regions are markedly enriched in both compacted di-nucleosome footprints, as well as large patches of accessible chromatin (**Figures 4A-4C**), indicating that these features of centromere core chromatin architecture can occur irrespective of CENP-B occupancy. However, in contrast to studies using short-read sequencing methods[29], higher order chromatin organization within the centromere core region of chromosome Y did not mirror the underlying alpha-satellite repeat unit, but rather was a continuation of the canonical chromatin repeat unit observed throughout the rest of the Y chromosome (**Figure 4D**). Together, these findings indicate that chromatin architectures within human centromere cores are phased via CENP-B occupancy, but dichromatin formation within these regions can form irrespective of CENP-B.

Centromere core chromatin architecture is conserved despite divergence in alpha-satellite sequence

We next sought to evaluate the conservation of centromere core chromatin architecture between two human genomes, as the underlying DNA sequence and structure of each centromere can markedly differ between individuals (Logsdon et al., in preparation). To accomplish this, we applied Fiber-seq to CHM1 cells, as a reference genome for all CHM1 centromeres was recently resolved (Logsdon et al., in preparation) (**Figure 5A**). Notably, centromere core regions (*i.e.,* hypo-CpG methylated regions) within each CHM1 centromere markedly diverge in both their sequence content and location when compared to CHM13 cells (Logsdon et al., in preparation) (**Figures 5B, 5C and S5**). However, despite the sequence divergence of centromere cores within CHM1, these regions are still marked by dichromatin – exhibiting both compacted di-nucleosome footprints as well as large patches of accessible chromatin intermingled along the same chromatin fiber (**Figures 5C, 5D and S5**). In addition, chromatin higher order organization along the autosomes and X chromosome within CHM1 cells similarly displayed a chromatin repeat unit of 170 bp selectively within centromere cores (**Figure 5E**). Together, these findings demonstrate that dichromatin is a conserved feature of chromatin compaction within the centromere core in humans.

Centromere core dichromatin architecture does not require alpha satellite repeats

To further disentangle the relationship between centromere sequence, chromatin, and CENP-B occupancy, we sought to study chromatin architectures within gibbon centromeres. Gibbons have undergone rapid karyotypic evolution accompanied by the inactivation of many ancestral centromeres and the formation of evolutionary new centromeres[47] that are largely composed of transposable elements that lack CENP-B boxes and alpha-satellite repeats[48]. We first assembled and validated the sequence of five centromeres from the eastern hoolock gibbon (*Hoolock leuconedys*) by applying Oxford Nanopore (ONT) sequencing, Hi-C, Fiber-seq, and CENP-A CUT&RUN to a 46,XX lymphoblastoid cell line derived from Betty (**Figure 6A**). These centromeres contained hypo-CpG methylated regions corresponding to sites of CENP-A occupancy (**Figure 6B and S6**). However, the sequence of these centromere cores was largely derived from transposable elements, not alpha satellite repeats. Notably, despite the lack of alpha-satellite repeats within these centromere cores, these regions were still

marked by dichromatin – exhibiting both compacted di-nucleosome footprints as well as clustered patches of accessible chromatin intermingled along the same chromatin fibers (**Figure 6C, 6D and S6E**). Unlike human autosomes, the higher order chromatin organization within the eastern hoolock gibbon centromere core regions appeared to be a continuation of the canonical chromatin repeat unit observed within flanking regions, albeit less organized owing to the dichromatin features (**Figure 6E**). Together, these findings demonstrate that dichromatin is a conserved feature of centromere core chromatin compaction irrespective of the underlying DNA sequence.

Discussion

We demonstrate that pairing complete reference centromere genome assemblies with single-molecule chromatin fiber sequencing (Fiber-seq) enables the delineation of chromatin architectures within centromeres at single-molecule and near single-nucleotide resolution. Application of Fiber-seq to four human and non-human primate samples with complete centromere assemblies exposed a unique form of chromatin compaction within the centromere cores that is not observed elsewhere along the genome. Specifically, centromere cores simultaneously contain both the most accessible and compacted chromatin in the genome, with these dichotomous chromatin features both present along individual chromatin fibers. In addition, unlike traditional euchromatic regulatory elements, accessible chromatin patches within the centromere core are heterogeneously positioned across individual chromatin fibers. These dichotomous features of chromatin compaction within the centromere core diverge from the standard model of heterochromatin and euchromatin compaction and is hence termed 'dichromatin'.

These accessible chromatin patches that punctuate the centromere core create a highly plastic chromatin topology at the site of kinetochore attachment that is readily exposed to proteins essential for maintaining the centromere core[49]. Specifically, whereas the surrounding chromatin within the centromere is marked by tightly compacted nucleosome arrays that often stretch for >20kb in length[17], the centromere core contains highly heterogeneous chromatin with clustered accessible chromatin patches often disrupting the ability of compacted nucleosome arrays to extend beyond ~1kb in length.

The clustering of these patches along human autosome and chromosome X centromeres potentially reflects the action of bound CENP-B homodimers[50] binding to and bridging distinct CENP-B boxes in close physical proximity along the alpha-satellite array[40,51]. However, these accessible chromatin patches nonetheless cluster along centromeres that lack CENP-B occupancy, albeit to a lesser degree, indicating that CENP-B occupancy may simply exist as a mechanism for stabilizing this inherent property of chromatin within the centromere core. Notably, although CENP-B boxes selectively phase surrounding chromatin to mirror the underlying alpha-satellite repeat, this phasing phenomenon was only observed on centromeres containing occupied CENP-B boxes and was not necessary for dichromatin formation. The basic alpha-satellite repeat unit across species (*i.e.,* 171 bp in primates, 186 bp in the fish *Sparus aurata*, 155 bp in both Arabidopsis and maize, and 340 bp in pigs)[52] is well documented to reflect mono- or di-nucleosome lengths, and our findings suggest that the evolutionary pressure for this may in fact be to mediate the clustering of these accessible chromatin patches within the centromere core via the stochastic occupancy of dimeric CENP-B proteins at nearby CENP-B boxes along the chromatin fiber.

Centromeres comprise the most rapidly evolving DNA sequences in eukaryotic genomes[53], and the marked heterogeneity of chromatin organization within the centromere core likely enables kinetochore attachment to be particularly immune to sequence and structural variability within centromeres. Consistent with this, we find that the dichromatin architecture of centromere cores is highly conserved between individuals, despite the marked divergence of the underlying alpha-satellite organization. Furthermore, dichromatin can assemble within centromere cores that lack alpha-satellite repeats,

indicating that functional conservation within centromeres is mediated at the level of chromatin, not DNA. These findings are consistent with the centromere paradox[52], and raise additional questions as to how the centromere core remains faithfully localized to a specific region along the genome despite its predilection to form along divergent sequences.

Acknowledgements

We thank Steven Henikoff, Sue Biggins, Stirling Churchman and John Stamatoyannopoulos for their helpful comments and feedback. We are grateful to Nam Pho and the research computing group at the University of Washington for computational assistance. We thank the Telomere-to-Telomere (T2T) consortium for generating a complete reference genome of CHM13 cells, as well as the Y chromosome of HG002. This research was supported by NIH grant 1DP5OD029630 and a 2021 Catalytic Collaborations pilot grant from the Brotman Baty Institute for Precision Medicine. A.B.S. holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund and is a Pew Biomedical Scholar.

Author contributions

A.B.S., J.R. G.A.H., K.M.M., and B.M. designed and performed the experiments. A.B.S., D.D., A.E.S.C., G.A.H., G.A.L., E.E.E., R.J.O and T.R. performed the computational analyses. A.B.S. and D.D. wrote the manuscript.

Declaration of interests

A.B.S. is a co-inventor on a patent relating to the Fiber-seq method.

Figure legends

Figure 1. Dichotomous accessible chromatin patches mark the centromere core

(A) Schematic of Single-molecule chromatin fiber sequencing (Fiber-seq) for mapping protein occupancy events along multi-kilobase DNA molecules.

(B) Application of Fiber-seq to CHM13 cells, which have a paired complete T2T reference genome.

(C and D) Genome coverage and mapping quality of Fiber-seq reads to euchromatic and satellite DNA regions.

(E) Genomic locus of chromosome 5 centromere showing satellite repeats, bulk CpG methylation, Fiber-seq identified di-nucleosome footprint density, as well as Fiber-seq identified chromatin accessibility density. Below are individual Fiber-seq reads (grey bars) with m6A-modified bases in purple delineating single-molecule chromatin architectures within euchromatic and heterochromatic regions.

(F) Average density of di-nucleosome footprints and accessible chromatin patches within various genomic regions (* p-value <0.01 Mann-Whitney).

(G) Hexbin plot showing the single-molecule density of di-nucleosome footprints and accessible chromatin patches within various genomic regions.

(H) Swarm and box-and-whisker plots showing the distance between accessible chromatin patches along the same molecule of DNA within the centromere core, as well as the expected distance based on the density of accessible chromatin patches within each chromosome's centromere core (* p-value <0.01 Mann-Whitney).

(I) Estimated total number of accessible chromatin patches along each chromosome's centromere core versus the length of that chromosome. Chromosomes with high rates of mis-segregation[34] are in red.

See also Figure S1.

Figure 2. Centromere core chromatin mirrors the alpha-satellite DNA repeat

(A) Box-and-whisker plots of the chromatin repeat lengths from various genomic regions. Specifically, m6A-marked chromatin features from individual chromatin fibers were subjected to Fourier transform and the per-molecule spectral densities were then aggregated across different fibers from the same genomic region.

(B) Heatmap of the median spectral density for various chromatin repeat lengths, and the peak chromatin repeat length(s) from fibers contained within the euchromatic genome, or various satellite repeat regions. The satellite DNA repeat unit for each region is also indicated.

See also Figure S2.

Figure 3. CENP-B selectively occupies dichromatin within the centromere core

(A) Genomic locus of showing bulk CpG methylation determined using either Nanopore sequencing or PacBio sequencing of Fiber-seq library. Below are per-molecule m6A-marked chromatin architectures and CpG methylations demonstrating single-molecule footprints at centromere core CENP-B boxes.

(B) Structure of CENP-B bound to a CENP- box (PDB 1HLV), as well as in vitro DNaseI footprint of a CENP-B-occupied CENP-B box relative to aggregate m6A methylation profile at centromere core CENP-B boxes containing various levels of methylated CpGs (mCpGs).

(C) Heatmap showing the genome-wide similarity of CpG methylation calculated using Nanopore sequencing or PacBio sequencing of the CHM13 Fiber-seq library.

(D) Box-and-whisker plots of footprint scores at CENP-B boxes within various satellite regions as a function of mCpG status at the CENP-B box. Scores higher than 1 quantitatively indicate greater CENP-B occupancy.

(E-G) Directly comparing m6A-accessibility (F) and nucleosome footprint boundaries (G) at individual CENP-B boxes based on whether the read does or does not contain a mCpG within that CENP-B box.

Above is the aggregate m6A methylation (F) or nucleosome boundary positions (G) at centromere core and non-core alpha satellite CENP-B boxes containing no mCpG. Below is the aggregate difference in these features between reads with and without mCpG overlapping the CENP-B box.
See also Figure S3.

Figure 4. CENP-B boxes are essential for dichromatin phasing along alpha-satellite repeats
(A) Schematic for Fiber-seq in HG002/GM24385 cells, which contain a T2T Y chromosome reference. Below is the Fiber-seq identified di-nucleosome footprint density, and chromatin accessibility density along the entire chromosome Y. Bottom left inset shows these features and CpG methylation within the centromere alpha-satellite HOR, as well as the single-molecule chromatin architecture along individual alpha-satellite repeats.
(B-C) Average density of di-nucleosome footprints and accessible chromatin patches within various genomic regions along the Y chromosome in GM24385 cells (* p-value <0.01 Mann-Whitney).
(D) (left) Plot showing the average distance between CENP-B boxes within the centromere along each autosome and chromosome X in CHM13 cells. Chromosome Y centromere in GM24385 cells lacks CENP-B boxes. Heatmaps of the median spectral density for various chromatin repeat lengths derived from fibers within the centromere core or non-alpha satellite regions on each autosome and chromosome X in CHM13 cells, as well as chromosome Y centromere in GM24385 cells.
See also Figure S4.

Figure 5. Dichromatin is a conserved feature of centromere core chromatin in humans
(A) Schematic for Fiber-seq in CHM1 cells, which contain complete centromere reference genome assemblies for all autosomes and chromosome X.
(B-C) Genomic locus showing chromosome 5 centromere in both CHM13 (B) and CHM1 (C) cells, including CpG methylation, di-nucleosome footprint density, and chromatin accessibility. (bottom) Single-molecule chromatin architecture along individual alpha-satellite repeats within centromere core and non-core alpha-satellite repeats.
(D) Average density of di-nucleosome footprints and accessible chromatin patches within various genomic regions in CHM1 cells (* p-value <0.01 Mann-Whitney).
(E) Heatmap of the median spectral density for various chromatin repeat lengths derived from fibers within the centromere core or other regions in CHM1 cells.
See also Figure S5.

Figure 6. Centromere core dichromatin architecture does not require alpha-satellite repeats
(A) Schematic for Fiber-seq in a lymphoblastoid cell line from the eastern hoolock gibbon (*Hoolock leuconedys*) Betty. Ultralong read Nanopore sequencing was used for *de novo* genome assembly, which was validated using the contigs assembled using Fiber-seq sequencing data. CENP-A CUT&RUN was used to identify centromere cores along validated assembly regions.
(B) Genomic locus showing chromosome 8 centromere, including different repeat classes, CpG methylation, CENP-A CUT&RUN, and Fiber-seq derived di-nucleosome footprint density, and chromatin accessibility. (bottom) Single-molecule chromatin architecture from pericentromeric gene regulatory elements, centromere core, and heterochromatin.
(C) Average density of mCpG, di-nucleosome footprints, and accessible chromatin patches within and flanking centromere cores (* p-value <0.01 Mann-Whitney).
(D) Hexbin plot showing the single-molecule density of di-nucleosome footprints and accessible chromatin patches within centromere core and flanking regions.

(E) Box-and-whisker plots of the chromatin repeat lengths from centromere core and flanking regions. Below are indicated the peak chromatin repeat unit as well as the medial spectral density at that repeat unit showing that chromatin is more disorganized within the centromere core. See also Figure S6.

METHODS

Single-molecule chromatin fiber sequencing

CHM13 cells were grown in AmnioMax C-100 Basal Medium (Invitrogen 12558011), which includes the AmnioMAX™ C-100 Supplement to approximately 90% confluency at 37°C and 5% CO_2. Growth media was changed daily, and cells were split using 0.25% trypsin (GIbco 25200056). One million CHM13 cells per sample were pelleted at 250 x g for 5 minutes (4 samples were processed in parallel), and washed once with PBS and then pelleted again at 250 x g for 5 minutes. Each cell pellet was resuspended in 60 µl Buffer A (15 mM Tris, pH 8.0; 15 mM NaCl; 60 mM KCl; 1mM EDTA, pH 8.0; 0.5 mM EGTA, pH 8.0; 0.5 mM Spermidine) and 60 µl of cold 2X Lysis buffer (0.1% IGEPAL CA-630 in Buffer A) was added and mixed by gentle flicking then kept on ice for 10 minutes. Samples were then pelleted at 4°C for 5 min at 350 x g and the supernatant was removed. The nuclei pellets were gently resuspended individually with wide bore pipette tips in 57.5 µl Buffer A and moved to a 25°C thermocycler. 1 µl of Hia5 MTase (200 U)[19] and 1.5 µl 32 mM S-adenosylmethionine (NEB B9003S) (0.8 mM final concentration) were added, then carefully mixed by pipetting the volume up and down 10 times with wide bore tips. The reactions were incubated for 10 minutes at 25°C then stopped with 3 µl of 20% SDS (1% final concentration) and transferred to new 1.5 mL microfuge tubes. The sample volumes were adjusted to 100 µl by the addition of 37 µl PBS. To that, 500 µl of HMW Lysis Buffer A (Promega Wizard HMW DNA Extraction Kit A2920) and 3 µl RNAse A (ThermoFisher EN0531) were added. The tubes were mixed by inverting 7 times and incubated 15 minutes at 37°C. 20 µl of Proteinase K (Promega Wizard HMW DNA Extraction Kit A2920) was added and the samples mixed by inverting 10 times and incubated 15 minutes at 56°C followed by chilling on ice for 1 minute. Protein was precipitated by the addition of 200 µl Protein Precipitation Solution (Promega Wizard HMW DNA Extraction Kit A2920). Using a wide bore tip the samples were mixed by drawing up contents from the bottom of the tube and then expelled on the side of the tube 5 times. Tubes were centrifuged at 16,000g for 5 minutes. The supernatant was poured into a new 1.5 ml tube containing 600 µl isopropanol. This was mixed by gentle inversion 10 times, incubated 1 minute at room temperature then inverted an additional 10 times. DNA was precipitated by centrifugation at 16,000g for 5 minutes. The supernatant was decanted and the pellet washed with the addition of 70% ethanol and centrifuged again at 16,000 g 5 minutes. After the supernatant was decanted, a quick spin was performed to facilitate removal of any residual ethanol. Open tubes were air dried on the bench for 15 minutes. The DNA was resuspended by the addition of 25 µl 10mM Tris pH 8.0. Tubes were stored overnight in 4°C and the next day were mixed gently by using a wide bore pipette tip, drawing up the sample 5 times and gently expelling the contents on the side of the tube. Samples were then stored at -80°C prior to library construction.

DNA shearing, library construction, and PacBio Sequel II sequencing were performed as previously described[19] with the exception that 15-20kb fragments were targeted when shearing using the Megaruptor (Diagenode Diagnostics). In addition, we performed a high-pass size selection of the SMRTbell library using the Sage Science PippinHT platform (Sage Science cat. no. ELF0001) according to the manufacturer's protocol, using a high-pass cutoff of 10-15 kb to target an average library size of 17-20 kb.

Fiber-seq on CHM1 cells was performed in a similar manner, and CHM1 cells were similarly grown in AmnioMax C-100 Basal Medium (Invitrogen 12558011), which includes the AmnioMAX™ C-100 Supplement. Fiber-seq on GM24385 and HLE cells were performed in a similar manner with the exception that cells were grown in the following conditions. GM24385 cells were gown in RPMI-1640 supplemented with 2mM L-Glutamine, 15% fetal bovine serum (FBS) and 1% Penicillin/Streptomycin. HLE cells were grown in RPMI-1640 supplemented with 10% FBS, 1% Pen/Strep, 1% l-glutamine, 1% non-essential amino acids, and 1% sodium pyruvate.

Mapping single-molecule chromatin Features

For parallelization, subreads were split into multiple files while keeping ZMWs together with *bamsieve* (v0.2.0). Subreads were generated into CCS reads with the *ccs –hifi-kinetics* (v6.4.0) command. CCS reads were then aligned to the appropriate genome with *pbmm2 -N 1* (v1.9.0). Fiber-seq reads from CHM13 cells were aligned to CHM13 v1.0. Fiber-seq reads from CHM1 cells was aligned to chm1_cens_v21.fa. Fiber-seq reads from GM24385/HG002 were aligned to a custom genome containing the Verkko-assembled autosomes and X chromosome from GM24385/HG002[54], as well as the Y chromosome from CHM13 v2.0, which is the fully assembled GM24385/HG002 Y chromosome[10]. Fiber-seq reads from the eastern hoolock gibbon (*Hoolock leuconedys*) lymphoblastoid cell line (HLE) were aligned to a de novo genome assembly produced from that cell line as described below. Subreads were then aligned to CCS reads with *actc* (v0.2.0). CCS reads were converted into fasta format with *samtools fasta* (v1.15.1). *ipdSummary* (v 3.0) was then run with the fasta formatted CCS reads as the reference genome and the following flags:*--pvalue .001 --identify m6A.*

We then trained a Gaussian Mixture Model (GMM) with *per_molecule_train_gmm.v3.py* using the output .csv from *ipdSummary,* in a manner similar to Abdulhay et. al.. All ipdRatio values associated with adenines are used to train a molecule-specific 2-state GMM, and that GMM is subsequently applied to the molecule to determine methylated adenines. To generate the GMM we use the GMM module from *sklearn* with the following parameters: *n_components = 2, n_init = 3, max_iter = 500, covariance_type = full, tol = 1E-5.* The CCS fibers with m6A calls were then converted to bed12 format, where the m6A location is encoded in the blockSize/blockStart field. The blockSize/blockStart fields have "1" appended to both ends of the arrays to denote the end of the molecule when visualizing on the UCSC genome browser.

To identify CpG methylation status on a per-fiber basis we used *primrose* (v.1.3.0) *--min-passes 3* with the aligned CCS bam file as input. The per-molecule CpG information was converted to bed12 format in the same manner above.

Identifying single-molecule chromatin features

To identify nucleosomes along a single chromatin fiber we built and applied a Hidden Markov Model (HMM) in conjunction with a simple nucleosome caller approach.

For each sample, we first train an HMM using the *trainHMM.py* script. Briefly, this script takes in single-molecule m6A calls in bed12 format, a genome fasta file, and the number of fibers to train the HMM. The chromatin fibers are subsampled to the specified number of fibers without replacement, then the m6A content of each fiber is encoded in a binary format in A/T space, using the alignment to the reference genome to define the sequence content of each fiber. Following this, a HMM is trained using the binarized m6A fiber with the *pomegranate* python library using the Discrete Distribution(*n_components = 2, max_iterations = 250, n_init = 10*). The trained HMM was then output in JSON format.

Following training, we applied the sample-specific HMM to the entire dataset with the *applyHMM.py* script. First, each fiber was converted into A/T space and then a simple nucleosome caller was run, identifying all m6A-free patches longer than 65 bp. Following this, the HMM was applied to the same fiber. We prioritize simple-called nucleosomes over HMM identified nucleosomes. HMM identified nucleosomes are only kept if they are flanked by regions not identified by the simple nucleosome caller. The ends of the HMM called nucleosomes are adjusted to abut the nearest m6A call.

To identify methyltransferase free patches (MSPs), we take the inverse of the nucleosome calls along each fiber with the *invertBed.py* script.

Identifying the centromere core/centromere dip region

For CHM13, annotations of various satellite DNA were downloaded from the UCSC T2T hub (http://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.0/cenSatAnnotation.bigBed), and putative kinetochore bindings regions for CHM13 were searched for within 'hor' satellite regions using CpG methylation data that was similarly downloaded from the UCSC T2T hub (http://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.0/methyFreq.bigWig). CHM1 annotations were generated as previously described (Logsdon et al., In preparation). For CHM13 and CHM1, alpha satellite repeats less than 25 bp apart were merged together and we only kept alpha satellite stretches >= 100kb. We then made windows of 1190 bp across this region with step size of 170bp (approximate alpha satellite repeat length). We then took the mean CpG methylation frequency across all CpG sites that fall in the window. We identified all windows that were below the 35th percentile of CpG methylation window frequency, and merged neighboring windows that passed this threshold. Merged windows that were 1) at least 15kb in length, 2) had a mean CpG methylation window frequency < 20th percentile, and 3) were not located in the terminal 70kb of the alpha satellite repeat region were identified as CDRs. For HG002, CDRs were identified using the same approach, but the threshold to the terminus of the alpha satellite array was lowered to 15kb. For HLE, we used CENPA CUT&RUN enriched regions to identify the centromere core, as we did not know whether to expect this region to be hypo-CpG methylated.

Identification of dinucleosomes and large MSPs

The cutoff for identifying dinucleosomes and large MSPs was 210 and 150, respectively. For each individual fiber, features (Nucleosomes or MSPs) were compared against their respective thresholds, and a percentage value was calculated for each individual chromatin fiber.

Measuring Distance Between Neighboring Accessible Patches

To calculate the observed distance between large MSPs we calculated the distance between large MSP start positions. To calculate the expected distance between large MSPs in the CDR, we first calculated the number of large MSPs fully contained in the CDR and divide by the total number of sequenced bases per CDR to get a large MSP / sequenced bp value. We then multiply the resulting value by the length of the respective CDR to get the expected number of large MSPs per CDR. We divide the length of the CDR by this value to get the expected number of bp between large MSPs within the CDR.

Determining chromatin repeat lengths

To determine the chromatin repeat length of individual chromatin fibers, the per-base m6A pattern on each fiber was converted into a binary vector (m6a=1, no-m6a=0). These vectors were then independently analyzed using the *scipy.signal.periodogram* tool. This results in spectral density estimations across a vector of frequencies, which is the inverse of the nucleosome repeat length in base-pairs. Spectral densities derived from multiple fibers originating from similar genomic loci were combined and displayed.

Delineating CENP-B binding elements and occupancy

CENP-B boxes were identified in each annotation group using the 17-bp CENP-B box NTTCGNNNNANNCGGGN motif. Specifically, genomic regions were scanned with this motif using *fimo* (version 4.11.2)[55] with the *--parse-genomic-coord --qv-thresh --thresh 0.05 --max-stored-scores 1000000* flags. We further filtered motifs to those which contained 2 CpG sites. To delineate the occupancy of these CENP-B boxes, we intersected CENP-B boxes per-fiber methylation calls. Specifically, the m6A methylation profile of every region along every fiber overlapping a CENP-B box was converted into a binary profile (m6a=1, no-m6a=0) and stored in an array. Based on genomic context and/or CpG

methylation contexts these arrays were collapsed to form m6A frequency tracks with a rolling average of 10bp.

To calculate footprint score distributions, we took all fibers fulfilling a filtering requirement (CpG methylation status / genomic region), subsampled to 10% of the fibers 1000 times and collapsed these to generate m6A frequency tracks smoothed by a 10bp window. We used these fibers to calculate a footprint score defined by the m6A frequency in the flanking region divided m6A frequency in the core CENP-B box. The flanking region score was defined as the mean of the smoothed m6A content -35 - -27 bases upstream and +22 - +27 bases downstream of the CENPB box center. The CENP-B core score was defined by the mean of the smoothed m6A frequency in the -7 - +1 positions relative to the center of the CENP-B box.

Quantifying nucleosome positioning surrounding CENP-B boxes

To visualize the association of CENP-B occupancy and surrounding nucleosome positioning we first classified every instance of a chromatin fiber overlapping a CENP-B box by its CpG methylation state (0,1 or 2 mCpGs). For each overlap, we took the closest (Relative to center of CENPB box) boundary of the both the upstream and downstream nucleosome. For each position upstream or downstream of the CENP-B box we then calculated the percentage of upstream or downstream nucleosomes beginning at that position, respectively.

***Hoolock leuconedys* Oxford Nanopore Technologies (ONT) Sequencing**

Cells were collected from a transformed lymphoblastoid cell line (LCL) derived from a female *Hoolock leuconedys* individual (Betty). To isolate high molecular weight (HMW) DNA, a phenol/chloroform/isoamyl alcohol extraction was performed per standard protocols for the isolation of high molecular weight DNA. Library preparation was performed using the Ligation Sequencing Kit (LSK109). HMW DNA was sequenced on the PromethION platform from Oxford Nanopore Technologies using a PromethION R9.4.1 FLOPRO002 flow cell and basecalled using Guppy (v5.0.16).

To isolate ultra-high molecular weight (UHMW) DNA, the Circulomics Nanobind UHMW DNA Extraction for cultured cells (EXT-CLU-001) protocol was followed according to manufacturer's instructions with the Circulomics Nanobind CBB Big DNA Kit (NB-900-001-01). Library preparation was performed using the Circulomics Nanobind Library Prep protocol for ultra-long sequencing (LBP-ULN-001) using the Circulomics Nanobind UL Library Prep Kit (NB-900-601-01) and the Oxford Nanopore Technologies Ultra-Long DNA Sequencing Kit (SQK-ULK001). UHMW DNA was sequencing on the PromethION platform using a R9.4.1 FLOPRO002 flow cell and basecalled using Guppy (v5.0.16).

***Hoolock leuconedys* Dovetail™ Omni-C™ Sequencing**

Roughly 1.5 million cells were collected from the previously described *Hoolock leuconedys* lymphoblastoid cell line and processed according to the Dovetail™ Omni-C™ Proximity Ligation Assay protocol for mammalian samples (v1.4) as written. Lysate quantification was performed using the Qubit® dsDNA HS kit for the Qubit® Fluorometer and the D5000 HS kit for the Agilent TapeStation 2200. 150 bp paired end sequencing was performed on the Illumina NextSeq 550 V2 platform to a depth of ~274M reads.

***Hoolock leuconedys* CUT&RUN Sequencing**

The CUT&RUN Assay Kit (#86652) from Cell Signaling Technology® was used to assess CENP-A protein-DNA interactions with 250,000 cells per condition following manufacturer's instructions. To assess CENPA-DNA interactions, the CENP-A monoclonal antibody (ADI-KAM-CC006-E) was used at a dilution of

1:50. Tri-methyl-histone H3 (Lys4) (C42D8) rabbit monoclonal antibody (#9751) at a dilution of 1:50 was used as a positive control; rabbit (DA1E) monoclonal antibody IgG XP® isotype control (#66362) at a dilution of 1:10 was used as a negative control. Input chromatin samples were sheared to ~100-700 bases using a Covaris S2 sonicator prior to purification. DNA purification was performed using the Cell Signaling® DNA purification with spin columns kit (#14209). DNA concentration was assessed using the Qubit® dsDNA HS kit for the Qubit® Fluorometer and the High Sensitivity D1000 kit for the Agilent TapeStation 2200. CENP-A and Input libraries were prepared using the NEBNext Ultra II DNA Library Prep Kit for Illumina (#E7645S) and sequenced using the Illumina NovaSeq 150 bp paired end settings to a depth of ~15M reads.

***Hoolock leuconedys* genome assembly**

Flye (v2.9)[56] was used to assemble the raw Oxford Nanopore reads using an estimated genome size of 2.9 Gb, the size of the previously assembled *Nomascus leucogenys* genome[57]. Medaka (v1.4.3) was used for long read polishing using default settings and the r941_prom_sup_g507 model.
Publicly available Illumina WGS sequencing reads for the same individual (SAMN12702557)[58] were used to polish the assembly by mapping the reads with Burrow's Wheeler Aligner (v0.7.17) using the bwa mem algorithm and processed using Samtools (v1.7). Short read error correction was performed using Pilon (v1.22) with default parameters. Haplotype redundancies and assembly artefacts based on read coverage were removed using minimap2 (v2.15) and PURGEhaplotigs (v1.0). Omni-C sequences were used to scaffold the assembly with Juicer (v1.6)[59] and 3D-DNA (v180922)[60] following the default protocols outlined by developers. Juicebox with Assembly Tools (v1.11.08) was used for manual review of the produced scaffolds. Gap filling was performed using TGS-GapCloser (v1.0.1). The gap filled assembly was polished with Illumina reads using Pilon (v.1.22) with default parameters. The reformat.sh module of BBMap was used to impose a 3kb limit on the genome.

LASTZ (v1.04.15) was used to align the assembly to the CHM13 v1.1 genome. Resulting alignments were validated according to predicted syntenic regions and large-scale chromosome misassemblies and misorientations were manually corrected using Emboss (6.6.0). To reduce any misassemblies associated with manual curation and scaffolding, pre-scaffolded contigs (the "query") were aligned to the curated assembly (the "reference") and scaffolded using RagTag (v2.1.0). The resulting scaffolded assembly (built from "query" contigs) was gap filled using TGS-GapCloser (v1.0.1). The gap-filled assembly was polished with Illumina reads using Pilon (v.1.22) with default parameters.

***Hoolock leuconedys* centromere annotation**

A two-step approach was used to identify contiguous centromeres for Fiber-seq analysis. First, centromeric regions were identified by enrichment of CENP-A binding. Low quality reads (Phred score < 20, length < 50 nt) and adapters were trimmed from CENP-A and Input CUT&RUN sequences using cutadapt (v3.5) and mapped using Bowtie2 with default parameters (v2.5.0). Peaks were called using MACS (v2.1.2) using the --keep-dup=2 --broad -q 0.01 -g 2.8e9 flags. These peaks were labeled as putative centromere regions. Second, hifiasm was run directly on the Fiber-seq reads using default parameters. Hifiasm contigs were then mapped to the Hoolock leuconedys genome assembly generated using ONT and HiC data above. Putative centromere regions that were fully encompassed within a hifiasm contig that mapped to that region were then considered to be validated as correctly assembled centromeres, as these centromeres were independently assembled using two orthologous technologies. Five centromeres passed these filters and all five were validated by manual review of read coverage and selected for subsequent Fiber-seq analysis. Repeats in the genome were annotated with RepeatMasker (v4.1.2-p1) using the Crossmatch search engine (v1.090518) and a combined gibbon (*Hylobates* sp.) Dfam (v3.6) and Repbase (20181026) repeat library. CpG methylation was detected in the ONT reads

using the Bonito (v0.3.2) -Remora (v1.1.1) basecalling pipeline with default parameters and the dna_r9.4.1_e8_sup@v3.3 model. Resulting files were converted to a bedMethyl file using modbam2bed (v0.6.2).

QUANTIFICATION AND STATISTICAL ANALYSIS

Unless noted otherwise, significance was tested using a Mann-Whitney U test. In box plots, box extends from Q1 to Q3, and whiskers represent the 10th and 90th percentile.

Supplemental information

Supplemental Figure legends

Figure S1. Single-molecule chromatin figure sequencing of human centromeres, related to Figure 1
(A) Bar plot showing the percentage of short-read DNaseI-seq data mapping to euchromatic genomic regions as well as alpha-satellite DNA that are uniquely mapping to these regions.
(B-C) Resolution of GpC methyltransferase-based and non-specific m6A-methyltransferase-based methods for mapping chromatin architectures within the centromere (B) as well as CENP-B occupancy along CENP-B boxes (C).
(D) Histogram of nucleosome footprint sizes within the centromere core, as well as in other genomic regions.
(E) Genomic locus of chromosome 3 centromere showing satellite repeats, bulk CpG methylation, Fiber-seq identified di-nucleosome footprint density, as well as Fiber-seq identified chromatin accessibility density. Below are individual Fiber-seq reads (grey bars) with m6A-modified bases in purple delineating single-molecule chromatin architectures within centromere core and HSat3 heterochromatic regions.

Figure S2. Higher-order chromatin architectures of heterochromatin, related to Figure 2
(A) Genomic loci showing single-molecule chromatin architectures within six separate satellite regions.
(B) Box-and-whisker plots of the chromatin repeat lengths from each of these six distinct satellite regions.

Figure S3. Single-molecule CENP-B occupancy within centromere core, related to Figure 3
(A) Aggregate m6A methylation profile at all CENP-B boxes located within the centromere core. Regions used for calculating footprint score are highlighted, and the calculation of the footprint score is enumerated below.
(B) Box-and-whisker plots showing the footprint scores at CENP-B boxes within the putative kinetochore binding region as a function of bulk mCpG at the CENP-B box, as well as the p-value match of the CENP-B box to the consensus sequence NTTCGNNNNANNCGGGN.
(C) Box-and-whisker plots showing the size of MSPs overlapping CENP-B boxes identified within different genomic regions.
(D) Histogram of the distance between CENP-B boxes within various genomic regions
(E) Directly comparing large MTase-sensitive chromatin patch (MSP) boundary locations at individual CENP-B boxes based on whether the read does or does not contain a mCpG within that CENP-B box. Above is the aggregate MSP boundary location at centromere core CENP-B boxes containing no mCpG. Below is the aggregate difference in this feature between reads with and without mCpG overlapping the CENP-B box.

Figure S4. Chromatin architectures along the entire chromosome Y, related to Figure 4
Genomic locus showing CpG methylation signal along the entire chromosome Y in HG002/GM24385 cells. Below are insets of single-molecule chromatin architectures at a promoter within the euchromatic region of chromosome Y, as well as an HSat3-HSat1B boundary within the heterochromatic q arm of chromosome Y.

Figure S5. Conservation of dichromatin architecture between humans, related to Figure 5
Genomic loci showing CpG methylation signal as well as di-nucleosome density and chromatin accessibility along the chromosome 12 and 15 centromeres in both CHM13 and CHM1 cells.

Figure S6. Alpha-satellite DNA is not necessary for dichromatin formation, related to Figure 6

(A-D) Genomic loci showing repeats, CpG methylation, CENP-A CUT&RUN, and Fiber-seq derived di-nucleosome density and chromatin accessibility along four assembled centromeres from a lymphoblastoid cell line from the eastern hoolock gibbon (*Hoolock leuconedys*) Betty. Below inset showing single-molecule chromatin architectures within the centromere core region, a heterochromatic region, and a gene regulatory element along chromosome 1.

(E) Swarm and box-and-whisker plots showing the distance between accessible chromatin patches along the same molecule of DNA within the centromere core, as well as the expected distance based on the density of accessible chromatin patches within each chromosome's centromere core (* p-value <0.01 Mann-Whitney).

References

1. Rosenberg, H., Singer, M., and Rosenberg, M. (1978). Highly reiterated sequences of SIMIANSIMIANSIMIANSIMIANSIMIAN. Science (1979) *200*, 394–402. 10.1126/science.205944.

2. Altemose, N., Logsdon, G.A., Bzikadze, A. V., Sidhwani, P., Langley, S.A., Caldas, G. V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J., et al. (2022). Complete genomic and epigenetic maps of human centromeres. Science *376*. 10.1126/SCIENCE.ABL4178.

3. Muro, Y., Masumoto, H., Yoda, K., Nozaki, N., Ohashi, M., and Okazaki, T. (1992). Centromere protein B assembles human centromeric α-satellite DNA at the 17-bp sequence, CENP-B box. Journal of Cell Biology *116*, 585–596. 10.1083/jcb.116.3.585.

4. 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. Journal of Cell Biology *109*, 1963–1973. 10.1083/jcb.109.5.1963.

5. Palmer, D.K., O'Day, K., Trong, H.L.E., Charbonneau, H., and Margolis, R.L. (1991). Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. Proc Natl Acad Sci U S A *88*, 3734–3738. 10.1073/pnas.88.9.3734.

6. Warburton, P.E., Cooke, C.A., Bourassa, S., Vafa, O., Sullivan, B.A., Stetten, G., Gimelli, G., Warburton, D., Tyler-Smith, C., Sullivan, K.F., et al. (1997). Immunolocalization of CENP-A suggests a distinct nucleosome structure at the inner kinetochore plate of active centromeres. Current Biology *7*, 901–904. 10.1016/S0960-9822(06)00382-4.

7. Fachinetti, D., Han, J.S., McMahon, M.A., Ly, P., Abdullah, A., Wong, A.J., and Cleveland, D.W. (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.

8. Zhang, W., Lee, H.R., Koo, D.H., and Jiang, J. (2008). Epigenetic modification of centromeric chromatin: Hypomethylation of DNA sequences in the CENH3-associated chromatin in Arabidopsis thaliana and maize. Plant Cell *20*, 25–34. 10.1105/tpc.107.057083.

9. Logsdon, G.A., Vollger, M.R., Hsieh, P., Mao, Y., Liskovykh, M.A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P.C., Rhie, A., et al. (2021). The structure, function and evolution of a complete human chromosome 8. Nature 2021 593:7857 *593*, 101–107. 10.1038/s41586-021-03420-7.

10. Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., Hook, P.W., Koren, S., Rautiainen, M., Alexandrov, I.A., et al. (2022). The complete sequence of a human Y chromosome. bioRxiv, 2022.12.01.518724. 10.1101/2022.12.01.518724.

11. Sullivan, B.A., and Karpen, G.H. (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.

12. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837. 10.1016/j.cell.2007.05.009.

13. Gershman, A., Sauria, M.E.G., Guitart, X., Vollger, M.R., Hook, P.W., Hoyt, S.J., Jain, M., Shumate, A., Razaghi, R., Koren, S., et al. (2022). Epigenetic patterns in a complete human genome. Science *376*. 10.1126/SCIENCE.ABJ5089.

14. Panchenko, T., Sorensen, T.C., Woodcock, C.L., Kan, Z.Y., Wood, S., Resch, M.G., Luger, K., Englander, S.W., Hansen, J.C., and Black, B.E. (2011). Replacement of histone H3 with CENP-A directs global nucleosome array condensation and loosening of nucleosome superhelical termini. Proc Natl Acad Sci U S A *108*, 16588–16593. 10.1073/PNAS.1113621108/-/DCSUPPLEMENTAL/PNAS.1113621108_SI.PDF.

15. Thakur, J., and Henikoff, S. (2016). CENPT bridges adjacent CENPA nucleosomes on young human α-satellite dimmers. Genome Res *26*, 1178–1187. 10.1101/gr.204784.116.

16. Ando, S., Yang, H., Nozaki, N., Okazaki, T., and Yoda, K. (2002). CENP-A, -B, and -C Chromatin Complex That Contains the I-Type α-Satellite Array Constitutes the Prekinetochore in HeLa Cells. Mol Cell Biol *22*, 2229–2241. 10.1128/mcb.22.7.2229-2241.2002.

17. Zhou, K., Gebala, M., Woods, D., Sundararajan, K., Edwards, G., Krzizike, D., Wereszczynski, J., Straight, A.F., and Luger, K. (2022). CENP-N promotes the compaction of centromeric chromatin. Nat Struct Mol Biol *29*, 403–413. 10.1038/S41594-022-00758-Y.

18. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53. 10.1126/SCIENCE.ABJ6987.

19. Stergachis, A.B., Debo, B.M., Haugen, E., Churchman, L.S., and Stamatoyannopoulos, J.A. (2020). Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science *368*, 1449–1454. 10.1126/science.aaz1646.

20. Lee, I., Razaghi, R., Gilpatrick, T., Molnar, M., Gershman, A., Sadowski, N., Sedlazeck, F.J., Hansen, K.D., Simpson, J.T., and Timp, W. (2020). Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat Methods *17*, 1191–1199. 10.1038/s41592-020-01000-7.

21. Abdulhay, N.J., McNally, C.P., Hsieh, L.J., Kasinathan, S., Keith, A., Estes, L.S., Karimzadeh, M., Underwood, J.G., Goodarzi, H., Narlikar, G.J., et al. (2020). Massively multiplex single-molecule oligonucleosome footprinting. Elife *9*, 1–23. 10.7554/ELIFE.59404.

22. Shipony, Z., Marinov, G.K., Swaffer, M.P., Sinnott-Armstrong, N.A., Skotheim, J.M., Kundaje, A., and Greenleaf, W.J. (2020). Long-range single-molecule mapping of chromatin accessibility in eukaryotes. Nat Methods, 1–9. 10.1038/s41592-019-0730-2.

23. Drozdz, M., Piekarowicz, A., Bujnicki, J.M., and Radlinska, M. (2012). Novel non-specific DNA adenine methyltransferases. Nucleic Acids Res *40*, 2119–2130. 10.1093/nar/gkr1039.

24. Luo, G.Z., and He, C. (2017). DNA N 6 -methyladenine in metazoans: Functional epigenetic mark or bystander? Nat Struct Mol Biol *24*, 503–506. 10.1038/nsmb.3412.

25. Kong, Y., Cao, L., Deikus, G., Fan, Y., Mead, E.A., Lai, W., Zhang, Y., Yong, R., Sebra, R., Wang, H., et al. (2022). Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. Science *375*, 515–522. 10.1126/SCIENCE.ABE7489.

26. Steinberg, K.M., Schneider, V.A., Graves-Lindsay, T.A., Fulton, R.S., Agarwala, R., Huddleston, J., Shiryev, S.A., Morgulis, A., Surti, U., Warren, W.C., et al. (2014). Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res *24*, 2066–2076. 10.1101/gr.180893.114.

27. Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., Project, 1000 Genomes, et al. (2010). Diversity of Human Copy Number Variation and Multicopy Genes. Science (1979) *330*, 641–646. 10.1126/science.1197005.

28. Tachiwana, H., Kagawa, W., Shiga, T., Osakabe, A., Miya, Y., Saito, K., Hayashi-Takanaka, Y., Oda, T., Sato, M., Park, S.Y., et al. (2011). Crystal structure of the human centromeric nucleosome containing CENP-A. Nature *476*, 232–235. 10.1038/NATURE10258.

29. Hasson, D., Panchenko, T., Salimian, K.J., Salman, M.U., Sekulic, N., Alonso, A., Warburton, P.E., and Black, B.E. (2013). The octamer is the major form of CENP-A nucleosomes at human centromeres. Nat Struct Mol Biol *20*, 687–695. 10.1038/NSMB.2562.

30. Canzio, D., Chang, E.Y., Shankar, S., Kuchenbecker, K.M., Simon, M.D., Madhani, H.D., Narlikar, G.J., and Al-Sady, B. (2011). Chromodomain-mediated oligomerization of HP1 suggests a nucleosome-bridging mechanism for heterochromatin assembly. Mol Cell *41*, 67–81. 10.1016/j.molcel.2010.12.016.

31. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82. 10.1038/nature11232.

32. Kiewisz, R., Fabig, G., Conway, W., Baum, D., Needleman, D., and Müller-Reichert, T. (2022). Three-dimensional structure of kinetochore-fibers in human mitotic spindles. Elife *11*. 10.7554/ELIFE.75459.

33. O'Toole, E., Morphew, M., and Richard McIntosh, J. (2020). Electron tomography reveals aspects of spindle structure important for mechanical stability at metaphase. Mol Biol Cell *31*, 184–195. 10.1091/MBC.E19-07-0405.

34. Dumont, M., Gamba, R., Gestraud, P., Klaasen, S., Worrall, J.T., de Vries, S.G., Boudreau, V., Salinas-Luypaert, C., Maddox, P.S., Lens, S.M., et al. (2020). Human chromosome-specific aneuploidy is influenced by DNA-dependent centromeric features. EMBO J *39*. 10.15252/EMBJ.2019102924.

35. Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and DNA. Science *184*, 868–871. 10.1126/SCIENCE.184.4139.868.

36. Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *474*, 516–522. 10.1038/NATURE10002.

37. Vafa, O., and Sullivan, K.F. (1997). Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. Curr Biol *7*, 897–900. 10.1016/S0960-9822(06)00381-2.

38. Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90. 10.1038/nature11212.

39. Mirny, L.A. (2010). Nucleosome-mediated cooperativity between transcription factors. Proc Natl Acad Sci U S A *107*, 22534–22539. 10.1073/pnas.0913805107.

40. Yoda, K., Ando, S., Okuda, A., Kikuchi, A., and Okazaki, T. (1998). In vitro assembly of the CENP-B/alpha-satellite DNA/core histone complex: CENP-B causes nucleosome positioning. Genes to cells *3*, 533–548. 10.1046/J.1365-2443.1998.00210.X.

41. Tanaka, Y., Nureki, O., Kurumizaka, H., Fukai, S., Kawaguchi, S., Ikuta, M., Iwahara, J., Okazaki, T., and Yokoyama, S. (2001). Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA. EMBO J *20*, 6612–6618. 10.1093/EMBOJ/20.23.6612.

42. Tanaka, Y., Kurumizaka, H., and Yokoyama, S. (2005). CpG methylation of the CENP-B box reduces human CENP-B binding. FEBS Journal *272*, 282–289. 10.1111/j.1432-1033.2004.04406.x.

43. Okada, T., Ohzeki, J. ichirou, Nakano, M., Yoda, K., Brinkley, W.R., Larionov, V., and Masumoto, H. (2007). CENP-B Controls Centromere Formation Depending on the Chromatin Context. Cell *131*, 1287–1300. 10.1016/j.cell.2007.10.045.

44. Ikeno, M., Masumoto, H., and Okazaki, T. (1994). Distribution of CENP-B boxes reflected in CREST centromere antigenic sites on long-range α-satellite DNA arrays of human chromosome 21. Hum Mol Genet *3*, 1245–1257. 10.1093/hmg/3.8.1245.

45. Tanaka, Y., Tachiwana, H., Yoda, K., Masumoto, H., Okazaki, T., Kurumizaka, H., and Yokoyama, S. (2005). Human centromere protein B induces translational positioning of nucleosomes on α-satellite sequences. Journal of Biological Chemistry *280*, 41609–41618. 10.1074/jbc.M509666200.

46. Earnshaw, W.C., Sullivan, K.F., Machlin, P.S., Cooke, C.A., Kaiser, D.A., Pollard, T.D., Rothfield, N.F., and Cleveland, D.W. (1987). Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen. J Cell Biol *104*, 817–829. 10.1083/JCB.104.4.817.

47. Capozzi, O., Carbone, L., Stanyon, R.R., Marra, A., Yang, F., Whelan, C.W., de Jong, P.J., Rocchi, M., and Archidiacono, N. (2012). A comprehensive molecular cytogenetic analysis of chromosome rearrangements in gibbons. Genome Res *22*, 2520–2528. 10.1101/GR.138651.112.

48. Hartley, G.A., Okhovat, M., O'Neill, R.J., and Carbone, L. (2021). Comparative Analyses of Gibbon Centromeres Reveal Dynamic Genus-Specific Shifts in Repeat Composition. Mol Biol Evol *38*, 3972–3992. 10.1093/MOLBEV/MSAB148.

49. Walfridsson, J., Bjerling, P., Thalen, M., Yoo, E.J., Park, S.D., and Ekwall, K. (2005). The CHD remodeling factor Hrp1 stimulates CENP-A loading to centromeres. Nucleic Acids Res *33*, 2868–2879. 10.1093/NAR/GKI579.

50. Tawaramoto, M.S., Park, S.Y., Tanaka, Y., Nureki, O., Kurumizaka, H., and Yokoyama, S. (2003). Crystal structure of the human centromere protein B (CENP-B) dimerization domain at 1.65-A resolution. J Biol Chem *278*, 51454–51461. 10.1074/JBC.M310388200.

51. Chardon, F., Japaridze, A., Witt, H., Velikovsky, L., Chakraborty, C., Wilhelm, T., Dumont, M., Yang, W., Kikuti, C., Gangnard, S., et al. (2022). CENP-B-mediated DNA loops regulate activity and stability of human centromeres. Mol Cell *82*, 1751-1767.e8. 10.1016/J.MOLCEL.2022.02.032.

52. Henikoff, S., Ahmad, K., and Malik, H.S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. Science *293*, 1098–1102. 10.1126/SCIENCE.1062939.

53. Logsdon, G.A., and Eichler, E.E. (2022). The Dynamic Structure and Rapid Evolution of Human Centromeric Satellite DNA. Genes (Basel) *14*, 92. 10.3390/GENES14010092.

54. Rautiainen, M., Nurk, S., Walenz, B.P., Logsdon, G.A., Porubsky, D., Rhie, A., Eichler, E.E., Phillippy, A.M., and Koren, S. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. 10.1038/S41587-023-01662-6.

55. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: Scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018. 10.1093/bioinformatics/btr064.

56. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol *37*, 540–546. 10.1038/S41587-019-0072-8.

57. Carbone, L., Alan Harris, R., Gnerre, S., Veeramah, K.R., Lorente-Galdos, B., Huddleston, J., Meyer, T.J., Herrero, J., Roos, C., Aken, B., et al. (2014). Gibbon genome and the fast karyotype evolution of small apes. Nature *513*, 195–201. 10.1038/NATURE13679.

58. Okhovat, M., Nevonen, K.A., Davis, B.A., Michener, P., Ward, S., Milhaven, M., Harshman, L., Sohota, A., Fernandes, J.D., Salama, S.R., et al. (2020). Co-option of the lineage-specific LAVA retrotransposon in the gibbon genome. Proc Natl Acad Sci U S A *117*, 19328–19338. 10.1073/PNAS.2006038117/-/DCSUPPLEMENTAL.

59. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst *3*, 95–98. 10.1016/J.CELS.2016.07.002.

60. Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science *356*, 92–95. 10.1126/SCIENCE.AAL3327.

Figure 1

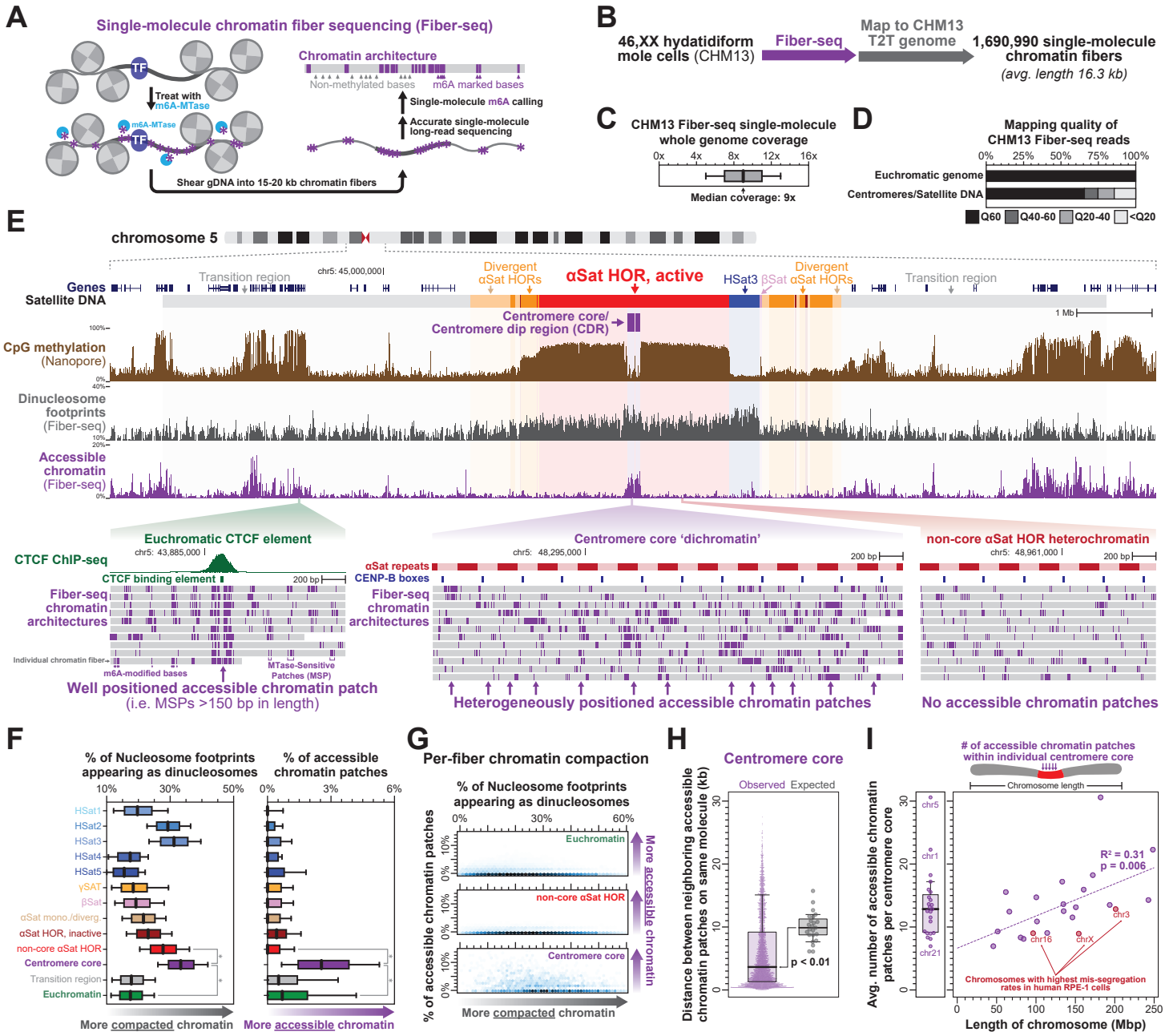

A
Single-molecule chromatin fiber sequencing (Fiber-seq)
Treat with m6A-MTase
Shear gDNA into 15-20 kb chromatin fibers
Chromatin architecture
Non-methylated bases
m6A marked bases
Single-molecule m6A calling
Accurate single-molecule long-read sequencing
B
46,XX hydatidiform mole cells (CHM13)
Fiber-seq
Map to CHM13 T2T genome
1,690,990 single-molecule chromatin fibers
(avg. length 16.3 kb)
C
CHM13 Fiber-seq single-molecule whole genome coverage
Median coverage: 9x
D
Mapping quality of CHM13 Fiber-seq reads
Euchromatic genome
Centromeres/Satellite DNA
Q60
Q40-60
Q20-40
<Q20
E
chromosome 5
Genes
Satellite DNA
Transition region
Divergent αSat HORs
αSat HOR, active
HSat3
βSat
Centromere core/ Centromere dip region (CDR)
CpG methylation (Nanopore)
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
Euchromatic CTCF element
CTCF ChIP-seq
CTCF binding element
Fiber-seq chromatin architectures
Individual chromatin fiber
m6A-modified bases
MTase-Sensitive Patches (MSP)
Well positioned accessible chromatin patch
(i.e. MSPs >150 bp in length)
Centromere core 'dichromatin'
αSat repeats
CENP-B boxes
Heterogeneously positioned accessible chromatin patches
non-core αSat HOR heterochromatin
No accessible chromatin patches
F
% of Nucleosome footprints appearing as dinucleosomes
% of accessible chromatin patches
HSat1
HSat2
HSat3
HSat4
HSat5
γSAT
βSat
αSat mono./diverg.
αSat HOR, inactive
non-core αSat HOR
Centromere core
Transition region
Euchromatin
More compacted chromatin
More accessible chromatin
G
Per-fiber chromatin compaction
% of Nucleosome footprints appearing as dinucleosomes
% of accessible chromatin patches
Euchromatin
non-core αSat HOR
Centromere core
More accessible chromatin
More compacted chromatin
H
Centromere core
Observed
Expected
Distance between neighboring accessible chromatin patches on same molecule (kb)
p < 0.01
I
# of accessible chromatin patches within individual centromere core
Chromosome length
Avg. number of accessible chromatin patches per centromere core
R² = 0.31
p = 0.006
Chromosomes with highest mis-segregation rates in human RPE-1 cells
Length of chromosome (Mbp)

Figure 2

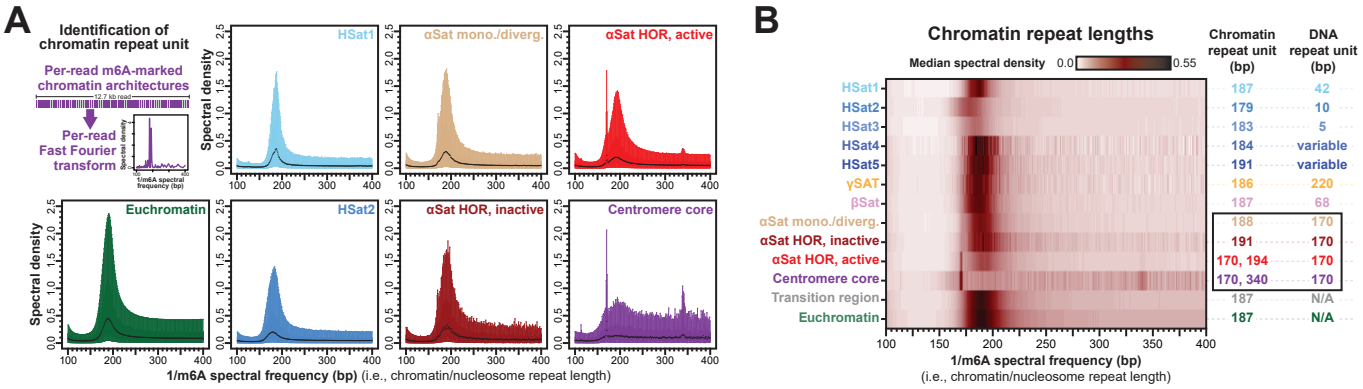

A
Identification of chromatin repeat unit
Per-read m6A-marked chromatin architectures
Per-read Fast Fourier transform
Spectral density
HSat1
αSat mono./diverg.
αSat HOR, active
Euchromatin
HSat2
αSat HOR, inactive
Centromere core
1/m6A spectral frequency (bp) (i.e., chromatin/nucleosome repeat length)
B
Chromatin repeat lengths
Median spectral density 0.0 0.55
Chromatin repeat unit (bp)
DNA repeat unit (bp)
HSat1 187 42
HSat2 179 10
HSat3 183 5
HSat4 184 variable
HSat5 191 variable
γSAT 186 220
βSat 187 68
αSat mono./diverg. 188 170
αSat HOR, inactive 191 170
αSat HOR, active 170, 194 170
Centromere core 170, 340 170
Transition region 187 N/A
Euchromatin 187 N/A
1/m6A spectral frequency (bp)
(i.e., chromatin/nucleosome repeat length)

Figure 3

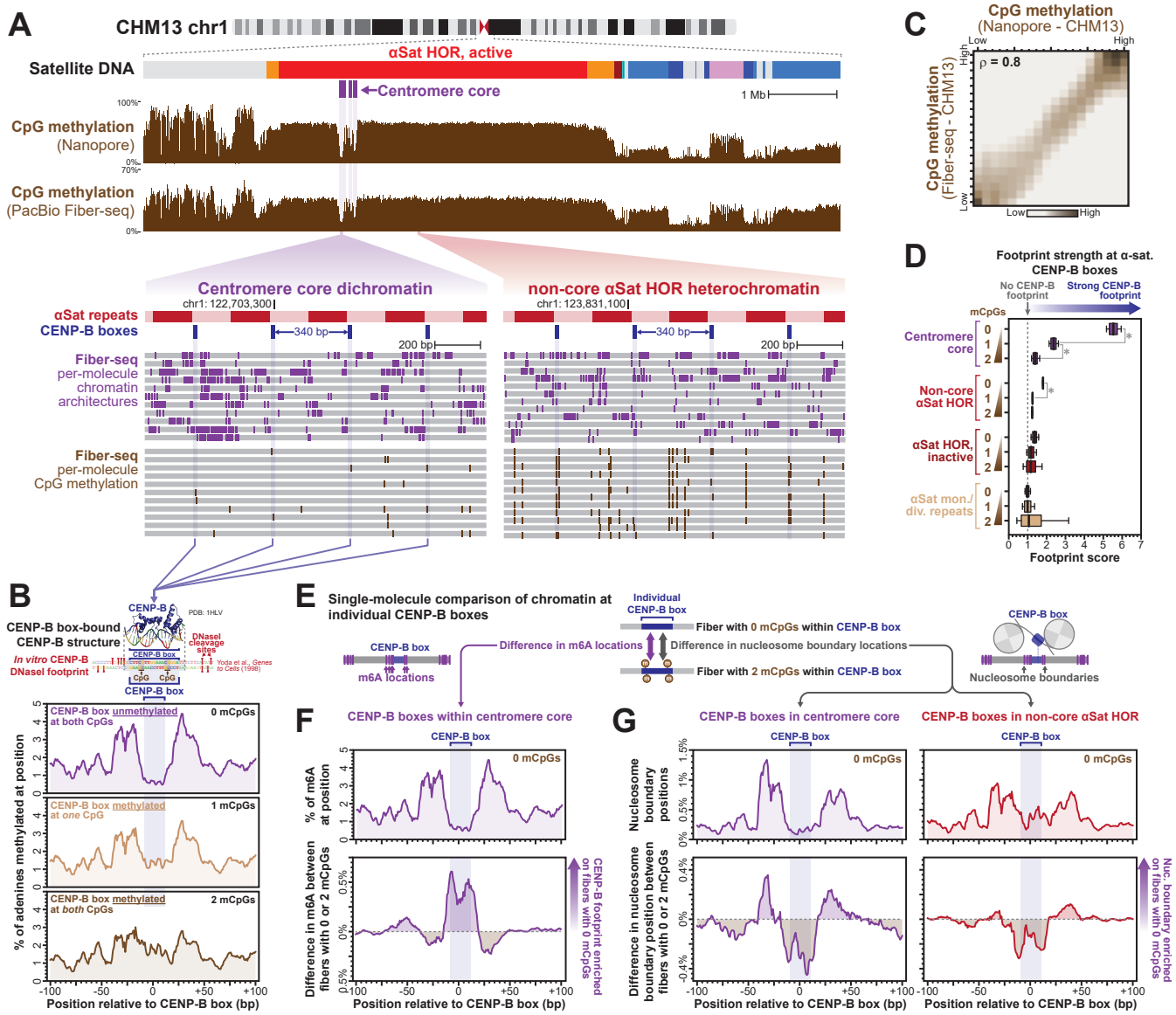

A
CHM13 chr1
Satellite DNA
αSat HOR, active
Centromere core
1 Mb
CpG methylation (Nanopore)
CpG methylation (PacBio Fiber-seq)
Centromere core dichromatin
non-core αSat HOR heterochromatin
chr1: 122,703,300
chr1: 123,831,100
αSat repeats
CENP-B boxes
340 bp
200 bp
Fiber-seq per-molecule chromatin architectures
Fiber-seq per-molecule CpG methylation
B
CENP-B box-bound CENP-B structure
PDB: 1HLV
DNaseI cleavage sites
In vitro CENP-B DNaseI footprint
Yoda et al., Genes to Cells (1998)
CENP-B box
CpG
CENP-B box unmethylated at both CpGs
0 mCpGs
CENP-B box methylated at one CpG
1 mCpGs
CENP-B box methylated at both CpGs
2 mCpGs
% of adenines methylated at position
Position relative to CENP-B box (bp)
C
CpG methylation (Nanopore - CHM13)
CpG methylation (Fiber-seq - CHM13)
ρ = 0.8
Low
High
D
Footprint strength at α-sat. CENP-B boxes
No CENP-B footprint
Strong CENP-B footprint
mCpGs
Centromere core
Non-core αSat HOR
αSat HOR, inactive
αSat mon./ div. repeats
Footprint score
E
Single-molecule comparison of chromatin at individual CENP-B boxes
Individual CENP-B box
Fiber with 0 mCpGs within CENP-B box
Fiber with 2 mCpGs within CENP-B box
Difference in m6A locations
Difference in nucleosome boundary locations
m6A locations
Nucleosome boundaries
F
CENP-B boxes within centromere core
% of m6A at position
Difference in m6A between fibers with 0 or 2 mCpGs
CENP-B footprint enriched on fibers with 0 mCpGs
Position relative to CENP-B box (bp)
G
CENP-B boxes in centromere core
CENP-B boxes in non-core αSat HOR
Nucleosome boundary positions
Difference in nucleosome boundary position between fibers with 0 or 2 mCpGs
Nuc. boundary enriched on fibers with 0 mCpGs
Position relative to CENP-B box (bp)

Figure 4

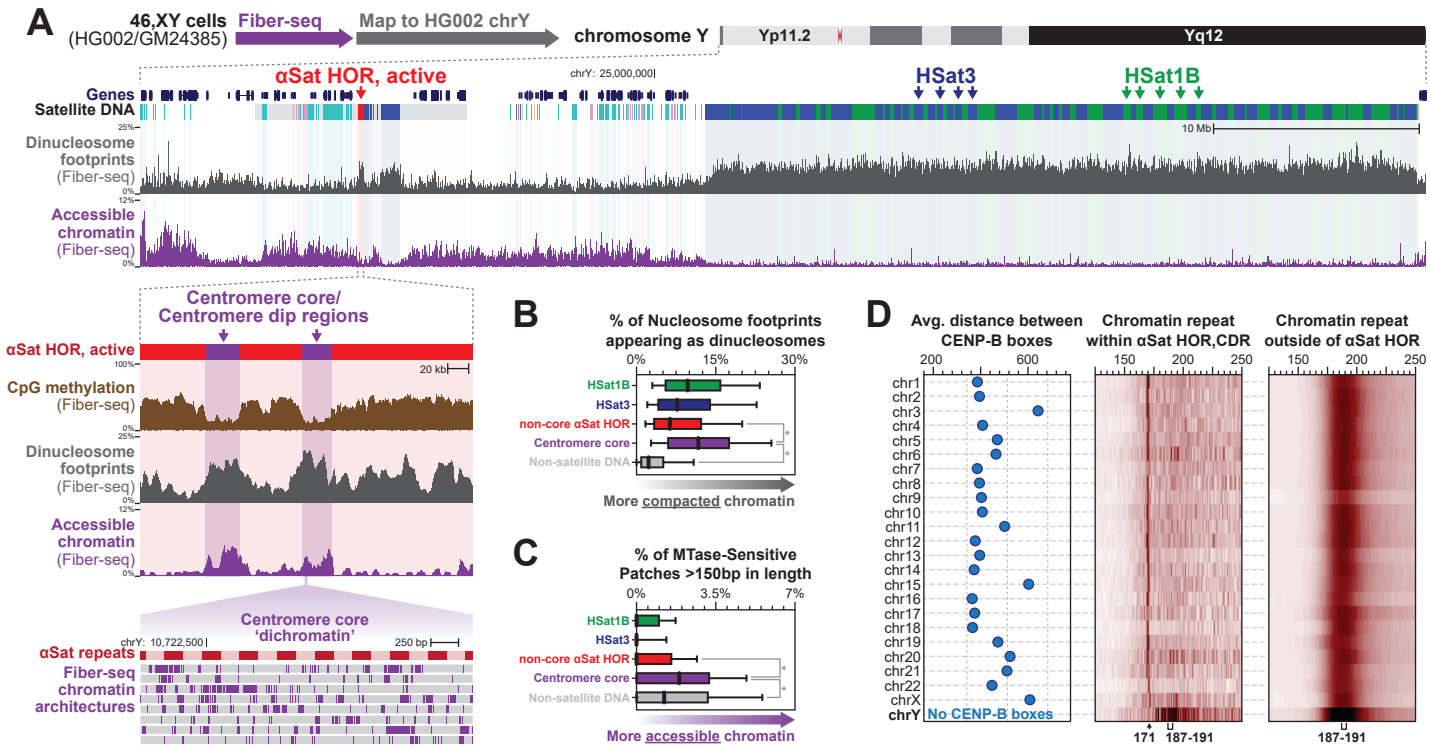

A
46,XY cells
(HG002/GM24385)
Fiber-seq
Map to HG002 chrY
chromosome Y
Yp11.2
Yq12
αSat HOR, active
chrY: 25,000,000
HSat3
HSat1B
Genes
Satellite DNA
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
10 Mb
Centromere core/ Centromere dip regions
αSat HOR, active
CpG methylation (Fiber-seq)
20 kb
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
Centromere core 'dichromatin'
chrY: 10,722,500
250 bp
αSat repeats
Fiber-seq chromatin architectures
B
% of Nucleosome footprints appearing as dinucleosomes
0%
15%
30%
HSat1B
HSat3
non-core αSat HOR
Centromere core
Non-satellite DNA
More compacted chromatin
C
% of MTase-Sensitive Patches >150bp in length
0%
3.5%
7%
HSat1B
HSat3
non-core αSat HOR
Centromere core
Non-satellite DNA
More accessible chromatin
D
Avg. distance between CENP-B boxes
200
600
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
No CENP-B boxes
Chromatin repeat within αSat HOR,CDR
150
200
250
171
187-191
Chromatin repeat outside of αSat HOR
150
200
250
187-191

Figure 5

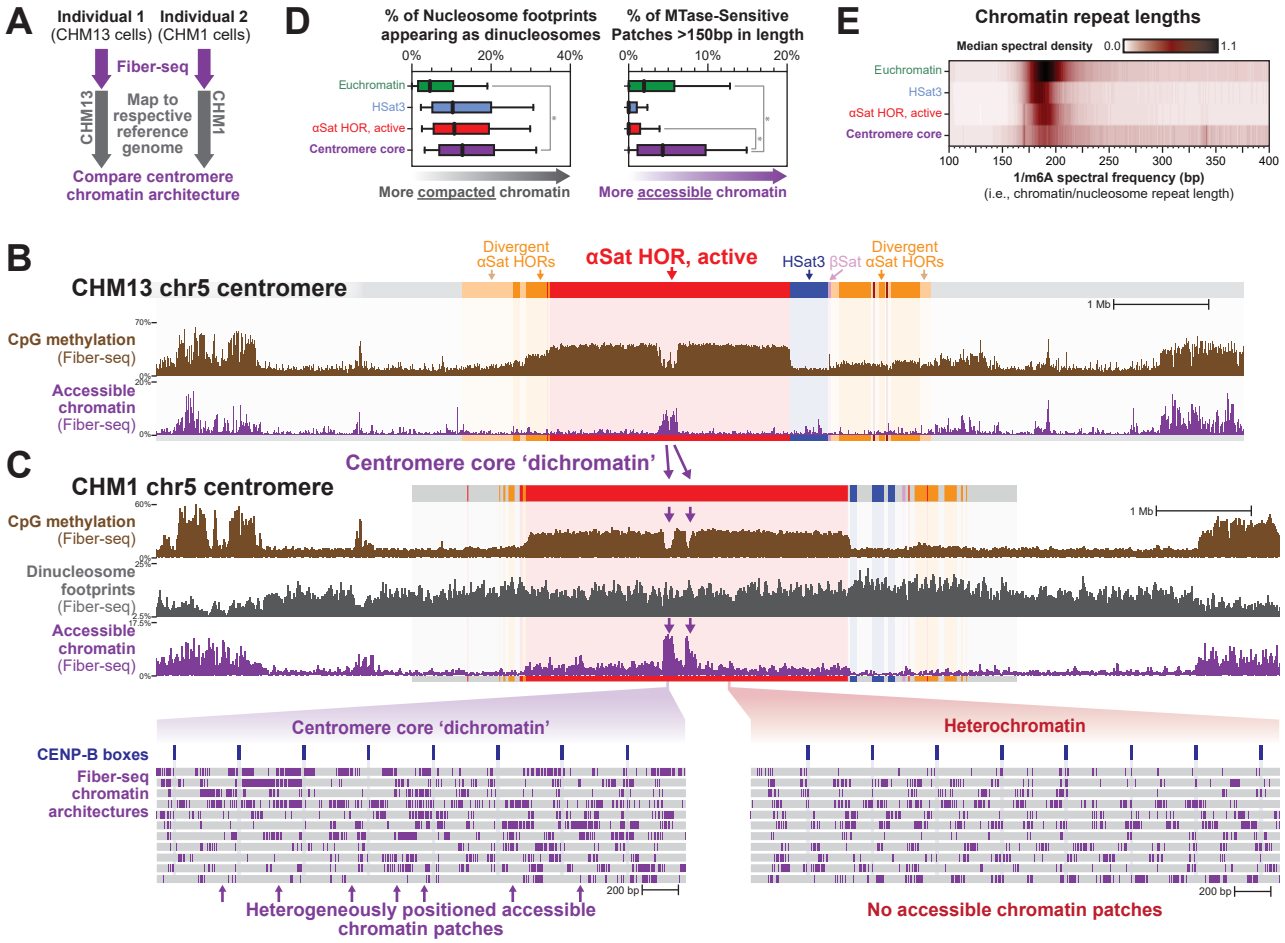

A
Individual 1 (CHM13 cells)
Individual 2 (CHM1 cells)
Fiber-seq
CHM13
Map to respective reference genome
CHM1
Compare centromere chromatin architecture
D
% of Nucleosome footprints appearing as dinucleosomes
% of MTase-Sensitive Patches >150bp in length
Euchromatin
HSat3
αSat HOR, active
Centromere core
More compacted chromatin
More accessible chromatin
E
Chromatin repeat lengths
Median spectral density
1/m6A spectral frequency (bp)
(i.e., chromatin/nucleosome repeat length)
B
CHM13 chr5 centromere
Divergent αSat HORs
αSat HOR, active
HSat3
βSat
CpG methylation (Fiber-seq)
Accessible chromatin (Fiber-seq)
1 Mb
C
CHM1 chr5 centromere
Centromere core 'dichromatin'
Dinucleosome footprints (Fiber-seq)
Centromere core 'dichromatin'
Heterochromatin
CENP-B boxes
Fiber-seq chromatin architectures
200 bp
Heterogeneously positioned accessible chromatin patches
No accessible chromatin patches

Figure 6

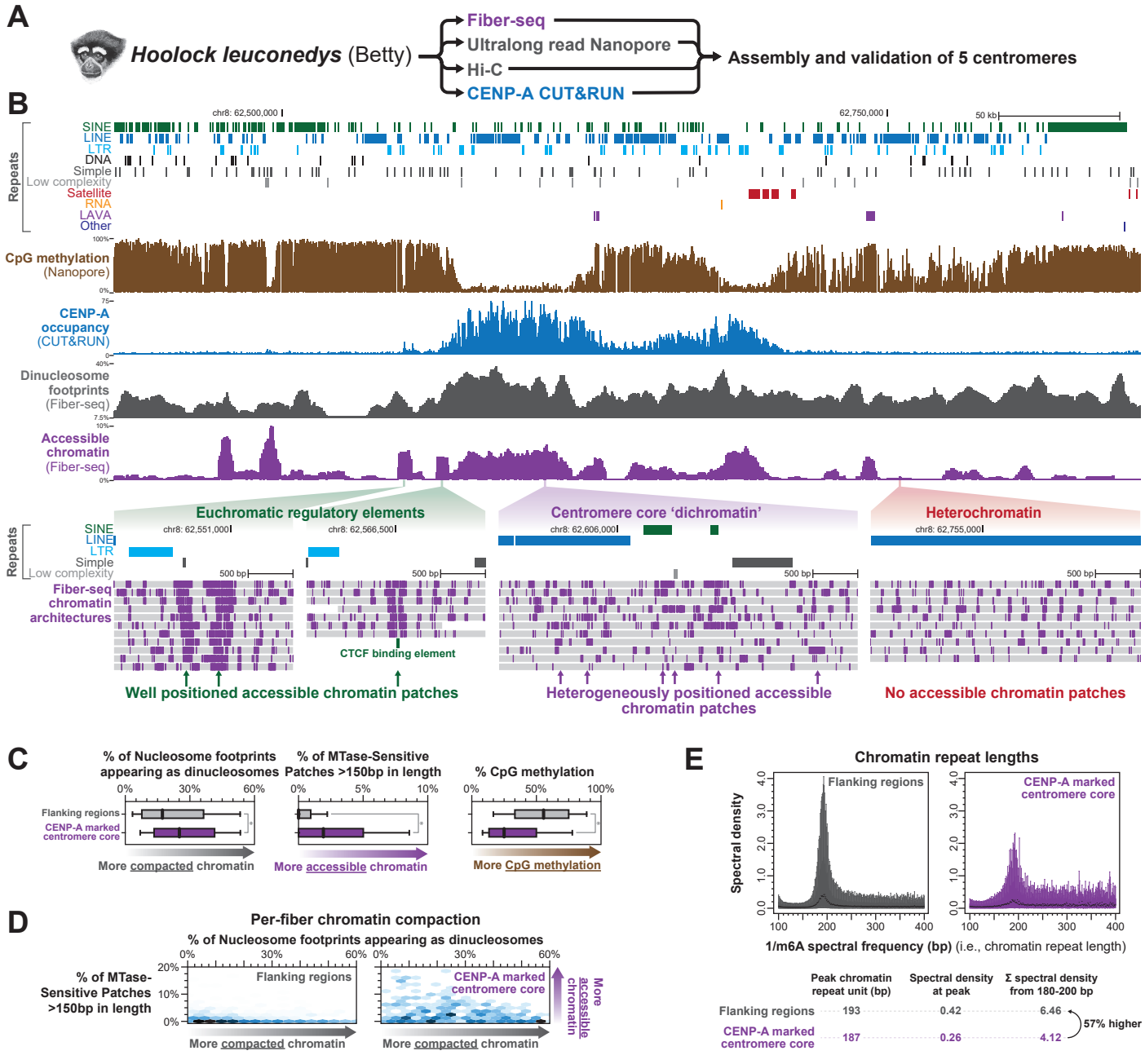

A
Hoolock leuconedys (Betty)
Fiber-seq
Ultralong read Nanopore
Hi-C
CENP-A CUT&RUN
Assembly and validation of 5 centromeres
B
chr8: 62,500,000
62,750,000
50 kb
Repeats
SINE
LINE
LTR
DNA
Simple
Low complexity
Satellite
RNA
LAVA
Other
CpG methylation (Nanopore)
CENP-A occupancy (CUT&RUN)
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
Euchromatic regulatory elements
Centromere core 'dichromatin'
Heterochromatin
chr8: 62,551,000
chr8: 62,566,500
chr8: 62,606,000
chr8: 62,755,000
500 bp
Fiber-seq chromatin architectures
CTCF binding element
Well positioned accessible chromatin patches
Heterogeneously positioned accessible chromatin patches
No accessible chromatin patches
C
% of Nucleosome footprints appearing as dinucleosomes
% of MTase-Sensitive Patches >150bp in length
% CpG methylation
Flanking regions
CENP-A marked centromere core
More compacted chromatin
More accessible chromatin
More CpG methylation
D
Per-fiber chromatin compaction
% of Nucleosome footprints appearing as dinucleosomes
% of MTase-Sensitive Patches >150bp in length
Flanking regions
CENP-A marked centromere core
More compacted chromatin
More accessible chromatin
E
Chromatin repeat lengths
Spectral density
1/m6A spectral frequency (bp) (i.e., chromatin repeat length)
Peak chromatin repeat unit (bp)
Spectral density at peak
Σ spectral density from 180-200 bp
Flanking regions 193 0.42 6.46
CENP-A marked centromere core 187 0.26 4.12
57% higher

Figure S1

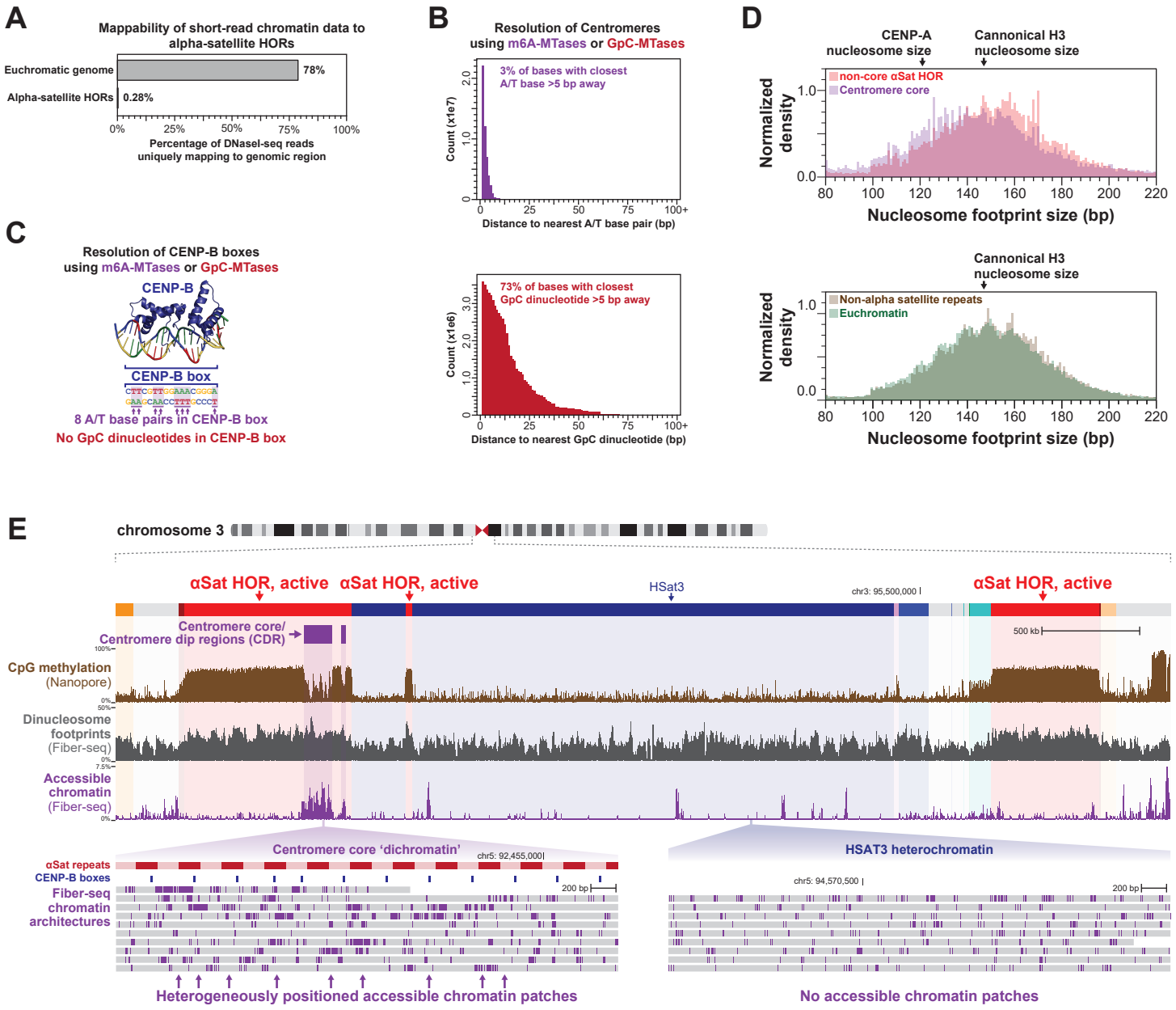

A
Mappability of short-read chromatin data to alpha-satellite HORs
Euchromatic genome
78%
Alpha-satellite HORs
0.28%
0%
25%
50%
75%
100%
Percentage of DNaseI-seq reads uniquely mapping to genomic region
B
Resolution of Centromeres using m6A-MTases or GpC-MTases
3% of bases with closest A/T base >5 bp away
Count (x1e7)
Distance to nearest A/T base pair (bp)
73% of bases with closest GpC dinucleotide >5 bp away
Count (x1e6)
Distance to nearest GpC dinucleotide (bp)
C
Resolution of CENP-B boxes using m6A-MTases or GpC-MTases
CENP-B
CENP-B box
8 A/T base pairs in CENP-B box
No GpC dinucleotides in CENP-B box
D
CENP-A nucleosome size
Cannonical H3 nucleosome size
non-core αSat HOR
Centromere core
Normalized density
Nucleosome footprint size (bp)
Non-alpha satellite repeats
Euchromatin
E
chromosome 3
αSat HOR, active
αSat HOR, active
HSat3
chr3: 95,500,000
αSat HOR, active
Centromere core/ Centromere dip regions (CDR)
500 kb
CpG methylation (Nanopore)
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
Centromere core 'dichromatin'
chr5: 92,455,000
αSat repeats
CENP-B boxes
Fiber-seq chromatin architectures
200 bp
Heterogeneously positioned accessible chromatin patches
HSAT3 heterochromatin
chr5: 94,570,500
200 bp
No accessible chromatin patches

Figure S2

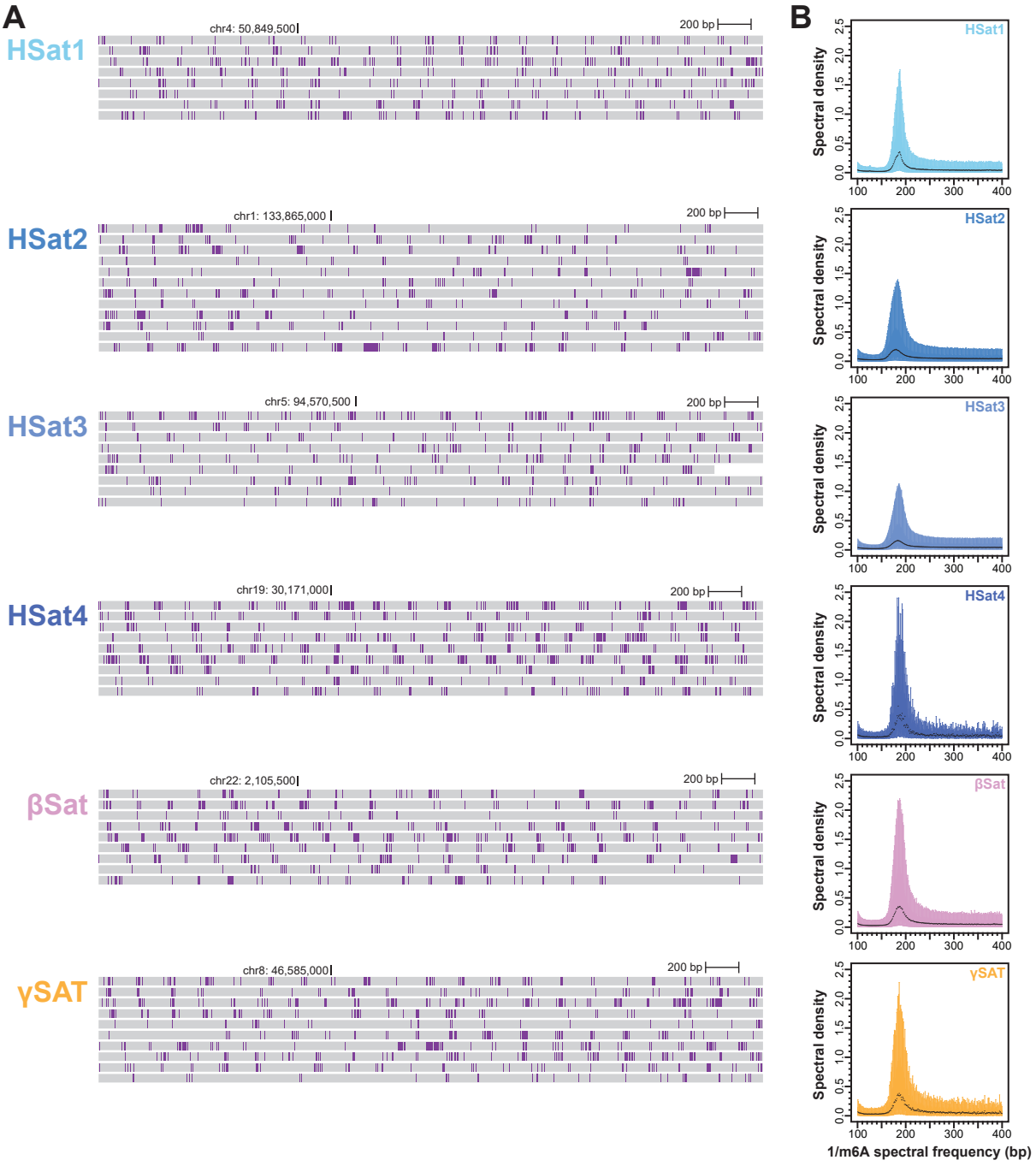

A
HSat1
chr4: 50,849,500
200 bp
HSat2
chr1: 133,865,000
200 bp
HSat3
chr5: 94,570,500
200 bp
HSat4
chr19: 30,171,000
200 bp
βSat
chr22: 2,105,500
200 bp
γSAT
chr8: 46,585,000
200 bp
B
HSat1
HSat2
HSat3
HSat4
βSat
γSAT
Spectral density
0.0
0.5
1.0
1.5
2.0
2.5
100
200
300
400
1/m6A spectral frequency (bp)

Figure S3

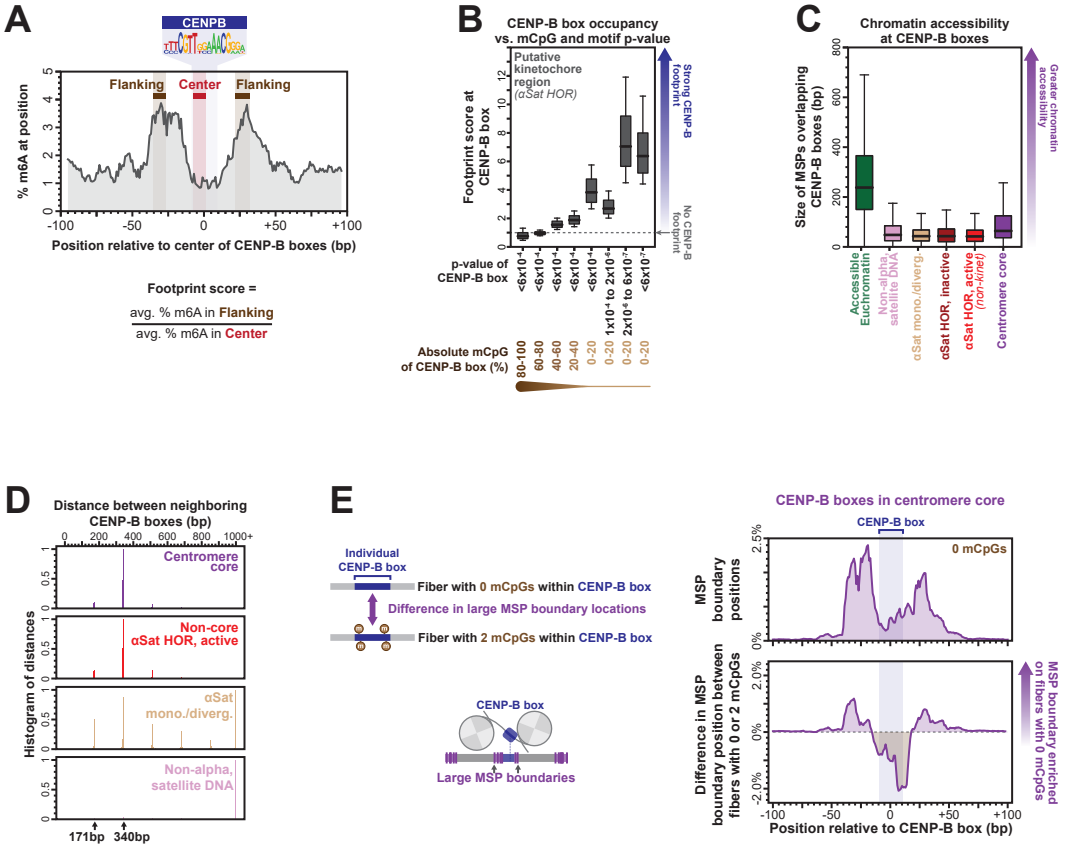

A
CENPB
Flanking
Center
Flanking
% m6A at position
Position relative to center of CENP-B boxes (bp)
Footprint score =
avg. % m6A in Flanking
avg. % m6A in Center
B
CENP-B box occupancy
vs. mCpG and motif p-value
Putative kinetochore region (αSat HOR)
Footprint score at CENP-B box
Strong CENP-B footprint
No CENP-B footprint
p-value of CENP-B box
Absolute mCpG of CENP-B box (%)
C
Chromatin accessibility at CENP-B boxes
Size of MSPs overlapping CENP-B boxes (bp)
Greater chromatin accessibility
Accessible Euchromatin
Non-alpha, satellite DNA
αSat mono./diverg.
αSat HOR, inactive
αSat HOR, active (non-kinet)
Centromere core
D
Distance between neighboring CENP-B boxes (bp)
Histogram of distances
Centromere core
Non-core αSat HOR, active
αSat mono./diverg.
Non-alpha, satellite DNA
171bp
340bp
E
Individual CENP-B box
Fiber with 0 mCpGs within CENP-B box
Difference in large MSP boundary locations
Fiber with 2 mCpGs within CENP-B box
CENP-B box
Large MSP boundaries
CENP-B boxes in centromere core
CENP-B box
0 mCpGs
MSP boundary positions
Difference in MSP boundary position between fibers with 0 or 2 mCpGs
MSP boundary enriched on fibers with 0 mCpGs
Position relative to CENP-B box (bp)

Figure S4

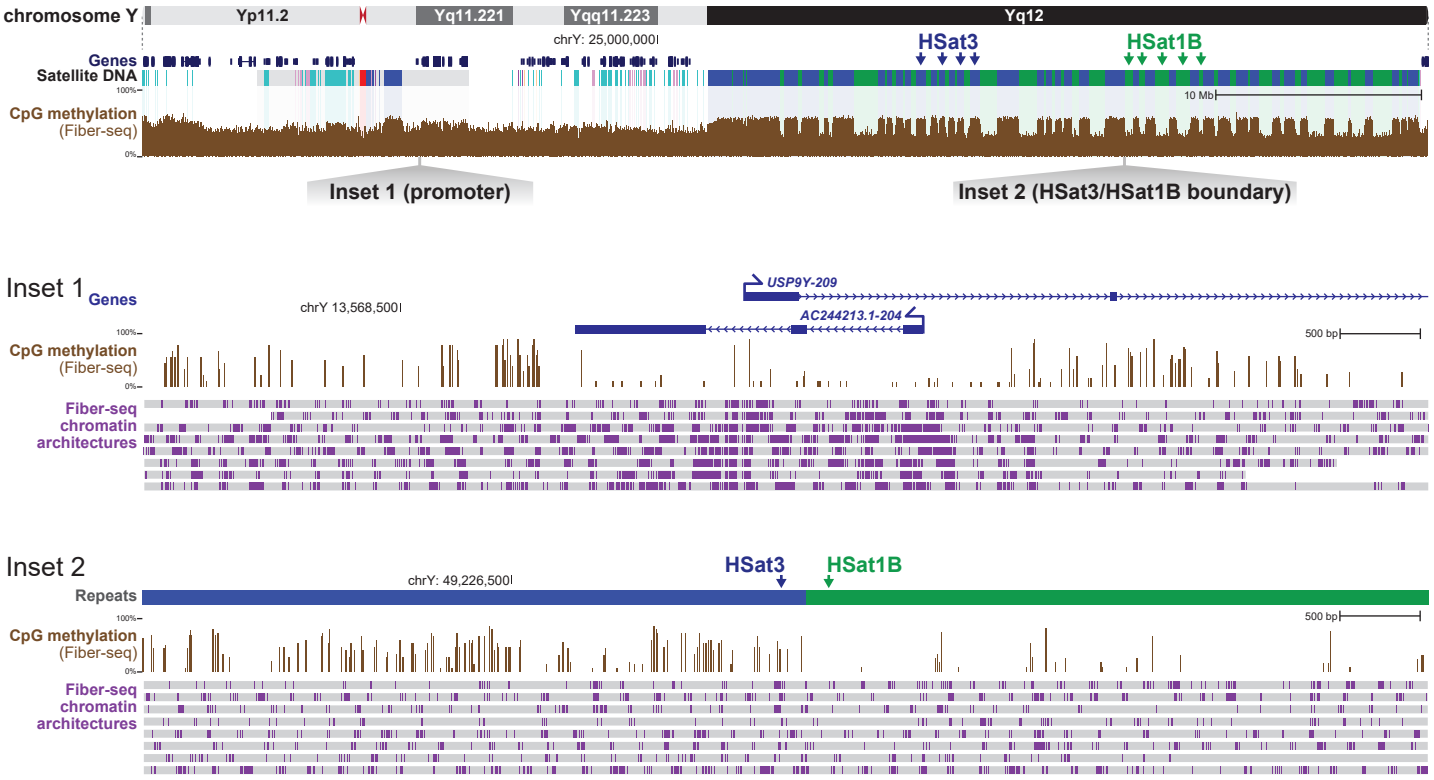

chromosome Y
Yp11.2
Yq11.221
Yqq11.223
Yq12
chrY: 25,000,000
HSat3
HSat1B
Genes
Satellite DNA
100%
10 Mb
CpG methylation
(Fiber-seq)
0%
Inset 1 (promoter)
Inset 2 (HSat3/HSat1B boundary)
Inset 1
Genes
chrY 13,568,500
USP9Y-209
AC244213.1-204
100%
500 bp
CpG methylation
(Fiber-seq)
0%
Fiber-seq
chromatin
architectures
Inset 2
HSat3
HSat1B
chrY: 49,226,500
Repeats
100%
500 bp
CpG methylation
(Fiber-seq)
0%
Fiber-seq
chromatin
architectures

Figure S5

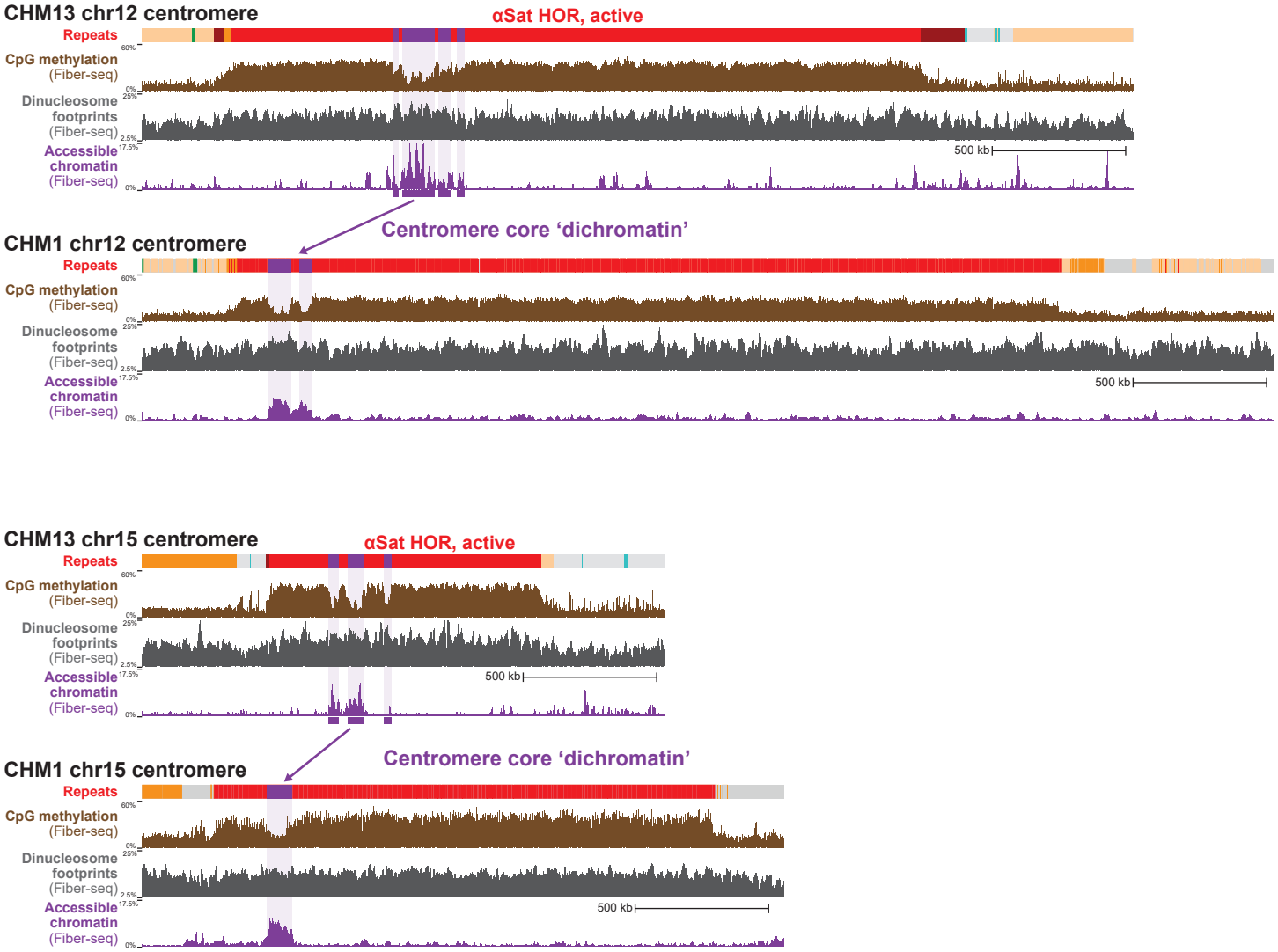
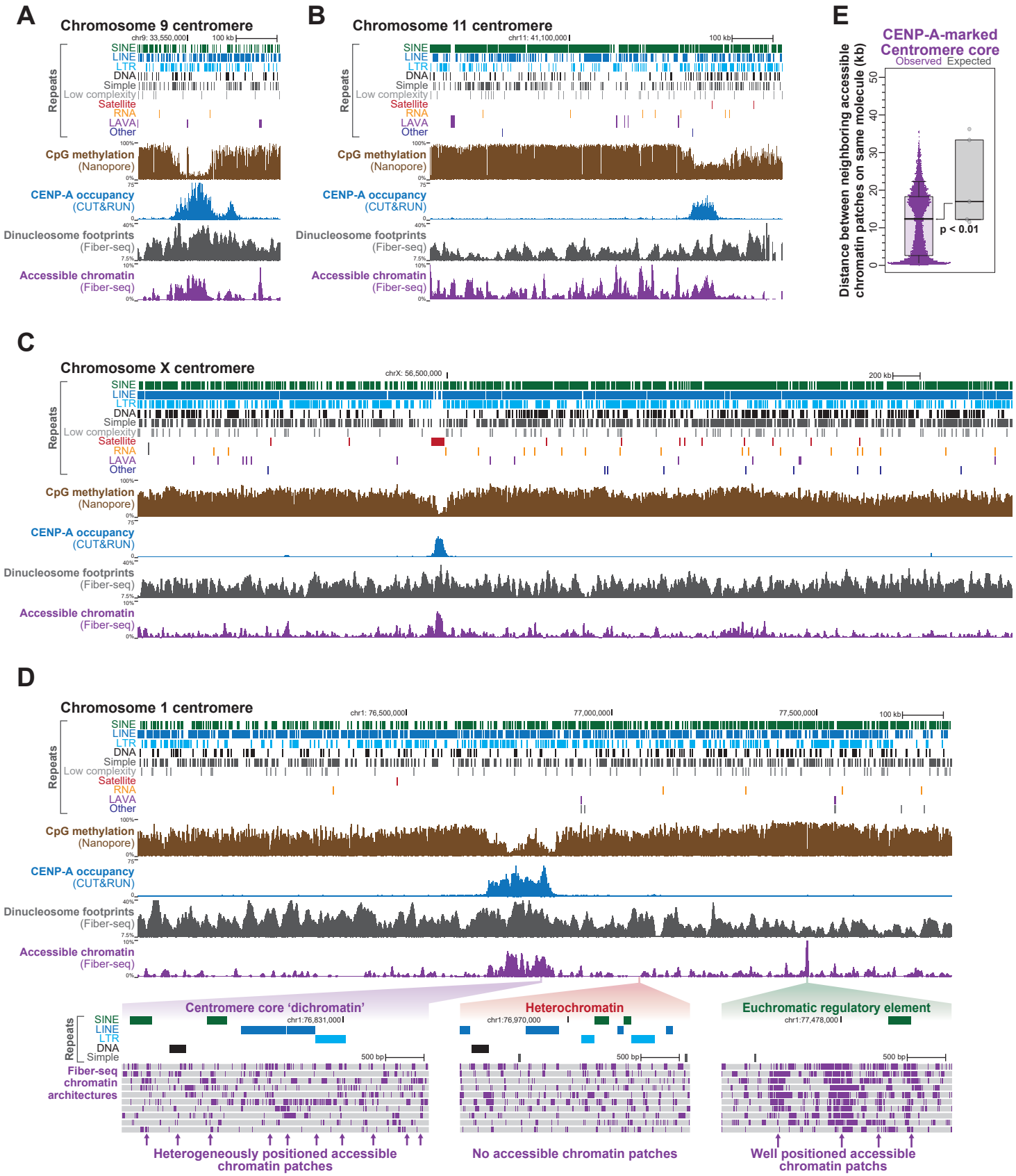

CHM13 chr12 centromere
αSat HOR, active
Repeats
CpG methylation (Fiber-seq)
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
500 kb
Centromere core 'dichromatin'
CHM1 chr12 centromere
Repeats
CpG methylation (Fiber-seq)
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
500 kb
CHM13 chr15 centromere
αSat HOR, active
Repeats
CpG methylation (Fiber-seq)
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
500 kb
Centromere core 'dichromatin'
CHM1 chr15 centromere
Repeats
CpG methylation (Fiber-seq)
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
500 kb

Figure S6


A
Chromosome 9 centromere
chr9: 33,550,000
100 kb
Repeats
SINE
LINE
LTR
DNA
Simple
Low complexity
Satellite
RNA
LAVA
Other
CpG methylation (Nanopore)
CENP-A occupancy (CUT&RUN)
Dinucleosome footprints (Fiber-seq)
Accessible chromatin (Fiber-seq)
B
Chromosome 11 centromere
chr11: 41,100,000
100 kb
E
CENP-A-marked Centromere core
Observed Expected
Distance between neighboring accessible chromatin patches on same molecule (kb)
p < 0.01
C
Chromosome X centromere
chrX: 56,500,000
200 kb
D
Chromosome 1 centromere
chr1: 76,500,000
77,000,000
77,500,000
100 kb
Centromere core 'dichromatin'
chr1:76,831,000
Heterochromatin
chr1:76,970,000
Euchromatic regulatory element
chr1:77,478,000
500 bp
Fiber-seq chromatin architectures
Heterogeneously positioned accessible chromatin patches
No accessible chromatin patches
Well positioned accessible chromatin patchs