Non-B-form DNA structures mark centromeres

Sivakanthan Kasinathan\textsuperscript{1,2} & Steven Henikoff\textsuperscript{2,3*}

1. Medical Scientist Training Program, University of Washington School of Medicine, Seattle, WA 98195
2. Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109
3. Howard Hughes Medical Institute, Seattle, WA 98109

*Correspondence: steveh@fhih.org

Keywords
Centromere, neocentromere, CENP-B, CENP-A, HJURP/Scm3, cruciform
Abstract

Animal and plant centromeres are embedded in repetitive “satellite” DNA, but are thought to be epigenetically specified. To define genetic characteristics of centromeres, we surveyed satellite DNA from diverse eukaryotes and identified variation in <10-bp dyad symmetries predicted to adopt non-B-form conformations. Organisms lacking centromeric dyad symmetries had binding sites for sequence-specific DNA binding proteins with DNA bending activity. For example, human and mouse centromeres are depleted for dyad symmetries, but are enriched for non-B DNA and are associated with binding sites for the conserved DNA-binding protein CENP-B, which is required for artificial centromere function but is paradoxically non-essential. We also detected dyad symmetries and predicted non-B-form DNA structures at neocentromeres, which form at ectopic loci. We propose that centromeres form at non-B-form DNA because of dyad symmetries or are strengthened by sequence-specific DNA binding proteins. Our findings resolve the CENP-B paradox and provide a general basis for centromere specification.

Introduction

Centromeres are nearly universally marked by the presence of nucleosomes containing the histone H3 variant CenH3 (CENP-A), which is sufficient to direct formation of the proteinaceous kinetochore that links chromosomes to the mitotic spindle (Mendiburo, et al. 2011). CENP-A is partitioned equally between sister chromatids during cell division and is deposited coincident with mitotic exit (Jansen, et al. 2007) by the conserved chaperone Holliday junction binding protein (HJURP/Scm3) (Kato, et al. 2007; Foltz, et al. 2009; Sanchez-Pulido, et al. 2009). Mechanisms dictating selection of particular loci as centromeres in complex genomes have remained elusive. One possibility is that centromeric DNA directs CenH3 deposition (De Rop, et al. 2012). Indeed, budding yeast ‘point’ centromeres are completely determined by ~120 bp sequences (Meraldi, et al. 2006; Gordon, et al. 2011). However, the centromeres of most other eukaryotes are made up of large tandem arrays of repetitive genetic elements whose role in centromere formation is unclear.

Centromeric repeats act as selfish genetic elements by driving non-Mendelian chromosome transmission during meiosis (Henikoff, et al. 2001) in turn spurring the rapid evolution of centromeric proteins to restore meiotic parity (Malik and Henikoff 2009). Incompatibilities between centromeric proteins and selfish DNA repeats may therefore serve as a molecular basis for speciation (Henikoff, et al. 2001). This centromere drive model is supported by observations confirming that centromeric alterations distort chromosome transmission in humans (Daniel 2002), mice (Chmatal, et al. 2014; Iwata-Otsubo, et al. 2017) and plants (Fishman and Saunders 2008). Critically, centromere drive supposes that centromeres are genetically defined at least through a transient stage of their evolution (Dawe and Henikoff 2006). However, the identification of neocentromeres that form at ectopic loci devoid of alphoid DNA (du Sart,
et al. 1997) has led to the widely held view that centromere identity is specified independently of DNA sequence through “epigenetic” mechanisms involving CENP-A (Karpen and Allshire 1997; Ekwall 2007; Allshire and Karpen 2008).

Consistent with a role for genetic factors in centromere specification, primate centromeres have remained embedded in megabases of ~170-bp α-satellite tandem repeats over nearly 60 million years of evolution (Alexandrov, et al. 2001). The identification of CENP-B, a sequence-specific DNA-binding protein at centromeres that is highly conserved in mammals (Sullivan and Glass 1991), suggested a possible mechanism for DNA-encoded centromere specification (Masumoto, et al. 1989). However, although CENP-B is present in all primate genomes (Schueler, et al. 2010), the 17-nt CENP-B box sequence bound by the protein is not present on all centromeres within a species and is not found in all primates (Masumoto, et al. 1989; Haaf, et al. 1995). Further complicating this “CENP-B paradox” (Goldberg, et al. 1996; Kipling and Warburton 1997), CENP-B binding is required for de novo centromere formation on artificial chromosomes (Ohzeki, et al. 2002) and has been shown to enhance the fidelity of chromosome segregation (Fachinetti, et al. 2015). Reconciling these observations concerning variation in centromeric DNA with evidence that centromeres are epigenetically defined remains an outstanding challenge.

Here, we reconsider the role of DNA sequence in specification of centromere identity. We mined publicly available whole-genome sequencing datasets for centromeric DNAs from great apes, Old World Monkeys (OWMs), mouse, chicken, plants, and yeasts and characterized clade-specific variation in abundance of dyad symmetries predicted to adopt non-B-form DNA conformations. We found a highly restricted distribution of CENP-B boxes limited to great apes and mouse. Indeed, we could show that absence of CENP-B boxes in OWM α-satellite corresponds to lack of CENP-B binding. We discovered that this loss of CENP-B binding was correlated with an increased tendency of centromeric satellites to form predicted non-B-form DNA structures such as cruciforms. Using experimental datasets, we were able to detect these non-B-form DNA structures at functional human and mouse centromeres. In budding yeasts, we identified a similar association between centromeric dyad symmetry and exaptation of a CENP-B –like DNA-binding protein. We also predicted non-B-form DNA formation at the human Y chromosome centromere and human and chicken neocentromeres, which are not associated with CENP-B binding. Based on these data, we advance a unifying model for centromere specification based on recognition of non-B-form DNA structures, either aided or unaided by sequence-specific DNA binding proteins.

**Results**

Dyad symmetries are common features of centromeres
Sequence analysis of centromeres has been challenging because of the intractability of repeats; however, studies in the fission yeast *Schizosaccharomyces pombe*, which has regional centromeres similar to vertebrates, and analyses of primate α-satellite consensus sequences and budding yeast centromeres have suggested the existence of sequence-encoded features that are important for centromere function (Koch 2000; Catania, et al. 2015). To identify potential DNA sequence determinants that might have been overlooked in recent studies of centromere specification, we first catalogued <10-nt dyad symmetries. Using published consensus sequences or libraries of species-specific satellite consensus sequences derived from *de novo* tandem repeat detection (Figure S1A), we scanned deep (>10X coverage) publicly available paired-end, whole genome sequencing datasets from a sampling of vertebrates, fission yeast, and plants to identify centromeric sequences, which are typically poorly represented in genome assemblies. In chicken, which has both repetitive and unique centromeres (Shang, et al. 2010), we restricted our analyses to well-characterized unique sequences. Visual examination of centromeric sequences identified dyad symmetries with species-specific variation (Figure 1A). For species with high-quality genome assemblies, we compared dyad symmetries in centromeric regions and dinucleotide composition-matched background genomic regions without known centromeres and confirmed the pattern of species-specific enrichment of dyad symmetries (Figure 1B). To determine whether dyad symmetries may have been selected for during centromere evolution, we compared centromeric sequences from each species to random permutations of the same sequences to account for nucleotide composition. Enumeration of dyad symmetries over varying palindrome lengths revealed enrichment of >3-bp dyad symmetries in OWMs, horse, chicken, stickleback, fission yeast, and plants, but not in great apes or mouse (Figure 1C, S2A). Based on these analyses, we conclude that dyad symmetries are a unique feature of many eukaryotic centromeres and may have been selected for during centromere evolution.

Predicted non-B-form DNA structures at dyad-enriched centromeres

Given the possibility that regions of dyad symmetry may adopt non-B DNA conformations (Pearson, et al. 1996) and reports of stem-loop, hairpin, and triplex nucleic acid at human centromeres (Ohno, et al. 2002; Jonstrup, et al. 2008; Garavis, et al. 2015; Aze, et al. 2016), we sought to characterize the theoretical secondary structure formation potential of centromeric DNAs. We used a computational method that models stress-induced structural transitions (SIST) in DNA (Zhabinskaya, et al. 2015) to determine whether variation in dyad symmetry corresponds to differing predispositions for adopting non-B conformations. Comparing SIST DNA melting and cruciform extrusion scores for centromeric sequences and dinucleotide composition-matched non-centromeric background genomic intervals for select species, we found that dyad-enriched centromeres were associated with significantly higher levels of predicted non-B-form DNA (Figure 2A). We then used an independent approach to predict folding
free energies for centromeric DNAs and found that species predicted by SIST to adopt non-B-form DNA tended to form more stable secondary structures (Figure 2B). For example, consistent with the SIST predictions, the distributions of predicted free energies for dyad-enriched α-satellite from OWMs were substantially left-shifted compared to dyad-depleted great ape α-satellite, suggesting that OWM centromeres may adopt more stable secondary structures than great ape centromeres (two-sample Kolmogorov-Smirnov p-value << 1e-5; Figures 2B,C, S2B). Similar trends were observed in other species (Figure 2C). From these analyses, we conclude that centromeres enriched for dyad symmetries may adopt stable non-B-form DNA structures such as cruciforms.

CENP-B-associated enrichment of non-B-form DNA at dyad-depleted human and mouse centromeres

We next sought to verify computational structure predictions using publicly available data from genome-wide mapping of non-B-form DNA using potassium permanganate treatment (permanganate-seq) in mouse and human cells (Kouzine, et al. 2013; Kouzine, et al. 2017). In mouse, minor satellite (MiSat) sequences constitute the functional centromere (Joseph, et al. 1989), while adjacent major satellite (MaSat) domains are heterochromatic (Horz and Altenburger 1981). We confirmed the relative abundances of these satellites (Figure 3A) and, in agreement a previous report (Guenatri, et al. 2004), we detected CENP-B boxes in MiSat arrays while MaSat sequences contained very few CENP-B boxes (Figure 3A). Importantly, unlike centromeric MiSat sequences, heterochromatic MaSat arrays were predicted to more favorably adopt non-B-form structures (Figure 3B). MaSat sequences were enriched for permanganate-seq reads concordant with structure predictions (Figure 3C). Surprisingly, MiSat sequences were also highly enriched for permanganate-seq signal (Figure 3C). Permanganate-seq signal increased in activated B-cells, which are undergoing cell division, relative to quiescent resting B-cells (Figure 3C). To determine whether non-B-form DNA was associated with functional MiSat sequences occupied by CENP-A, we analyzed chromatin immunoprecipitation and sequencing (ChIP-seq) data (Iwata-Otsubo, et al. 2017). CENP-A occupancy was positively correlated with permanganate-seq signal (Spearman’s ρ = 0.45; p << 1e-10), with MiSat sequences associated with low-scoring CENP-B sites having the least CENP-A and permanganate-seq alignments (Figure 3D,E).

Similarly, contrary to the predicted poor tendency for human alphoid DNA to adopt non-B-form DNA structures (Figure 2C), more than half the human permanganate-seq data aligned to α-satellite, representing a ~10-fold enrichment for alphoid DNA (Figure 3F). Regions enriched for non-B-form DNA appeared to colocalize with CENP-A, particularly at functional α-satellite dimers with CENP-B boxes that we previously identified by ChIP-seq (Henikoff, et al. 2015), but were depleted at presumably non-functional α-satellite arrays that are not at centromeres (Slee, et al. 2012) and not associated with CENP-B boxes (Figure 3G). Indeed, CENP-A occupancy and permanganate-seq signal (Spearman’s ρ =

5
0.35; \( p << 1 \times 10 \), with alphoid arrays associated with low-scoring CENP-B sites having the least CENP-A and permanganate-seq signal (Figure 3H). We conclude that non-B-form DNA is characteristic of functional, CENP-B-associated human and mouse centromeric satellites and may form in a cell cycle-dependent manner.

Functional Old World Monkey centromeres enriched for non-B-DNA are not bound by CENP-B

Given the inverse relationship between CENP-B binding and non-B DNA, we speculated that CENP-B boxes might be specific to great-ape centromeres. To test this hypothesis, we searched for matches to the multiple alignment-based consensus CENP-B box sequence (TTCGNNNNANCGGG) required to support CENP-B DNA binding (Iwahara, et al. 1998) in randomly sampled whole-genome sequencing reads. Although OWMs tended to have more \( \alpha \)-satellite (Figure 4A), CENP-B boxes were highly enriched in great apes whereas OWM genomes contained negligible matches to the minimal CENP-B box (Figure 4B), consistent with a previous report (Goldberg, et al. 1996). We further verified great ape-specific enrichment of CENP-B boxes using the SELEX-defined CENP-B motif (Jolma, et al. 2013) (Figure S1B). To more finely characterize functional centromeric sequences in divergent primates, we mapped genomic binding of CENP-A and CENP-B in K562 (human) and Cos-7 (African green monkey) cell lines using CUT&RUN, which profiles DNA fragments released by antibody-targeted nuclease cleavage \textit{in situ} (Skene and Henikoff 2017b). We detected substantial enrichment for alphoid sequence in CENP-A CUT&RUN in both human and African green monkey (Figure 4B). Consistent with a paucity of CENP-B boxes in OWMs, CENP-B protein binding within alphoid sequence was observed in human but not African green monkey (Figure 4C). CENP-B boxes were highly enriched in CENP-A CUT&RUN reads from K562 cells, but depleted in CENP-A associated sequences from Cos-7 cells (Figure 4D). Taken together, these analyses suggest that CENP-B protein and binding sites are specific to dyad-depleted great ape centromeres.

CENP-B-depleted human Y centromere and vertebrate neocentromeres are enriched for non-B-form DNA

The observations that the human Y chromosome centromere and neocentromeres are devoid of CENP-B boxes (Haaf, et al. 1995; Saffery, et al. 2000) and that the CENP-B protein is non-essential (Kapoor, et al. 1998) contradict the requirement for CENP-B in \textit{de novo} centromere assembly (Ohzeki, et al. 2002). To reexamine this “CENP-B paradox” (Goldberg, et al. 1996; Kipling and Warburton 1997) in light of DNA-encoded structural features at centromeres, we asked whether DYZ3 alphoid repeats from the human Y chromosome and vertebrate neocentromere sequences are associated with non-B DNA. First, we compared predicted folding free energies of fragments derived from DYZ3 and non-DYZ3 alphoid DNA and found that DYZ3 sequences are associated with thermodynamically favorable non-B structures.
(Figure 5A,B), suggesting that the CENP-B binding may not be required for de novo centromerization. Next, we analyzed CENP-A ChIP-seq data from human cell lines bearing neocentromeres (Hasson, et al. 2013) and detected enrichment for dyad symmetries predicted to form stable secondary structures in CENP-A-associated neocentromere domains (Figures 5C). To determine whether cruciform structures, a specific class of non-B-form DNA, are associated with neocentromeres, we used data from genome-wide analysis of palindrome formation (GAP-seq) based on DNA renaturation and S1-nuclease treatment (Yang, et al. 2014), and found regions that may form cruciforms in vivo at neocentromeres (Figures 5C). Neocentromeres were also markedly enriched for dyad symmetries relative to base composition-matched randomly selected genomic regions and native centromeric sequences (Figure 5C). We also analyzed data from a chicken neocentromere cell line to determine whether a similar trend is generally observed in vertebrates. Like other mammalian centromeres, chicken neocentromeres were also enriched for short dyad sequences and predicted to undergo strand separation and cruciform transitions (Figure 5D). Taken together, these analyses suggest that native centromeres depleted for CENP-B boxes such as the human Y chromosome centromere and neocentromeres are enriched for non-B-form DNA structures.

Exaptation of a CENP-B-like protein at dyad-depleted budding yeast centromeres

In contrast to most eukaryotes, the saccharomycetes have among the simplest known centromeres, which are fully determined by a ~120 bp sequence composed of three centromere determining elements (CDEs I-III) (Clarke and Carbon 1985). To gain insight into whether dyad symmetries are a conserved feature of centromeres in diverse eukaryotes, we analyzed sequences of well-characterized budding yeast centromeres. We determined the extent of CENP-A binding at annotated centromeres using published datasets (Henikoff, et al. 2014; Thakur, et al. 2015) and quantified the extent of DNA melting and cruciform transition predicted by SIST (Zhabinskaya, et al. 2015). We first analyzed the average ChIP-seq signal for the CENP-A homologue Cse4 (Henikoff, et al. 2014), dyad symmetry, and SIST DNA melting and cruciform extrusion scores from Saccharomyces cerevisiae. Similar to what we observed in vertebrates, we found higher levels of predicted non-B DNA at S. cerevisiae centromeres that were enriched for dyad symmetries compared to composition-matched non-centromeric genomic regions (Figures 6A and S6B). We detected a similar pattern in enrichment for dyad symmetries and non-B-form DNA at the centromeres of other sensu strictu yeasts; however, despite similar sequence composition to sensu strictu saccharomycetes the sensu lato species S. castellii and S. darienensis had comparatively less dyad symmetry and lower SIST DNA melting and cruciform extrusion scores (Figures 6B,C and S3C,D). Recently, S. castellii and S. darienensis were shown to have divergent point centromeres (Kobayashi, et al. 2015), with a substantially different CDEI region devoid of a binding site for the basic helix-loop-helix transcription factor Cbf1, which is found at CDEI sequences of sensu strictu centromeres.
We found that the consensus site at CDEI of dyad-depleted sensu lato yeasts is strongly predicted ($p < 1e^{-5}$) to bind the DNA-bending general regulatory factor Reb1 (Figures 6B, C and S3E) based on searching a database of 203 yeast transcription factors (MacIsaac, et al. 2006) and comparison to a consensus motif from high-resolution mapping (Kasinathan, et al. 2014). These analyses suggest that non-B-form DNA at centromeres may represent an ancient mechanism for centromere specification in eukaryotes and that DNA-binding proteins such as CENP-B and Reb1 may serve an important role in shaping the evolution of centromeric DNA.

**Discussion**

We used comparative analysis of whole genome sequencing and functional genomic datasets to define evolutionary transitions in centromeres (Figure 7A). We found that short dyad symmetries that are predicted to adopt non-B-form structures are characteristic of OWM, chicken, and fission yeast centromeres, while great ape and mouse centromeres were comparatively depleted for dyad symmetries and not predicted to form non-B DNA. Surprisingly, both human and mouse centromeres were found to be associated with non-B DNA *in vivo*, with greater enrichment for non-B-form DNA at CENP-A-occupied satellite sequences associated with CENP-B boxes. Importantly, we did not detect CENP-B boxes at the OWM centromeres, which have a-satellite repeats predicted to adopt stable non-B-form structures. We also profiled CENP-A and CENP-B binding in human and African green monkey cell lines and demonstrated directly that functional OWM centromeres are not bound by CENP-B. Further, we found that the human Y chromosome centromere, which notably does not bind CENP-B, is predicted to form more thermodynamically favorable non-B-form DNA structures than other human centromeres. These observations resolve conflicting reports about the presence of CENP-B boxes in OWMs (Goldberg, et al. 1996; Yoda, et al. 1996).

Consistent with the presence of non-B-form DNA structures at functional centromeres, single-stranded DNA, hairpins, triplexes, and i-motifs have been observed in a-satellite *in vitro* and/or *in vivo* (Ohno, et al. 2002; Jonstrup, et al. 2008; Garavis, et al. 2015; Aze, et al. 2016). Taken together with our genomic analyses, these lines of evidence suggest testable models for the specification of centromere identity (Figure 7B, C). One possibility is that non-B-form DNA directly specifies centromere identity (Figure 7B). Because the CENP-A chaperone HJURP was named based on its *in vitro* Holliday junction-binding activity (Kato, et al. 2007), it is tempting to speculate that the four-way junction DNA structures recognized by HJURP are short cruciforms, which may form spontaneously or inducibly. Organisms such as OWMs may have satellites capable of adopting energetically favorable conformations recognized by HJURP and its ortholog Scm3, which is the CenH3 chaperone in both budding and fission yeast. In contrast, other species such as great apes may require the binding of a sequence-specific DNA-binding
protein such as CENP-B to promote formation of non-B DNA structures. It is intriguing that both mammalian CENP-B and yeast Reb1 bend DNA \(\sim 60^\circ\), raising the possibility that formation of non-B-form DNA in activated mouse and human B cells is directly or indirectly mediated by CENP-B-mediated DNA bending.

Alternatively, centromere specification may occur through a transcription-based mechanism that produces RNAs with secondary structure (Figure 7C). Transcription may occur readily at centromeres that adopt non-B-form structures, which include melted DNA, while sequences that are comparatively resistant to adopting non-B conformations may require the action of an architectural DNA-binding protein such as CENP-B or Reb1. Transcription in the case of dyad-enriched satellites may also occur through recognition by the rDNA transcription factor UBF, which has cruciform-binding activity (Copenhaver, et al. 1994), and RNA polymerase (Pol) I, whereas transcription at sequences that do not favorably adopt non-B structures could occur via RNA Pol II. In support of this model, Pol II transcription has been shown to occur at human centromeres (Quenet and Dalal 2014; McNulty and Sullivan 2017) and to be functionally important in budding yeast centromeres (Ohkuni and Kitagawa 2011). Pol I has similarly been suggested to be involved in human centromere function (Wong, et al. 2007).

These non-mutually exclusive mechanisms provide parsimonious explanations for a number of puzzling phenomena and are compatible with proposed functions for CENP-B in enhancing chromosome segregation fidelity (Fachinetti, et al. 2015). They resolve the CENP-B paradox by suggesting a context-specific role for helix-deforming DNA-binding proteins in facilitating DNA secondary structure formation and/or transcription. In addition to recruiting HJURP to centromeres, non-B-form DNA and/or active transcription may suppress the unscheduled incorporation of canonical H3-containing nucleosomes (Nickol and Martin 1983) and explain the enrichment of DNA breaks and some damage repair proteins at centromeres (Guerrero, et al. 2010; Crosetto, et al. 2013) (Lu, et al. 2015). Consistent with the expansion of centromeres by HJURP tethering (Perpelescu, et al. 2015) and the high frequency of neocentromere formation in chicken (Shang, et al. 2013), we observed that the chicken genome is predicted to form non-B-form DNA structures more favorably than mammalian genomes.

This view also unifies the genetic and epigenetic conceptions of the centromere by explaining why CENP-B is necessary for \textit{de novo} centromerization of artificial chromosomes (Ohzeki, et al. 2002), but not required to maintain native centromeres, which could be propagated by the presence of CENP-A (Fachinetti, et al. 2013) or Mis18 (Nardi, et al. 2016). Similarly, neocentromeres may be seeded in loci that adopt non-B conformations due to sequence and/or chromatin features. Centromere specification by recognition of nucleic acid structures would permit conservation of the general architecture of the centromere and kinetochore (Drinnenberg, et al. 2016) while providing a large sequence space that can be sampled during rapid evolution of DNA, suggesting a basis for driving genetic conflict at centromeres.
Therefore, genetically encoded structures may represent a common mechanism for specification of eukaryotic centromere identity.

**Author contributions**
S.K. performed the analyses, S.H. performed the experiments, and S.K. and S.H. wrote the manuscript.

**Acknowledgements**
We gratefully acknowledge P. Talbert, S. Ramachandran, J. Thakur, K Ahmad, and D. Melters for insightful discussions and suggestions and J. Henikoff for assistance with data analysis. We thank S. Biggins and H. Malik for comments on the manuscript. This work was funded by support from the Micki & Robert Flowers ARCS Endowment from the Seattle Chapter of the ARCS Foundation (S.K.) and the Howard Hughes Medical Institute (S.H.).

**Methods**

**Datasets.** NCBI Sequence Read Archive accession numbers and references for publicly available datasets from a variety of species used in this study are included in Table S1. Note that members of the *Chlorocebus* and *Macaca* genera included in this study may represent subspecies rather than bona fide species (Yan, et al. 2011; Warren, et al. 2015). Illumina whole-genome sequencing (WGS) data selected were paired-end ~100x100-bp datasets to facilitate analysis of repeat variation.

**Pre-processing of Illumina data.** Raw paired-end Illumina reads were subjected to adapter trimming and quality filtering using BBduk ([http://jgi.doe.gov/data-and-tools/bbtools/](http://jgi.doe.gov/data-and-tools/bbtools/)) with the following parameters:

```
ftm=5 qtrim=rl trimq=10 maq=15 minlen=85 ref=adapters.fa
```

The FASTA file ‘adapters.fa’ is part of the BBduk package and contains sequences of the Illumina TruSeq adapters. All subsequent analyses were performed on trimmed and filtered Illumina reads.

**Alignment of sequencing data.** Bowtie2 (v2.2.5) was used to perform alignments to published reference genomes or custom references as indicated. The following alignment parameters were used for paired-end reads:

```
--end-to-end --very-sensitive --no-unal --no-mixed --no-discardant --overlap --dovetail -I 10 -X 700
```

Alignment parameters used for single-end reads were:
Reference genomes for short read alignment. The following reference assemblies available from the UCSC Genome Browser were used: hg38 (human), mm10 (mouse), galGal5 (*Gallus gallus*), and sacCer2 (*S. cerevisiae*). For human and mouse, masked versions of the hg38 and mm10 assemblies were created using the hard-masked sequences available from the UCSC Genome Browser. For African green monkey, the RefSeq *Chlorocebus sabaeus* assembly (accession GCF_000409795.2) was hard-masked using RepeatMasker annotations available from RefSeq. The *S. pombe* assembly (ASM294v2) was downloaded from PomBase (McDowall, et al. 2015); the *S. mikatae* (IFO 1815^5^) and *S. kudriavzevii* (IFO 1815^7^) ultra-scaffolds were previously published (Scannell, et al. 2011) and are available online ([http://sss.genetics.wisc.edu/cgi-bin/s3.cgi](http://sss.genetics.wisc.edu/cgi-bin/s3.cgi)). The *S. castellii* assembly (NRRL-Y12630) was downloaded from the Saccharomyces Genome Database (Cherry, et al. 2012) and the *S. dairenensis* genome is available from the NCBI Assembly database (accession no. GCF_000227115.2). In all cases, Bowtie2 indexes were built using default parameters.

De novo definition of centromeric satellite units. Sanger reads, contigs from whole-genome assembly, and contigs from local assembly of Illumina reads were used to define centromeric satellites. Tandem Repeats Finder v5.02 (TRF) (Benson 1999) was used to identify all tandemly repeated sequences. Sequences corresponding to peaks in the resulting repeat length histograms that were not other abundant repeats (Alu, etc.) were classified as putative centromeric satellites. TRF was run with the following parameters:

2 7 7 80 10 50 1000 -h -ngs

Sequences from TRF peaks that passed a DUST complexity filter (implemented in PRINSEQ, [http://prinseq.sourceforge.net](http://prinseq.sourceforge.net); parameters: -lc_method dust -lc_threshold 7) were retained for subsequent analysis. In order to define unique monomers without shifting sequences to occupy similar registers, we took all tandem repeats corresponding to the major peak and subjected them to local alignment-based clustering using CD-HIT-EST (Li and Godzik 2006) with the following parameters: -c 0.8 -bak 1 -M 0 -d 0 -n 4 -G 0 -A 43

For each species, CD-HIT-EST-reported consensus sequences for clusters containing at least 1% of the input sequences were used to construct a BLAST database, which was then used to scan the Sanger reads and contigs and define new monomer locations. BLASTN searching was performed with the following options: -task blastn -num_alignments 1

Identification of satellite monomer fragments in Illumina datasets. Species-specific repeat databases produced as described above were used to identify fragments of monomers in paired-end Illumina
sequencing datasets using BLASTN with the following options: -task blastn -num_alignments 1 -outfmt "6 qseqid qstart qend sseqid evalue sstrand pident length qlen"
The high depth of genome coverage in the selected datasets necessitated randomly sampling up to $10^6$ reads for each species.

Principal components analysis (PCA) and clustering of Illumina reads. We analyzed intra- and inter-specific variation in centromeric satellite sequences using dimensionality reduction. At least 10,000 satellite-containing reads were sampled from each species and oriented relative to the species-specific consensus sequences based on the BLASTN-reported alignment strand (see above). For each read, a $k$-mer frequency vector was created by sliding a $k$-length window in 1-nt steps along the length of the read and enumerating the fractional representation of each observed $k$-mer. For each species, $k$-mers appearing in more than 10 reads and in less than 50% of reads were retained for subsequent analysis. PCA was then performed on the set of frequency vectors for each species (Figure S1C,D). To account for the possibility of clusters with non-convex topology, 2000 reads used for PCA analysis were subjected to spectral clustering with number of clusters varying from 1 to 19. The silhouette coefficient (Rousseeuw 1987) was computed for each clustering. The clustering that produced the maximum silhouette coefficient was selected as the optimal clustering for each species and $k$-mer length (Figure S1C,D).

CENP-B box abundance in primates and mouse. To estimate the abundance of CENP-B boxes, occurrences of matches to the multiple alignment consensus CENP-B box sequence (TTCGNNNNNNANNCGGG), which is required for DNA-binding, were counted in Illumina WGS data. We further used the SELEX-defined CENP-B motif (Jolma, et al. 2013)(JASPAR accession MA0637.1) to search for matches using FIMO (Grant, et al. 2011) with default parameters.

CUT&RUN profiling of CENP-A and CENP-B in human and African green monkey cell lines. CUT&RUN was performed as described (Skene and Henikoff 2017b, a). Briefly, K562 or Cos-7 cells were gently washed twice in room temperature Wash buffer [20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM spermidine and a Roche complete EDTA-free tablet (Sigma-Aldrich) per 50 ml], with 3 min centrifugation at 600 xg, mixed with activated Concanavalin A-coated magnetic beads (Bangs Laboratories) and rotated 5-10 min. Beads were captured by placing on a magnet stand, decanted and resuspended in Antibody buffer [2 mM EDTA and antibody at 1:100 in Dig-wash (Wash buffer supplemented with 0.05% digitonin (Calbiochem)]. Antibodies used were CENP-A (Abcam ab13939, mouse monoclonal), CENP-B (Abcam ab25734), Histone H3K27me3 (Cell Signaling Technologies 9733), and IgG (either Antibodies Online ABIN102961 guinea pig anti-rabbit, or GeneTex GTX105137,
rabbit anti-human mitochondrial RNA polymerase). After overnight binding at 4 °C with rotation, beads were captured and washed once or twice in Dig-wash. For the CENP-A samples, beads were resuspended in secondary antibody (Abcam Ab46540, Rabbit anti-mouse) in Dig-wash and incubated 1 hr at 4 °C and washed once in Dig-wash. Beads were resuspended in Protein A-MNase (Batch #5 360 µg/ml) 1:500 in Dig-wash, incubated 1 hr at 4 °C, washed twice in Dig-wash and resuspended in 100 µL Dig-wash. Tubes were placed 0 °C, mixed with 2 µL 100 mM CaCl₂, incubated at 0 °C for 30 min, and reactions were stopped by addition of 100 µL 340 mM NaCl, 20 mM EDTA pH8, 4 mM EGTA, 0.05% digitonin, 50 µg/ml glycoprotein and 200 µg mono-nucleosomal S. cerevisiae (spike-in) DNA. Samples were incubated 10 min at 37 °C and centrifuged 5 min 16,000 xg at 4 °C and the supernatant was treated with 25 µg/ml RNase A (Thermo) 10 min at 37 °C. After phenol-chloroform-isoamyl alcohol and chloroform extraction, DNA was precipitated by addition of 2.5 volumes 100% ethanol, chilled on ice, centrifuged 10 min at 4 °C at 16,000 xg and the pellets were rinsed in 100% ethanol and air-dried. Pellets were resuspended in 1 mM Tris pH8 0.1 mM EDTA and used for standard Illumina library preparation. Paired-end PE25x25 sequencing was performed by the Fred Hutch Shared Genomics Resource. Data have been deposited in GEO (GSEXXXXX).

Analysis of CUT&RUN data. CENP-A and CENP-B CUT&RUN data were aligned to human alphoid BAC reference sequences described previously (Henikoff, et al. 2015) or to an α- satellite-containing Chlorocebus aethiops BAC sequence (GenBank accession AC239401.3). Read length histograms were generated by counting the Bowtie2r-reported aligned paired-end fragment lengths and autocorrelation of the resulting read length distributions was performed using Numpy. H3K27me3 CUT&RUN data were aligned to repeat-masked versions of the hg38 and Chlorocebus sabaeus assemblies as described above.

Detection of perfect dyad symmetries in Illumina reads. A suffix array-based approach was used to detect dyad symmetries. Each sequence of interest was concatenated with its reverse complement and the suffix and longest common prefix (LCP) arrays were constructed. Traversal of the suffix array, which indexes the suffixes of the sequence and its reverse complement in lexicographic order, provides an efficient way to identify repeated patterns. Specifically, in the case of a pattern with dyad symmetry, the sequence on the forward and reverse strands should occupy adjacent positions in the suffix array. Dyad symmetries of various lengths (3-7 bp) separated by “spacer” sequences (0-20 bp) were discovered using this approach.

Detection of perfect and imperfect dyad symmetries in Illumina reads and genomic regions. We used EMBOSS palindrome (Rice, et al. 2000) to detect dyad symmetries with mismatches in the palindromic region with the following parameters (varying the number of mismatches):
For each position in a sequence of interest, we defined dyad density as the sum of the lengths of the palindromic regions that contain that position. For a sequence, the length-normalized dyad density was defined as the sum of the per-position values divided by the sequence length.

**DNA secondary structure prediction.** RNAfold from the ViennaRNA package (Lorenz, et al. 2011) was used to predict folding free energies. RNAfold was used with the following parameters for DNA secondary structure prediction:

```plaintext
-noGU -noconv -noPS -paramFile=dna_mathews2004.par -p --g
```

Predictions were performed on random samples of $10^4$ BLAST-defined monomers (Sanger data) or $10^4$ reads containing BLAST matches (Illumina data).

**Prediction of DNA melting and cruciform transitions.** Strand separation and cruciform extrusion propensities were predicted using SIST (Zhabinskaya, et al. 2015) with default parameters. For sequences greater than 10 kb in length (the maximum permissible length compatible with SIST), we slid a 5 kb window in 2.5 kb steps to generate short sequences that were analyzed using SIST. These SIST predictions were then reassembled for the full sequence by conservatively taking the maximum at each base (for bases spanned by multiple windows). For a given sequence, melt and cruciform scores were computed by summing the estimated transition probabilities for each position in the sequence and dividing by the length of the sequence.

**Selection of control genomic regions.** To account for sequence composition of centromere and neocentromere sequences, we selected random genomic regions without known centromere activity and with similar dinucleotide composition. Dinucleotide frequencies for a query sequence of interest were calculated and the Spearman rank correlation ($\rho$) was used to identify non-overlapping windows of the same length as the query in a genome with a similar dinucleotide frequency pattern. We considered two regions to be sufficiently similar if $\rho \geq 0.9$ and excluded regions overlapping annotated centromeric sequences. Up to 1,000 random sites defined using this procedure were used for comparisons.

**Statistical analyses.** The two-sample Kolmogorov-Smirnov test was used to compare distributions of values of interest (e.g., dyad density, SIST scores).

**Analysis of CENP-A ChIP-seq data.** Our published CENP-A ChIP-seq 100x100 bp Illumina data were subjected to the same adapter trimming and quality filtering steps described above prior to merging pairs
using SeqPrep as described previously (Henikoff, et al. 2015). Merged pairs were aligned to alphoid reference sequences using Bowtie2 using the single-end mapping parameters described above. Data from ChIP-seq of *S. cerevisiae* Cse4 (Henikoff, et al. 2014), *S. pombe* Cnp1 (Thakur, et al. 2015), and CENP-A in human neocentromeres cell lines (Hasson, et al. 2013) were aligned using the paired-end mapping parameters described above.

**Analysis of DNase-seq data.** Adapter trimmed and quality filtered human and mouse paired-end DNase-seq (DNase-FLASH) ENCODE datasets were aligned to Sanger reads containing satellite sequences. Fragment length vs. midpoint plots (V-plots) were constructed as previously described (Henikoff, et al. 2011).

**Analysis of ssDNA-seq data.** Raw reads were quality filtered and subjected to adapter trimming as described above.

**Analysis of motif similarity.** T.omtom v4.12.0 (Gupta, et al. 2007) with default parameters was used to determine similarity of yeast CDEI-derived motifs to a library of consensus sites for 203 yeast transcription factors (MacIsaac, et al. 2006) and to a Reb1 motif defined using high-resolution ChIP-seq (Kasinathan, et al. 2014).

**Source code.** Code for performing the described analyses is available from Github (https://github.com/sivakasinathan/cenpb).

**References**

Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y. 2001. Alpha-satellite DNA of primates: old and new families. Chromosoma 110:253-266.

Allshire RC, Karpen GH. 2008. Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nat Rev Genet 9:923-937.

Arnold C, Matthews LJ, Nunn CL. 2010. The 10kTrees website: A new online resource for primate phylogeny. Evolutionary Anthropology: Issues, News, and Reviews 19:114-118.

Aze A, Sannino V, Soffientini P, Bachi A, Costanzo V. 2016. Centromeric DNA replication reconstitution reveals DNA loops and ATR checkpoint suppression. Nat Cell Biol 18:684-691.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573-580.
Catania S, Pidoux AL, Allshire RC. 2015. Sequence features and transcriptional stalling within centromere DNA promote establishment of CENP-A chromatin. PLoS Genet 11:e1004986.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 40:D700-705.

Chmatal L, Gabriel SI, Mitsainas GP, Martinez-Vargas J, Ventura J, Searle JB, Schultz RM, Lampson MA. 2014. Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. Curr Biol 24:2295-2300.

Clarke L, Carbon J. 1985. The structure and function of yeast centromeres. Annu Rev Genet 19:29-55.

Copenhaver GP, Putnam CD, Denton ML, Pikaard CS. 1994. The RNA polymerase I transcription factor UBF is a sequence-tolerant HMG-box protein that can recognize structured nucleic acids. Nucleic Acids Res 22:2651-2657.

Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q, Karaca E, Chiarle R, Skrzypczak M, Ginalski K, et al. 2013. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat Methods 10:361-365.

Daniel A. 2002. Distortion of female meiotic segregation and reduced male fertility in human Robertsonian translocations: consistent with the centromere model of co-evolving centromere DNA/centromeric histone (CENP-A). Am J Med Genet 111:450-452.

Dawe RK, Henikoff S. 2006. Centromeres put epigenetics in the driver's seat. Trends Biochem Sci 31:662-669.

De Rop V, Padeganeh A, Maddox PS. 2012. CENP-A: the key player behind centromere identity, propagation, and kinetochore assembly. Chromosoma 121:527-538.

Drinnenberg IA, Henikoff S, Malik HS. 2016. Evolutionary Turnover of Kinetochore Proteins: A Ship of Theseus? Trends Cell Biol 26:498-510.

du Sart D, Cancilla MR, Earle E, Mao JI, Saffery R, Tainton KM, Kalitsis P, Martyn J, Barry AE, Choo KH. 1997. A functional neo-centromere formed through activation of a latent human centromere and consisting of non-alpha-satellite DNA. Nat Genet 16:144-153.

Ekwall K. 2007. Epigenetic control of centromere behavior. Annu Rev Genet 41:63-81.

Fachinetti D, Folco HD, Nechemia-Arbely Y, Valente LP, Nguyen K, Wong AJ, Zhu Q, Holland AJ, Desai A, Jansen LE, et al. 2013. A two-step mechanism for epigenetic specification of centromere identity and function. Nat Cell Biol 15:1056-1066.

Fachinetti D, Han JS, McMahon MA, Ly P, Abdullah A, Wong AJ, Cleveland DW. 2015. DNA Sequence-Specific Binding of CENP-B Enhances the Fidelity of Human Centromere Function. Dev Cell 33:314-327.

Fishman L, Saunders A. 2008. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. Science 322:1559-1562.
Foltz DR, Jansen LE, Bailey AO, Yates JR, 3rd, Bassett EA, Wood S, Black BE, Cleveland DW. 2009. Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. Cell 137:472-484.

Garavis M, Escaja N, Gabelica V, Villasante A, Gonzalez C. 2015. Centromeric Alpha-Satellite DNA Adopts Dimeric i-Motif Structures Capped by AT Hoogsteen Base Pairs. Chemistry 21:9816-9824.

Goldberg IG, Sawhney H, Pluta AF, Warburton PE, Earnshaw WC. 1996. Surprising deficiency of CENP-B binding sites in African green monkey alpha-satellite DNA: implications for CENP-B function at centromeres. Mol Cell Biol 16:5156-5168.

Gordon JL, Byrne KP, Wolfe KH. 2011. Mechanisms of chromosome number evolution in yeast. PLoS Genet 7:e1002190.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27:1017-1018.

Guenatri M, Bailly D, Maison C, Almouzni G. 2004. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. J Cell Biol 166:493-505.

Guerrero AA, Gamero MC, Trachana V, Futterer A, Pacios-Bras C, Diaz-Concha NP, Cigudosa JC, Martinez AC, van Wely KH. 2010. Centromere-localized breaks indicate the generation of DNA damage by the mitotic spindle. Proc Natl Acad Sci U S A 107:4159-4164.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. Genome Biol 8:R24.

Haaf T, Mater AG, Wienberg J, Ward DC. 1995. Presence and abundance of CENP-B box sequences in great ape subsets of primate-specific alpha-satellite DNA. J Mol Evol 41:487-491.

Hasson D, Panchenko T, Salimian KJ, Salman MU, Alonso A, Warburton PE, Black BE. 2013. The octamer is the major form of CENP-A nucleosomes at human centromeres. Nat Struct Mol Biol 20:687-695.

Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. 2011. Epigenome characterization at single base-pair resolution. Proc Natl Acad Sci U S A 108:18318-18323.

Henikoff JG, Thakur J, Kasinathan S, Henikoff S. 2015. A unique chromatin complex occupies young alpha-satellite arrays of human centromeres. Sci Adv 1.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293:1098-1102.

Henikoff S, Ramachandran S, Krassovsky K, Bryson TD, Codomo CA, Brogaard K, Widom J, Wang JP, Henikoff JG. 2014. The budding yeast Centromere DNA Element II wraps a stable Cse4 hemisome in either orientation in vivo. Elife 3:e01861.

Horz W, Altenburger W. 1981. Nucleotide sequence of mouse satellite DNA. Nucleic Acids Res 9:683-696.

Iwahara J, Kigawa T, Kitagawa K, Masumoto H, Okazaki T, Yokoyama S. 1998. A helix-turn-helix structure unit in human centromere protein B (CENP-B). EMBO J 17:827-837.
Iwata-Otsubo A, Dawicki-McKenna JM, Akera T, Falk SJ, Chmatal L, Yang K, Sullivan BA, Schultz RM, Lampson MA, Black BE. 2017. Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes that Drive in Female Meiosis. Curr Biol.

Jansen LE, Black BE, Foltz DR, Cleveland DW. 2007. Propagation of centromeric chromatin requires exit from mitosis. J Cell Biol 176:795-805.

Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. 2013. DNA-binding specificities of human transcription factors. Cell 152:327-339.

Jonstrup AT, Thomsen T, Wang Y, Knudsen BR, Koch J, Andersen AH. 2008. Hairpin structures formed by alpha satellite DNA of human centromeres are cleaved by human topoisomerase IIalpha. Nucleic Acids Res 36:6165-6174.

Joseph A, Mitchell AR, Miller OJ. 1989. The organization of the mouse satellite DNA at centromeres. Exp Cell Res 183:494-500.

Kapoor M, Montes de Oca Luna R, Liu G, Lozano G, Cummings C, Mancini M, Ouspenski I, Brinkley BR, May GS. 1998. The cenpB gene is not essential in mice. Chromosoma 107:570-576.

Karpen GH, Allshire RC. 1997. The case for epigenetic effects on centromere identity and function. Trends Genet 13:489-496.

Kasinathan S, Orsi GA, Zentner GE, Ahmad K, Henikoff S. 2014. High-resolution mapping of transcription factor binding sites on native chromatin. Nat Methods 11:203-209.

Kato T, Sato N, Hayama S, Yamabuki T, Ito T, Miyamoto M, Kondo S, Nakamura Y, Daigo Y. 2007. Activation of Holliday junction recognizing protein involved in the chromosomal stability and immortality of cancer cells. Cancer Res 67:8544-8553.

Kipling D, Warburton PE. 1997. Centromeres, CENP-B and Tigger too. Trends Genet 13:141-145.

Kobayashi N, Suzuki Y, Schoenfeld LW, Muller CA, Nieduszynski C, Wolfe KH, Tanaka TU. 2015. Discovery of an unconventional centromere in budding yeast redefines evolution of point centromeres. Curr Biol 25:2026-2033.

Koch J. 2000. Neocentromeres and alpha satellite: a proposed structural code for functional human centromere DNA. Hum Mol Genet 9:149-154.

Kouzine F, Wojtowicz D, Baranello L, Yamane A, Nelson S, Resch W, Kieffer-Kwon KR, Benham CJ, Casellas R, Przytycka TM, et al. 2017. Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. Cell Syst 4:344-356 e347.

Kouzine F, Wojtowicz D, Yamane A, Resch W, Kieffer-Kwon KR, Bandle R, Nelson S, Nakahashi H, Awasthi P, Feigenbaum L, et al. 2013. Global regulation of promoter melting in naive lymphocytes. Cell 153:988-999.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.
Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. Algorithms Mol Biol 6:26.

Lu S, Wang G, Bacolla A, Zhao J, Spitser S, Vasquez KM. 2015. Short Inverted Repeats Are Hotspots for Genetic Instability: Relevance to Cancer Genomes. Cell Rep.

Maclsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. 2006. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 7:113.

Malik HS, Henikoff S. 2009. Major evolutionary transitions in centromere complexity. Cell 138:1067-1082.

Masumoto H, Masukata H, Muro Y, Nozaki N, 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.

McDowall MD, Harris MA, Lock A, Rutherford K, Staines DM, Bahler J, Kersey PJ, Oliver SG, Wood V. 2015. PomBase 2015: updates to the fission yeast database. Nucleic Acids Res 43:D656-661.

McNulty SM, Sullivan BA. 2017. Centromere Silencing Mechanisms. Prog Mol Subcell Biol 56:233-255.

Mendiburo MJ, Padeken J, Fulop S, Schepers A, Heun P. 2011. Drosophila CENH3 is sufficient for centromere formation. Science 334:686-690.

Meraldi P, McInish AD, Rheinbay E, Sorger PK. 2006. Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. Genome Biol 7:R23.

Nardi IK, Zasadzinska E, Stellfox ME, Knippler CM, Foltz DR. 2016. Licensing of Centromeric Chromatin Assembly through the Mis18alpha-Mis18beta Heterotetramer. Mol Cell 61:774-787.

Nickol J, Martin RG. 1983. DNA stem-loop structures bind poorly to histone octamer cores. Proc Natl Acad Sci U S A 80:4669-4673.

Ohkuni K, Kitagawa K. 2011. Endogenous transcription at the centromere facilitates centromere activity in budding yeast. Curr Biol 21:1695-1703.

Ohno M, Fukagawa T, Lee JS, Ikemura T. 2002. Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. Chromosoma 111:201-213.

Ohzeki J, Nakano M, Okada T, Masumoto H. 2002. CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. J Cell Biol 159:765-775.

Pearson CE, Zorbas H, Price GB, Zannis-Hadjopoulos M. 1996. Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. J Cell Biochem 63:1-22.

Perpelescu M, Hori T, Toyoda A, Misu S, Monma N, Ikeo K, Obuse C, Fujiyama A, Fukagawa T. 2015. HJURP is involved in the expansion of centromeric chromatin. Mol Biol Cell 26:2742-2754.

Quenet D, Dalal Y. 2014. A long non-coding RNA is required for targeting centromeric protein A to the human centromere. Elife 3:e03254.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-277.

Rousseeuw PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53-65.

Saffery R, Irvine DV, Griffiths B, Kalitsis P, Wordeman L, Choo KH. 2000. Human centromeres and neocentromeres show identical distribution patterns of >20 functionally important kinetochore-associated proteins. Hum Mol Genet 9:175-185.

Sanchez-Pulido L, Pidoux AL, Ponting CP, Allshire RC. 2009. Common ancestry of the CENP-A chaperones Scm3 and HJURP. Cell 137:1173-1174.

Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, Rine J, Johnston M, Hittinger CT. 2011. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 (Bethesda) 1:11-25.

Schueler MG, Swanson W, Thomas PJ, Program NCS, Green ED. 2010. Adaptive evolution of foundation kinetochore proteins in primates. Mol Biol Evol 27:1585-1597.

Shang WH, Hori T, Martins NM, Toyoda A, Misu S, Monma N, Hiratani I, Maeshima K, Ikeo K, Fujiyama A, et al. 2013. Chromosome engineering allows the efficient isolation of vertebrate neocentromeres. Dev Cell 24:635-648.

Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, Sakakibara Y, Fujiyama A, Fukagawa T. 2010. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. Genome Res 20:1219-1228.

Skene PJ, Henikoff S. 2017a. CUT&RUN: Targeted in situ genome-wide profiling with high efficiency for low cell numbers. bioRxiv.

Skene PJ, Henikoff S. 2017b. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife 6:e21856.

Slee RB, Steiner CM, Herbert BS, Vance GH, Hickey RJ, Schwarz T, Christan S, Radovich M, Schneider BP, Schindelhauer D, et al. 2012. Cancer-associated alteration of pericentromeric heterochromatin may contribute to chromosome instability. Oncogene 31:3244-3253.

Sullivan KF, Glass CA. 1991. CENP-B is a highly conserved mammalian centromere protein with homology to the helix-loop-helix family of proteins. Chromosoma 100:360-370.

Thakur J, Talbert PB, Henikoff S. 2015. Inner Kinetochore Protein Interactions with Regional Centromeres of Fission Yeast. Genetics 201:543-561.

Warren WC, Jasinska AJ, Garcia-Perez R, Svardal H, Tomlinson C, Rocchi M, Archidiacono N, Capozzi O, Minx P, Montague MJ, et al. 2015. The genome of the vervet (Chlorocebus aethiops sabaeus). Genome Res 25:1921-1933.

Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, Northrop EL, Hannan R, Saffery R, Shaw ML, Williams E, et al. 2007. Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res 17:1146-1160.
Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, Cooper DN, Li Q, Li Y, van Gool AJ, et al. 2011. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol 29:1019-1023.

Yang H, Volfovsky N, Rattray A, Chen X, Tanaka H, Strathern J. 2014. GAP-Seq: a method for identification of DNA palindromes. BMC Genomics 15:394.

Yoda K, Nakamura T, Masumoto H, Suzuki N, Kitagawa K, Nakano M, Shinjo A, Okazaki T. 1996. Centromere protein B of African green monkey cells: gene structure, cellular expression, and centromeric localization. Mol Cell Biol 16:5169-5177.

Zhabinskaya D, Madden S, Benham CJ. 2015. SIST: stress-induced structural transitions in superhelical DNA. Bioinformatics 31:421-422.

Figure Legends

Figure 1 | Patterns of DNA dyad symmetry at eukaryotic centromeres. (A) Examples of dyad symmetries in centromeric DNA sampled randomly from human, African green monkey, mouse, chicken, and fission yeast whole-genome sequencing datasets. (B) Dyad density, which is defined for a given sequence as the total number of palindromic positions with palindrome length > 4 and spacer length < 20 normalized by sequence length, at centromeres relative to composition-matched background genomic regions. Asterisks indicate two-sample Kolmogorov-Smirnov p < 0.05. (C) Enrichment (relative to permuted sequence) of dyad symmetries over varying palindrome lengths in read ends mapping to centromeres or from sequences sampled from genome assemblies for a variety of organisms. The displayed phylogeny is based on NCBI Taxonomy annotations.

Figure 2 | Centromeric dyad symmetries are predicted to adopt non-B-form structures. (A) Scores from stress-induced structural transition (SIST) model predictions of DNA melting (left) and cruciform extrusion (right) for centromeric sequences and composition-matched background genomic regions from human, African green monkey, mouse, chicken, and fission yeast genomes. Asterisks indicate two-sample sample Kolmogorov-Smirnov p < 0.05. (B) Examples of minimum free energy secondary structure predictions for randomly selected α-satellite monomers from human and African green monkey. (C) DNA secondary structure folding free energy predictions for read ends mapping to centromeres or from sequences sampled from genome assemblies from the indicated species. The displayed phylogeny is based on NCBI Taxonomy annotations.

Figure 3 | Non-B-form DNA detected experimentally at dyad-depleted functional human and mouse centromeres. (A) Abundance of heterochromatin major (MaSat) and centromeric minor (MiSat) satellite
fragments (top) and the minimal CENP-B box (middle) in mouse whole-genome sequencing reads. (B) DNA secondary structure free energy distributions from computational prediction for MaSat and MiSat-containing Sanger reads. (C) Fraction of permanganate-seq reads from resting (R) and LPS-activated (A) cells mapping to the masked mouse genome (Masked mm10) or Sanger reads containing MiSat or MaSat monomers. (D) Examples of Sanger reads harboring MiSat and CENP-B box sequences (left) or devoid of predicted centromeric features (right) and signal from CENP-A ChIP-seq, permanganate-seq, and dyad symmetry analysis. (E) Correlation between total permanganate-seq signal from activated B cells (normalized to control) and input-normalized CENP-A occupancy for MiSat-containing Sanger reads. Scatter plot points are colored based on CENP-B box score tertile, where scores are defined as the sum of scores for all FIMO-defined CENP-B boxes occurring on a read. (F) Fraction of permanganate-seq reads aligning to the repeat-masked hg38 assembly and HuRef Sanger alphoid reads normalized to number of reads from whole-genome sequencing of HuRef mapping to the respective assemblies. (G) Examples of permanganate-seq, CENP-A ChIP, dyad symmetry, and predicted DNA melting and cruciform transition probabilities for functionally active (D5Z2) and inactive (D7Z2) α-satellite repeat arrays. Note that the CENP-A ChIP tracks are on different scales. (H) Correlation between total permanganate-seq signal and CENP-A occupancy for alphoid Sanger reads. Scatter plot points are colored based on CENP-B box score tertile.

Figure 4 | Primate CENP-B binding is restricted to dyad-depleted great ape centromeres. (A) Estimated abundance of α-satellite sequences in a sampling of simian primates. (B) Enrichment of minimal CENP-B box sequences in raw reads from whole-genome sequencing. The displayed phylogeny is a chronogram based on mitochondrial genomes and is adapted from the 10kTrees Project (Arnold, et al. 2010). Abundance of α-satellite sequence in aligned reads (C) and abundance of matches to the minimal CENP-B box sequence in raw reads (D) in CUT&RUN experiments performed in human (K562) and African green monkey (Cos-7) cell lines relative normalized to α-satellite abundance in whole-genome sequencing reads.

Figure 5 | The human Y centromere and vertebrate neocentromeres are associated with dyad symmetries and non-B-form DNA. (A) Predicted ensemble free energies for DYZ3 and non-DYZ3 alphoid satellites. (B) Examples of minimum free energy structures for a DYZ3 and D5Z2 alphoid fragments. A human chromosome 13 neocentromere (C) and a chicken chrZ neocentromere (D) with profiles from CENP-A ChIP-seq and SIST-predicted DNA melting and cruciform extrusion probabilities (left panels). Dyad symmetry and SIST DNA melting and cruciform extrusion scores for neocentromeres (“neo”) and composition-matched non-centromeric background genomic intervals (right panels). Data
from genome-wide analysis of palindrome formation with sequencing (GAP-seq), which was performed in human cell lines, is also included in (A). Asterisks indicate two-sample Kolmogorov-Smirnov $p < 0.05$.

**Figure 6 | Centromeric dyad symmetries are features of yeast centromeres.** Average CenH3 signal, dyad symmetry, and SIST melt and cruciform profiles (left panels) and comparison of dyad densities and SIST melt and cruciform scores for centromeres versus nucleotide composition-matched, randomly selected genomic intervals (right panels) in *S. cerevisiae* (A). (B) Predicted ensemble free energy distributions for centromeric sequences from *sensu strictu* and *sensu lato* saccharomycetes with well-annotated genomes. (C) Enriched CDEI motifs for saccharomycetes and average estimated dyad densities over CDEII.

**Figure 7 | Models for genetic centromere specification.** (A) Summary of centromeric DNA sequence type, association with helix-deforming DNA binding protein, dyad symmetry, and predicted secondary structure forming tendency for various eukaryotes. (B) Repetitive centromeres vary in their predilection for forming cruciform structures exemplified by alphoid sequences of Old World Monkeys, which are predicted to form stable non-B DNA structures, and great apes, which do not preferentially adopt non-B DNA structures. In great apes, CENP-B binding may facilitate formation of non-B-form DNA such as cruciforms. Cruciform structures are recognized by HJURP/Scm3 chaperones, which deposit CENP-A nucleosomes. (C) Alternatively, OWM AS units may be spontaneously transcribed, while CENP-B binding may facilitate transcription of great ape alphoid units, with the RNAs contributing to deposition of CENP-A.
Supplementary Figures & Tables

Figure S1 | Characterization of variation in vertebrate centromeres, Related to Figures 1,2,4, and 5.

(A) Heatmap representations of histograms of lengths of tandem repeats discovered de novo in datasets for great apes and Old World Monkeys (OWMs). Sequences from indicated peaks at ~170-bp and ~340-bp were used to generate species-specific repeat libraries. Where available, raw shotgun Sanger reads were used. If Sanger reads were unavailable, the genome assembly (including unplaced contigs) was scanned for tandem repeats (indicated by an asterisk); “ND” indicates neither Sanger reads nor an assembly were available. (B) Fraction of alphoid reads containing matches to the SELEX-defined CENP-B box motif deposited in JASPAR. Examples of dinucleotide profiles for centromeric regions, neocentromeres, and non-centromeric background genomic regions selected by k-mer matching from human (C) and chicken (D).

Figure S2 | Clade-specific secondary structure-forming dyad symmetries at vertebrate centromeres, Related to Figures 1 and 2.

(A) Enrichment (vs. permuted sequence) of dyad symmetries with varying palindrome and spacer lengths for great apes and Old World Monkeys (OWMs) allowing one mismatch in the palindromic region. (B) Two-sample Kolmogorov-Smirnov p-values for all pairwise comparisons of distributions of RNAfold-predicted ensemble free energies for great ape, OWM, mouse, and chicken centromeric sequences. Heatmap depicts -log(p-value) with warmer (red) colors indicating dissimilarity and colder (blue) colors indicating similarity.

Figure S3 | Analysis of yeast centromeric DNAs, Related to Figure 6.

(A) Dinucleotide composition at S. pombe centromeres and matched non-centromeric background regions sampled from the fission yeast genome assembly. (B) Dinucleotide composition at centromeres and randomly-selected, matched non-centromeric regions from the genomes of selected sensu strictu and sensu lato saccharomycete genomes. (C) Dyad symmetry density and DNA melting and cruciform extrusion predilections at centromeres and composition-matched non-centromeric random regions from the genomes of selected saccharomycetes. Asterisks indicate two-sample Kolmogorov-Smirnov p < 0.05. (D) Two-sample Kolmogorov-Smirnov p-values for all pairwise comparisons of RNAfold-predicted ensemble free energies for centromeric sequences from saccharomycetes. (E) Tomtom motif similarity analysis of CDEI motifs from two sensu lato yeasts with the highest scoring hit from scanning a database of binding sites for ~200 yeast transcription factors (MacIsaac, et al. 2006) and comparison to the Reb1 binding site defined by ORGANIC profiling, a high-resolution mapping approach (Kasinathan, et al. 2014).
Table S1 | Publicly available datasets analyzed in this study, Related to STAR Methods. NCBI Sequence Read Archive (SRA) accession numbers and references for whole-genome (WGS), chromatin immunoprecipitation (ChIP-seq), genome-wide analysis of palindrome formation (GAP-seq), permanganate-seq, and DNase-seq Illumina datasets are provided. Where references are unavailable, NCBI BioProject accessions for public reference genome projects are provided.
Kasinathan & Henikoff – Figure 1

A

B

C

Dyad symmetry

Palindrome length (bp)

Great apes

OWMs

Human
Chimpanzee
Bonobo
Western gorilla
Eastern gorilla
Bornean orangutan
Sumatran orangutan
African green monkey
Vervet
Green monkey
Rhesus macaque
Crab-eating macaque
Olive baboon
Mouse
Horse
Chicken
Stickleback
Fission yeast
Rice
Maize
Thale cress

0 0.25

Enrichment log2 ratio (vs. permuted control)
A. Predicted DNA secondary structure

B. DNA melting and Cruciform extrusion for different species:
- Human
- African Green Monkey
- Mouse
- Chicken
- Fission yeast

C. Heatmap showing DNA melting and Cruciform extrusion scores for various species and OMes.
Kasinathan & Henikoff – Figure 3

A. Mouse

- Satellite abundance
- CENP-B box abundance
- DNA secondary structure

B. CENP-B box score

- Mouse: 0.38
- Human: 0.95

C. Heterochromatin vs Centromeric

D. Non-B-form DNA

- Mouse: 0.45
- Human: 0.38

E. Permanganate-seq signal

F. Human

- Non-B-form DNA

G. CENP-B box score

- Mouse: 0.35
- Human: 0.95

H. Permanganate-seq signal

- Mouse: 0.45
- Human: 0.38
Kasinathan & Henikoff – Figure 5

A

Predicted secondary structure

Density

-0.2

-0.1

0

0.1

0.2

-90

-15

-5

Ensemble ∆G (kcal/mol)

B

DYZ3

ΔG = –15.89 kcal/mol

DSZ2

ΔG = –8.43 kcal/mol

C

Human

CENP-A ChIP

GAP-seq

Dyad symmetry

DNA melting

Cruc. extrusion

Neo Control

D

Chicken

CENP-A ChIP

Dyad symmetry

DNA melting

Cruc. extrusion

Neo Control
### A

| Centromere type | Native | Neocen. | Native |
|-----------------|--------|---------|--------|
| Species         | OWMs   |         |        |
|                 | Chicken|         | Mouse  |
|                 | Fission yeast |       | Yeast  |
|                 | Human  |         | Yeast  |
|                 | Human, great ape |   | Yeast |

Helix-deforming DNA-binding factor

Dyad symmetry enrichment

Predicted stable secondary structure

### B

**DNA**

**Spontaneous non-B DNA**

1. Non-B DNA formation
2. HJURP/Scm3 recruitment
3. CenH3 loading

**Induced non-B DNA**

- Tf
- ~60°

### C

**Spontaneous transcription**

1. Template melting
2. Transcription
3. CenH3 loading

**Induced transcription**

- RNA
- RNA pol

---

The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under CC-BY 4.0 International license.
Supplementary Figures & Tables

Figure S1 | Characterization of variation in vertebrate centromeres, Related to Figures 1, 2, 4, and 5.
(A) Heatmap representations of histograms of lengths of tandem repeats discovered de novo in datasets for great apes and Old World Monkeys (OWMs). Sequences from indicated peaks at ~170-bp and ~340-bp were used to generate species-specific repeat libraries. Where available, raw shotgun Sanger reads were used. If Sanger reads were unavailable, the genome assembly (including unplaced contigs) was scanned for tandem repeats (indicated by an asterisk); “ND” indicates neither Sanger reads nor an assembly were available. (B) Fraction of alphoid reads containing matches to the SELEX-defined CENP-B box motif deposited in JASPAR. Examples of dinucleotide profiles for centromeric regions, neocentromeres, and non-centromeric background genomic regions selected by k-mer matching from human (C) and chicken (D).

Figure S2 | Clade-specific secondary structure-forming dyad symmetries at vertebrate centromeres, Related to Figures 1 and 2. (A) Enrichment (vs. permuted sequence) of dyad symmetries with varying palindrome and spacer lengths for great apes and Old World Monkeys (OWMs) allowing one mismatch in the palindromic region. (B) Two-sample Kolmogorov-Smirnov p-values for all pairwise comparisons of distributions of RNAfold-predicted ensemble free energies for great ape, OWM, mouse, and chicken centromeric sequences. Heatmap depicts -log(p-value) with warmer (red) colors indicating dissimilarity and colder (blue) colors indicating similarity.

Figure S3 | Analysis of yeast centromeric DNAs, Related to Figure 6. (A) Dinucleotide composition at S. pombe centromeres and matched non-centromeric background regions sampled from the fission yeast genome assembly. (B) Dinucleotide composition at centromeres and randomly-selected, matched non-centromeric regions from the genomes of selected sensu strictu and sensu lato saccharomycete genomes. (C) Dyad symmetry density and DNA melting and cruciform extrusion predilections at centromeres and composition-matched non-centromeric random regions from the genomes of selected saccharomycetes. Asterisks indicate two-sample Kolmogorov-Smirnov p < 0.05. (D) Two-sample Kolmogorov-Smirnov p-values for all pairwise comparisons of RNAfold-predicted ensemble free energies for centromeric sequences from saccharomycetes. (E) Tomtom motif similarity analysis of CDEI motifs from two sensu lato yeasts with the highest scoring hit from scanning a database of binding sites for ~200 yeast transcription factors (MacIsaac, et al. 2006) and comparison to the Reb1 binding site defined by ORGANIC profiling, a high-resolution mapping approach (Kasinathan, et al. 2014).
Fission yeast

Control regions | Centromeric regions

Sequence composition

Budding yeasts

Sequence composition

S. cerevisiae
S. mikatae
S. kudriavzevii
S. dairenensis
S. castellii

Sensu strictu
Sensu lato

Similar
Dissimilar

Dyad density

DNA melting

Cruciform extrusion

Top MacIsaac motif match
Similarity to ORGANIC Reb1

S. dairenensis
Reb1 \( (p < 1e-5) \)
Yes \( (p < 1e-5) \)

S. castellii
Reb1 \( (p < 1e-5) \)
Yes \( (p < 1e-5) \)
Table S1 | Publicly available datasets analyzed in this study, Related to STAR Methods. NCBI Sequence Read Archive (SRA) accession numbers and references for whole-genome (WGS), chromatin immunoprecipitation (ChIP-seq), genome-wide analysis of palindrome formation (GAP-seq), permanganate-seq, and DNase-seq Illumina datasets are provided. Where references are unavailable, NCBI BioProject accessions for public reference genome projects are provided.
Table S1 | Publicly available datasets analyzed in this study. NCBI Sequence Read Archive (SRA) accession numbers and references for whole-genome (WGS), chromatin immunoprecipitation (ChIP-seq), genome-wide analysis of palindrome formation (GAP-seq), permanganate-seq, and DNase-seq Illumina datasets are provided. Where references are unavailable, NCBI BioProject accessions for public reference genome projects are provided.

| Species                         | Data type                      | SRA Accession(s)          | Reference                  |
|---------------------------------|--------------------------------|---------------------------|----------------------------|
| *Homo sapiens*                  | WGS (NA12878)                  | ERR091571, ERR091572      | (Redon et al., 2006)       |
|                                 | WGS (HeLa)                     | SRR5009881                | (Lee et al., 2014)         |
|                                 | WGS (K562)                     | SRR866898                 | (Mercer et al., 2013)      |
|                                 | CENP-A ChIP-seq & input (HuRef)| SRR766737, SRR766740      | (Henikoff et al., 2015)    |
|                                 | CENP-A ChIP-seq & input (IMS13g)| SRR766738, SRR766741      | (Hasson et al., 2013)      |
|                                 | CENP-A ChIP-seq & input (MS4221)| SRR766736, SRR766739     |                            |
|                                 | CENP-A ChIP-seq & input (PDNC4)| SRR766735, SRR766739     |                            |
|                                 | GAP-seq (IMR90)                | SRR942734                 | (Yang et al., 2014)        |
|                                 | GAP-seq (Colo320)              | SRR942724                 |                            |
|                                 | Permanganate-seq               | SRR823508-14              | (Kouzine et al., 2013)     |
| Chimpanzee (Pan troglodytes)    | WGS                            | SRR726233, SRR726241, SRR726242, SRR726275 | (Prado-Martinez et al., 2013) |
| Bonobo (Pan paniscus)           | WGS                            | SRR726612-4, SRR740772    |                            |
| Western gorilla (Gorilla gorilla) | WGS                          | SRR747963-6               |                            |
| Eastern gorilla (Gorilla beringei) | WGS                        | SRR747651, SRR747652, SRR747654, SRR747655 |                            |
| Sumatran orangutan (Pongo abelii) | WGS                        | SRR748024-7               |                            |
| Bornean orangutan (Pongo pygmaeus) | WGS                        | SRR747999, SRR748001, SRR748002, SRR748004 |                            |
| African green monkey (Chlorocebus aethiops) | WGS                  | SRR3709478-81, SRR556098, SRR556110, SRR556111, SRR556113 | PRJNA168472 |
| Green monkey (Chlorocebus sabaeus) | WGS                        | SRR1660260, SRR1660288, SRR1186326, SRR1186403, SRR1186633, SRR1186665, SRR1660261, SRR1186683 | (Warren et al., 2015)     |
| Species                        | Library Type          | SRA Numbers | Project Accession |
|-------------------------------|-----------------------|-------------|-------------------|
| Vervet (Chlorocebus pygerythrus) | WGS                   | SRR556129, SRR556168, SRR556091, SRR556092, SRR556099, SRR556101 | PRJNA168521 |
| Rhesus macaque (Macaca mulatta) | WGS                   | SRR1927121 | (Xue et al., 2016) |
| Crab-eating macaque (Macaca fascicularis) | WGS                   | SRR445517, SRR445602-5, SRR445610-13, SRR445630, SRR445633, SRR445635, SRR445658-62 | PRJNA20409 |
| Olive baboon (Papio anubis) | WGS                   | SRR927654, SRR927659, SRR1515161, SRR1515162, SRR1516196 | PRJNA54005 |
| Mouse (Mus musculus) | WGS                   | SRR091273, SRR091274, SRR091250, SRR091253 | PRJNA60381 |
|                                | Permanganate-seq     | SRR823425, SRR823427, SRR823428, SRR823430, SRR823433-8, SRR823440-6 | (Kouzine et al., 2013) |
| Chicken (Gallus gallus) | CENP-A ChIP-seq (DT40) | DRR018430, DRR018431 | (Hori et al., 2014) |
|                                | CENP-A ChIP-seq (BM23) |                      |                   |
| Budding yeast (Saccharomyces cerevisiae) | Cse4 ChIP-seq       | SRR1139246 | (Henikoff et al., 2014) |
| Fission yeast (Schizosaccharomyces pombe) | Cnp1 ChIP-seq       | SRR1722846 | (Thakur et al., 2015) |
References

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.

Henikoff, J.G., Thakur, J., Kasinathan, S., and Henikoff, S. (2015). A unique chromatin complex occupies young alpha-satellite arrays of human centromeres. Sci Adv 1.

Henikoff, S., Ramachandran, S., Krassovsky, K., Bryson, T.D., Codomo, C.A., Brogaard, K., Widom, J., Wang, J.P., and Henikoff, J.G. (2014). The budding yeast Centromere DNA Element II wraps a stable Cse4 hemisome in either orientation in vivo. Elife 3, e01861.

Hori, T., Shang, W.H., Toyoda, A., Misu, S., Monma, N., Ikeo, K., Molina, O., Vargiu, G., Fujiyama, A., Kimura, H., et al. (2014). Histone H4 Lys 20 monomethylation of the CENP-A nucleosome is essential for kinetochore assembly. Dev Cell 29, 740-749.

Kouzine, F., Wojtowicz, D., Yamane, A., Resch, W., Kieffer-Kwon, K.R., Bandle, R., Nelson, S., Nakahashi, H., Awasthi, P., Feigenbaum, L., et al. (2013). Global regulation of promoter melting in naive lymphocytes. Cell 153, 988-999.

Lee, M., Hills, M., Conomos, D., Stutz, M.D., Dagg, R.A., Lau, L.M., Reddel, R.R., and Pickett, H.A. (2014). Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. Nucleic Acids Res 42, 1733-1746.

Mercer, T.R., Edwards, S.L., Clark, M.B., Neph, S.J., Wang, H., Stergachis, A.B., John, S., Sandstrom, R., Li, G., Sandhu, K.S., et al. (2013). DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. Nat Genet 45, 852-859.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., et al. (2013). Great ape genetic diversity and population history. Nature 499, 471-475.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature 444, 444-454.

Thakur, J., Talbert, P.B., and Henikoff, S. (2015). Inner Kinetochore Protein Interactions with Regional Centromeres of Fission Yeast. Genetics 201, 543-561.

Warren, W.C., Jasinska, A.J., Garcia-Perez, R., Svardal, H., Tomlinson, C., Rocchi, M., Archidiacono, N., Capozzi, O., Minx, P., Montague, M.J., et al. (2015). The genome of the vervet (Chlorocebus aethiops sabaeus). Genome Res 25, 1921-1933.

Xue, C., Raveendran, M., Harris, R.A., Fawcett, G.L., Liu, X., White, S., Dahdouli, M., Rio Deiros, D., Below, J.E., Salerno, W., et al. (2016). The population genomics of rhesus macaques (Macaca mulatta) based on whole-genome sequences. Genome Res 26, 1651-1662.

Yang, H., Volfovsky, N., Rattray, A., Chen, X., Tanaka, H., and Strathern, J. (2014). GAP-Seq: a method for identification of DNA palindromes. BMC Genomics 15, 394.