Non-B-form DNA is enriched at centromeres

Sivakanthan Kasinathan[1,2] & Steven Henikoff[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@fhcrc.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-form 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. This may resolve the CENP-B paradox and provide a general basis for centromere specification.

Introduction

Centromeres are chromosomal regions that interact with the spindle apparatus during each cell division to ensure disjunction of chromosomes. In many eukaryotes, centromeres are made up of species-specific kilobase- to megabase-scale arrays of tandemly repeated satellite DNAs (Melters, et al. 2013). Centromeric repeats are thought to 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 DNAs may therefore serve as a molecular basis for speciation (Henikoff, et al. 2001). This centromere drive model is supported by observations confirming that alterations to centromeric DNA distort chromosome transmission in human (Daniel 2002), mouse (Chmatal, et al. 2014; Iwata-Otsubo, et al. 2017) and plants (Fishman and Saunders 2008). Critically, this hypothesis supposes that centromeres are genetically defined at least through a transient stage of their evolution (Dawe and Henikoff 2006).

Although the recognition of centromeres as chromosomal primary constrictions arguably predates genetics itself (Flemming 1882), centromere identity is thought to be determined independently of DNA sequence by the presence of nucleosomes containing the histone H3 variant CenH3 (CENP-A) (Karpen and Allshire 1997; Ekwall 2007; Allshire and Karpen 2008). Indeed, CENP-A has features of a heritable mark capable of self-propagation: it 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). This widely held "epigenetic" conception of the centromere is further supported by the existence

of neocentromeres, which form at ectopic chromosomal loci (du Sart, et al. 1997) and lack shared sequence features (Burrack and Berman 2012).

Mechanisms dictating selection of particular loci as centromeres have remained elusive in most organisms with the exception of budding yeasts, which have characteristic ~120-bp sequences that fully specify centromeres (Meraldi, et al. 2006; Gordon, et al. 2011). Although recent studies have highlighted a role for genetic variation in centromere function (Aldrup-MacDonald, et al. 2016), the search for shared genetic features has only led to more questions surrounding the role of DNA in centromere specification. There are two major ways in which centromere identity might be templated by DNA: recruitment of sequence-specific DNA binding proteins (as in budding yeasts) and/or by recognition of an emergent feature of the sequence itself such as DNA secondary structure. 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, while 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).

That centromeres may be specified by genetically-encoded structural features similar to G-quadruplexes of telomeres (Villasante, et al. 2007) remains highly speculative. In this context, studies in the fission yeast *Schizosaccharomyces pombe*, which has non-satellite centromeres of several kilobases, and limited analyses of primate and budding yeast centromeres have suggested the functional significance of DNA dyad symmetries (Koch 2000; Catania, et al. 2015), which may adopt non-B-form conformations such as stem-loops or cruciforms (Hamer and Thomas 1974; Pearson, et al. 1996). Indeed, various types of non-B-form structures such as single-stranded DNA, hairpins, triplexes, R-loops, and i-motifs have been observed *in vitro* and/or *in vivo* in centromeric DNA from a variety of organisms including human (Zhu, et al. 1996; Ohno, et al. 2002; Jonstrup, et al. 2008; Garavis, Escaja, et al. 2015; Garavis, Mendez-Lago, et al. 2015; Aze, et al. 2016; Kabeche, et al. 2017). Reconciling these observations concerning a role for DNA in specification of centromere identity with evidence that centromeres are epigenetically determined 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

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 annotations or libraries of species-specific satellite consensus sequences derived from *de novo* tandem repeat detection (**fig. S1A**), we scanned deep (>10X coverage) publicly available paired-end, whole genome sequencing datasets or genome assemblies from a sampling of vertebrates, fission yeast, and plants to identify centromeric sequences. Visual examination of centromeric sequences identified dyad symmetries with species-specific variation (**fig. 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 (**figs. 1B, S1B**). 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 (**fig. 1C**). 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-form DNA conformations (Pearson, et al. 1996) and reports of a variety of these structures at centromeres (Ohno, et al. 2002;

Jonstrup, et al. 2008; Garavis, Escaja, et al. 2015; Garavis, Mendez-Lago, 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-form conformations. Comparing SIST DNA melting and cruciform extrusion scores for centromeric sequences and dinucleotide composition-matched non-centromeric background genomic intervals for select species (**fig. S2**), we found that dyad-enriched centromeres were associated with significantly higher levels of predicted non-B-form DNA (**fig. 2A**). We then used RNAfold (Lorenz, et al. 2011), an independent approach for predict folding free energies, and found that species predicted by SIST to adopt non-B-form DNA tended to form more stable secondary structures (**fig. 2B**). For example, consistent with the SIST predictions, the distributions of predicted free energies for dyad-enriched centromeric 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; **figs. 2B,C and S3**). Similar trends were observed in other species (**fig. 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 (**fig. 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 (**fig. 3A**). Importantly, unlike centromeric MiSat sequences, heterochromatic MaSat arrays were predicted to more favorably adopt non-B-form structures (**fig. 3B**). MaSat sequences were enriched for permanganate-seq reads concordant with structure predictions (**fig. 3C**). Surprisingly, MiSat sequences were also highly enriched for permanganate-seq signal (**fig. 3C**). Permanganate-seq signal increased in activated B-cells, which are undergoing cell division, relative to quiescent resting B-cells (**fig. 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 $\rho = 0.45$; p << 1e-10),

with MiSat sequences associated with low-scoring CENP-B sites having the least CENP-A and permanganate-seq alignments (**fig. 3D–E**).

Similarly, contrary to the predicted poor tendency for human alphoid DNA to adopt non B-form DNA structures (**fig. 2C**), more than half the human permanganate-seq data aligned to α-satellite, representing a ~10-fold enrichment for alphoid DNA (**fig. 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 non-functional α-satellite arrays that are not at centromeres (Slee, et al. 2012) and show low CENP-A and permanganate-seq signals (**fig. 3G**). Indeed, CENP-A occupancy and permanganate-seq signal (Spearman's $\rho = 0.35$; $p << 1e\text{-}10$), with alphoid arrays associated with low-scoring CENP-B sites having the least CENP-A and permanganate-seq signal (**fig. 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-form 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 (TTCGNNNNANNCGGG) required to support CENP-B DNA binding (Iwahara, et al. 1998) in randomly sampled whole-genome sequencing reads. Although OWMs tended to have more α-satellite (**fig. 4A**), CENP-B boxes were highly enriched in great apes whereas OWM genomes contained negligible matches to the minimal CENP-B box (**fig. 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) (**fig. S4**). 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 *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 (**fig. 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 (**fig. 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 (**fig. 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 *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-form 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-form structures (**fig. 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 three human cell lines containing neocentromeres on chromosomes 4, 8, and 13 (Hasson, et al. 2013) and detected enrichment for dyad symmetries predicted to form stable secondary structures in CENP-A-associated neocentromere domains (**fig. 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 (**figs. 5C, S5**). Neocentromeres were also markedly enriched for dyad symmetries relative to base composition-matched randomly selected genomic regions and native centromeric sequences (**figs. 5C, S2**). We also analyzed CENP-A ChIP-seq data from a chicken cell line bearing a Z chromosome neocentromere (Hori, et al. 2014) to determine whether a similar trend is generally observed in vertebrates. Like other mammalian centromeres, this chicken neocentromere was also enriched for short dyad sequences and predicted to undergo strand separation and cruciform transitions (**fig. 5D**). Taken together, these analyses suggest that native centromeres depleted for CENP-B boxes such as the human Y chromosome centromere and vertebrate 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-form DNA at *S. cerevisiae* centromeres that were enriched for dyad symmetries compared to composition-matched non-centromeric genomic regions (**figs. 6A and S2**). 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. dairenensis* had comparatively less dyad symmetry and lower SIST DNA melting and cruciform extrusion scores (**figs. 6C and S6A,B**). Recently, *S. castellii* and *S. dairenensis* 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 (**fig. 6B,C**). We found that the consensus site at CDEI of dyad-depleted *sensu lato* yeasts is strongly predicted ($p < 1e\text{-}5$) to bind the DNA-bending general regulatory factor Reb1 (**figs. 6B,C and S6C)** 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 (**fig. 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-form 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 α-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 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 α-satellite *in vitro* and/or *in vivo* (Ohno, et al. 2002; Jonstrup, et al. 2008; Garavis, Escaja, et al. 2015; Garavis, Mendez-Lago, 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 (**fig. 7B,C**). One possibility is that non-B-form DNA directly specifies centromere identity (**fig. 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-form DNA structures. It is intriguing that both mammalian CENP-B and yeast Reb1 bend DNA ~60º (Tanaka, et al. 2001; Jaiswal, et al. 2016), 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 (**fig. 7C**). RNAs that adopt specific secondary structures may play an architectural role in the centromere or, the act of transcription itself may encourage histone turnover, permitting the cell-cycle-dependent incorporation of CENP-A. Transcription may occur readily at centromeres that adopt non-B-form structures such as melted DNA, while sequences that are comparatively resistant to non-B-form conformations may require the action of a 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-form 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). A context-specific role for helix-deforming DNA-binding proteins in facilitating DNA secondary structure formation and/or transcription suggests a possible resolution to the CENP-B paradox. 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.

The models proposed here could be tested using a variety of well-validated experimental approaches. For example, a strong prediction of these results is that OWM α-satellite, which lacks CENP-B boxes, will efficiently form centromeres upon transfer into great ape cells in an artificial chromosome/minichromosome assay (Harrington, et al. 1997). Interestingly, part of the converse experiment in which human α-satellite is introduced into OWM monkey cells has already been performed and demonstrated that the exogenous DNA may form centromeres (Haaf, et al. 1992). Using the same artificial chromosome assay, engineered synthetic sequences or mutation series of centromeric satellite with varying tendency to adopt different secondary structures could also be used to probe the role of non-B-form DNA in centromere identity. Modulation of transcript levels with RNA interference or control of transcription at specific satellite repeats using CRISPR-Cas9 transcriptional regulators (Didovyk, et al. 2016) may provide insight into the contribution of RNAs to centromere specification. These latter experiments may also aid in establishing the relative importance of DNA and RNA structures.

This view unifies the genetic and epigenetic conceptions of the centromere by explaining why CENP-B is necessary for *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-form conformations due to sequence and/or chromatin features. This also satisfies the requirement for transient genetic definition of centromeres required by the centromere drive hypothesis (Dawe and Henikoff 2006). Centromere specification by recognition of nucleic acid structures permits 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/) 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-discordant --overlap --dovetail -I 10 -X 700

Alignment parameters used for single-end reads were:

--very-sensitive --no-unal --non-deterministic

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^T) and *S. kudriavzevii* (IFO 1815^T) ultra-scaffolds were previously published (Scannell, et al. 2011) and are available online (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; 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.

Rationale for selection of functional centromeric sequences. Centromeric sequences are referred to as "functional" based on published interaction with CENP-A. In great apes, functional sequences account for

a majority of alphoid DNA. For example, in human, the two major CENP-A-associated alphoid variants account for ~70% of all α-satellite (Henikoff, et al. 2015). In Old World Monkeys, the alphoid fraction is highly homogenous with extremely low (~1-5%) inter-monomer or inter-dimer divergence (Musich, et al. 1980; Thayer, et al. 1981). In **figs. 1 and 2**, we proceeded with the heuristic that all alphoid sequences from primates are centromeric. In mouse, the functional sequence is minor satellite (Joseph, et al. 1989). Chicken has both repetitive and non-repetitive centromeres (Shang, et al. 2010); given quality of genome assembly/annotation and data availability, we analyzed only the non-repetitive chicken centromeric sequences. In fission yeast, we analyzed the "central core" region of the centromere, which binds CENP-A and is genetically required for centromere function (Polizzi and Clarke 1991; Takahashi, et al. 2000). Horse (Cerutti, et al. 2016), stickleback (Cech and Peichel 2015), rice (Cheng, et al. 2002), maize (Wolfgruber, et al. 2009), and *Arabidopsis thaliana* (Copenhaver, et al. 1999) centromeric sequences were identified by homology to published consensus sequences using BLAST as described above.

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 (TTCGNNNNANNCGGG), which is required for DNA-binding, were counted in Illumina WGS data. We further used the SELEX-defined CENP-B motif (JASPAR accession MA0637.1) (Jolma, et al. 2013) 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 ^{o}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 ^{o}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 ^{o}C, washed twice in Dig-wash and resuspended in 100 μL Dig-wash. Tubes

were placed 0 °C, mixed with 2 µL 100 mM $CaCl_2$, 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 glycogen and 200 pg 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 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):

-minpallen 5 -maxpallen 100 -gaplimit 20 -nummismatches X –overlap

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 (v2.3.5) (Lorenz, et al. 2011) was used to predict folding free energies. RNAfold was used with the following parameters for DNA secondary structure prediction:

–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).

Simulation of short-read sequencing. For the budding and fission yeasts and unique chicken centromere

sequence, given the absence of tandem repeats presenting an alignment challenge, we chose to use a simulation approach to permit direct comparison of these sequences with the data from organisms that have satellite centromeres. We simulate 100x100-bp paired-end Illumina reads using ART (version: MountRainier-2016-06-05) (Huang, et al. 2012) with the following parameters: -ss HS20 -l 100 -p -f 2

Analysis of human higher order repeats. To examine monomers from different human α-satellite higher-order repeats, we analyzed monomers from several human higher order repeats. Arrays that demonstrated centromere function in an artificial chromosome assay (Hayden, et al. 2013) were further classified as 'active' (D5Z2, DXZ1, D7Z1, and DYZ3) or 'competent' (D17Z1B, D11Z1, D7Z2, D4Z2bn) based on extent of CENP-A ChIP-seq enrichment (Henikoff, et al. 2015). Arrays that did not demonstrate centromerization in the artificial chromosome assay (D19Z1, Xmono, 3mono, and D5Z1) were classified as 'inactive' (Henikoff, et al. 2015). Monomers from these arrays were identified using BLAST as described above and subsequently used to retrieve sequences from the ChIP input in our previously published 100x100-bp paired-end CENP-A ChIP-seq data (Henikoff, et al. 2015). For Xmono and 3mono, few high-quality hits were recovered; the short-read simulation strategy described above was used with the HOR monomers used as reference sequences. Retrieved reads or simulated reads were then subjected to analysis with RNAfold as described above.

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 (ρ) 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 ssDNA-seq data. Raw reads were quality filtered and subjected to adapter trimming as described above with mapping performed as described earlier.

Analysis of motif similarity. Tomtom 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

Aldrup-MacDonald ME, Kuo ME, Sullivan LL, Chew K, Sullivan BA. 2016. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. Genome Res 26:1301-1311.

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.

Burrack LS, Berman J. 2012. Neocentromeres and epigenetically inherited features of centromeres. Chromosome Res 20:607-619.

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.

Cech JN, Peichel CL. 2015. Identification of the centromeric repeat in the threespine stickleback fish (Gasterosteus aculeatus). Chromosome Res 23:767-779.

Cerutti F, Gamba R, Mazzagatti A, Piras FM, Cappelletti E, Belloni E, Nergadze SG, Raimondi E, Giulotto E. 2016. The major horse satellite DNA family is associated with centromere competence. Mol Cytogenet 9:35.

Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, Gu M, Blattner FR, Jiang J. 2002. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell 14:1691-1704.

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, Nickel K, Kuromori T, Benito MI, Kaul S, Lin X, Bevan M, Murphy G, Harris B, Parnell LD, et al. 1999. Genetic definition and sequence analysis of Arabidopsis centromeres. Science 286:2468-2474.

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.

Didovyk A, Borek B, Tsimring L, Hasty J. 2016. Transcriptional regulation with CRISPR-Cas9: principles, advances, and applications. Curr Opin Biotechnol 40:177-184.

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.

Flemming W. 1882. Zellsubstanz, Kern und Zelltheilung. Leipzig: F. C. W. Vogel.

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.

Garavis M, Mendez-Lago M, Gabelica V, Whitehead SL, Gonzalez C, Villasante A. 2015. The structure of an endogenous Drosophila centromere reveals the prevalence of tandemly repeated sequences able to form i-motifs. Sci Rep 5:13307.

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.

Haaf T, Warburton PE, Willard HF. 1992. Integration of human alpha-satellite DNA into simian chromosomes: centromere protein binding and disruption of normal chromosome segregation. Cell 70:681-696.

Hamer DH, Thomas CA, Jr. 1974. Palindrome theory. J Mol Biol 84:139-144.

Harrington JJ, Van Bokkelen G, Mays RW, Gustashaw K, Willard HF. 1997. Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. Nat Genet 15:345-355.

Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, 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.

Hayden KE, Strome ED, Merrett SL, Lee HR, Rudd MK, Willard HF. 2013. Sequences associated with centromere competency in the human genome. Mol Cell Biol 33:763-772.

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.

Hori T, Shang WH, 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.

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

Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28:593-594.

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.

Jaiswal R, Choudhury M, Zaman S, Singh S, Santosh V, Bastia D, Escalante CR. 2016. Functional architecture of the Reb1-Ter complex of Schizosaccharomyces pombe. Proc Natl Acad Sci U S A 113:E2267-2276.

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, Whitington 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.

Kabeche L, Nguyen HD, Buisson R, Zou L. 2017. A mitosis-specific and R loop-driven ATR pathway promotes faithful chromosome segregation. Science.

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.

MacIsaac 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.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10.

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

Musich PR, Brown FL, Maio JJ. 1980. Highly repetitive component alpha and related alphoid DNAs in man and monkeys. Chromosoma 80:331-348.

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.

Polizzi C, Clarke L. 1991. The chromatin structure of centromeres from fission yeast: differentiation of the central core that correlates with function. J Cell Biol 112:191-201.

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.

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.

Takahashi K, Chen ES, Yanagida M. 2000. Requirement of Mis6 centromere connector for localizing a CENP-A-like protein in fission yeast. Science 288:2215-2219.

Tanaka Y, Nureki O, Kurumizaka H, Fukai S, Kawaguchi S, Ikuta M, Iwahara J, Okazaki T, 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.

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

Thayer RE, Singer MF, McCutchan TF. 1981. Sequence relationships between single repeat units of highly reiterated African Green monkey DNA. Nucleic Acids Res 9:169-181.

Villasante A, Abad JP, Mendez-Lago M. 2007. Centromeres were derived from telomeres during the evolution of the eukaryotic chromosome. Proc Natl Acad Sci U S A 104:10542-10547.

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.

Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, Shi J, Gao Z, Han F, Lee H, Xu R, et al. 2009. Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons. PLoS Genet 5:e1000743.

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.

Zhu L, Chou SH, Reid BR. 1996. A single G-to-C change causes human centromere TGGAA repeats to fold back into hairpins. Proc Natl Acad Sci U S A 93:12159-12164.

Figure Legends

Fig. 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.

Fig. 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.

Fig. 3 | Non-B-form DNA detected experimentally at dyad-depleted functional human and mouse centromeres. (**A**) Abundance of heterochromatic 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 (D5Z1) 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.

Fig. 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.

Fig. 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 classified based on CENP-A ChIP enrichment and centromere activity in artificial chromosome assays (Hayden, et al. 2013; Henikoff, et al. 2015). (**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 $p < 0.05$.

Fig. 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.

Fig. 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-form DNA structures, and great apes, which do not preferentially adopt non-B-form 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

Fig. S1 | Characterization of variation in primate centromeres. (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**) 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.

Fig. S2 | Dinucleotide frequencies of centromeric regions, neocentromeres, and matched control regions. Dinucleotide composition of centromeres, neocentromeres (human and chicken), and matched non-centromeric background regions sampled from the genomes of human, chicken, fission yeast, and various budding yeasts via *k*-mer matching.

Fig. S3 | Clade-specific patterns of predicted folding free energy. 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. The heatmap depicts -log(*p*-value) with warmer (red) colors indicating dissimilarity and colder (blue) colors indicating similarity.

Fig. S4 | Abundance of CENP-B boxes in selected primate genomes. Fraction of alphoid reads containing matches the minimal consensus CENP-B box sequence (left) or to the SELEX-defined CENP-B box motif deposited in JASPAR (right).

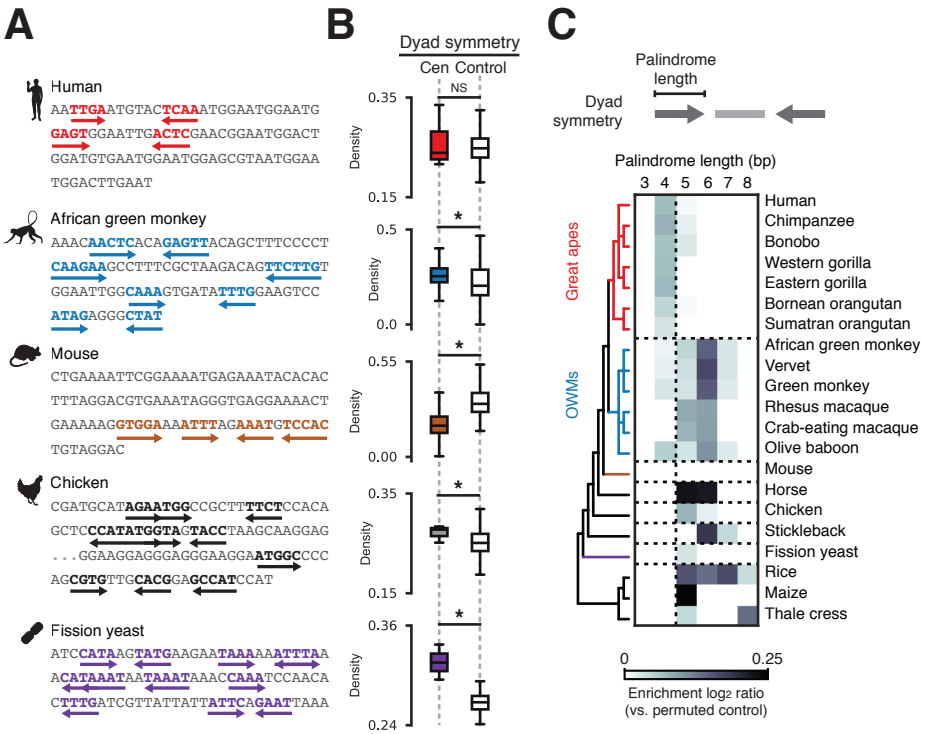
Fig. S5 | Human neocentromeres are enriched for non-B-form DNA. CENP-A ChIP-seq, GAP-seq, dyad symmetry, and SIST DNA melting and cruciform formation profiles for the human chromosome 8 and chromosome 13 neocentromeres characterized previously (Hasson, et al. 2013).

Fig. S6 | Analysis of yeast centromeric DNAs. (**A**) 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 $p < 0.05$. (**B**) Two-sample Kolmogorov-Smirnov *p*-values for all pairwise comparisons of RNAfold-predicted ensemble free energies for centromeric sequences from saccharomycetes. (**C**)

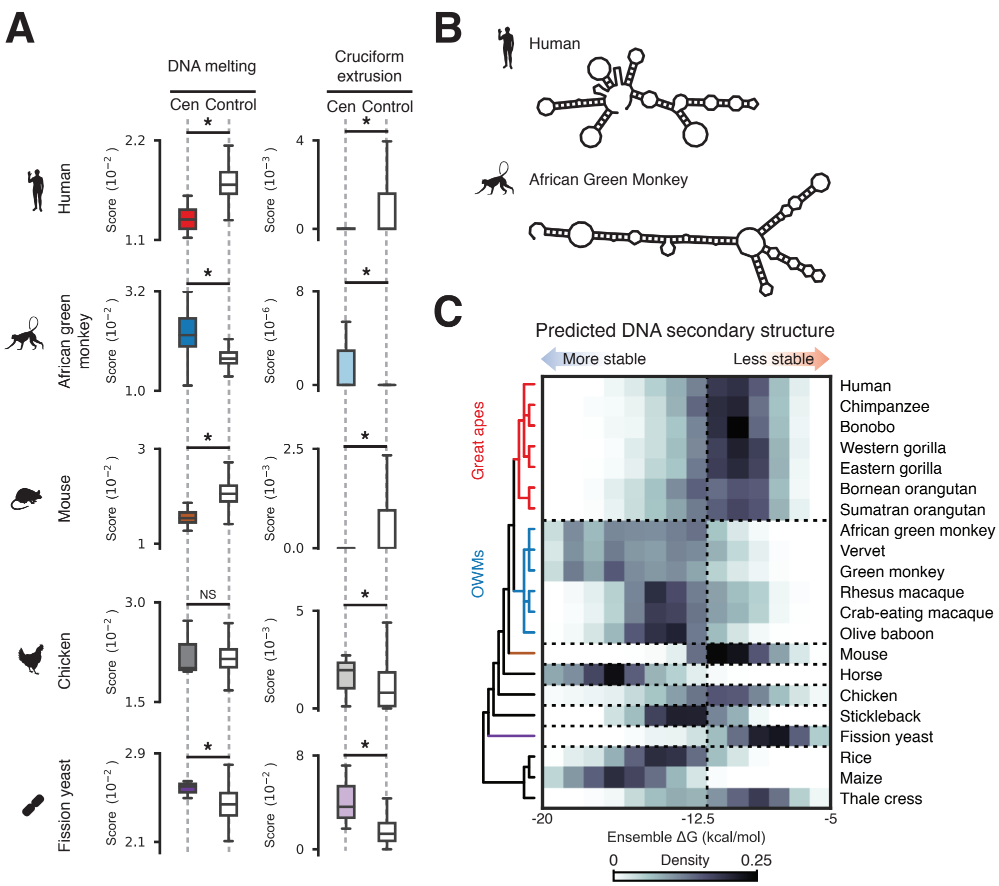
Tomtom motif similarity analysis of CDEI motifs from two *sensu lato* yeasts with the highest scoring hit from scanning a database of binding sites for ~200 yeast transcription factors (MacIsaac, et al. 2006) and comparison to the Reb1 binding site defined by ORGANIC profiling, a high-resolution mapping approach (Kasinathan, et al. 2014).

Table S1 | Publicly available datasets analyzed in this study. 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), and permanganate-seq Illumina datasets are provided. Where references are unavailable, NCBI BioProject accessions for public reference genome projects are provided.

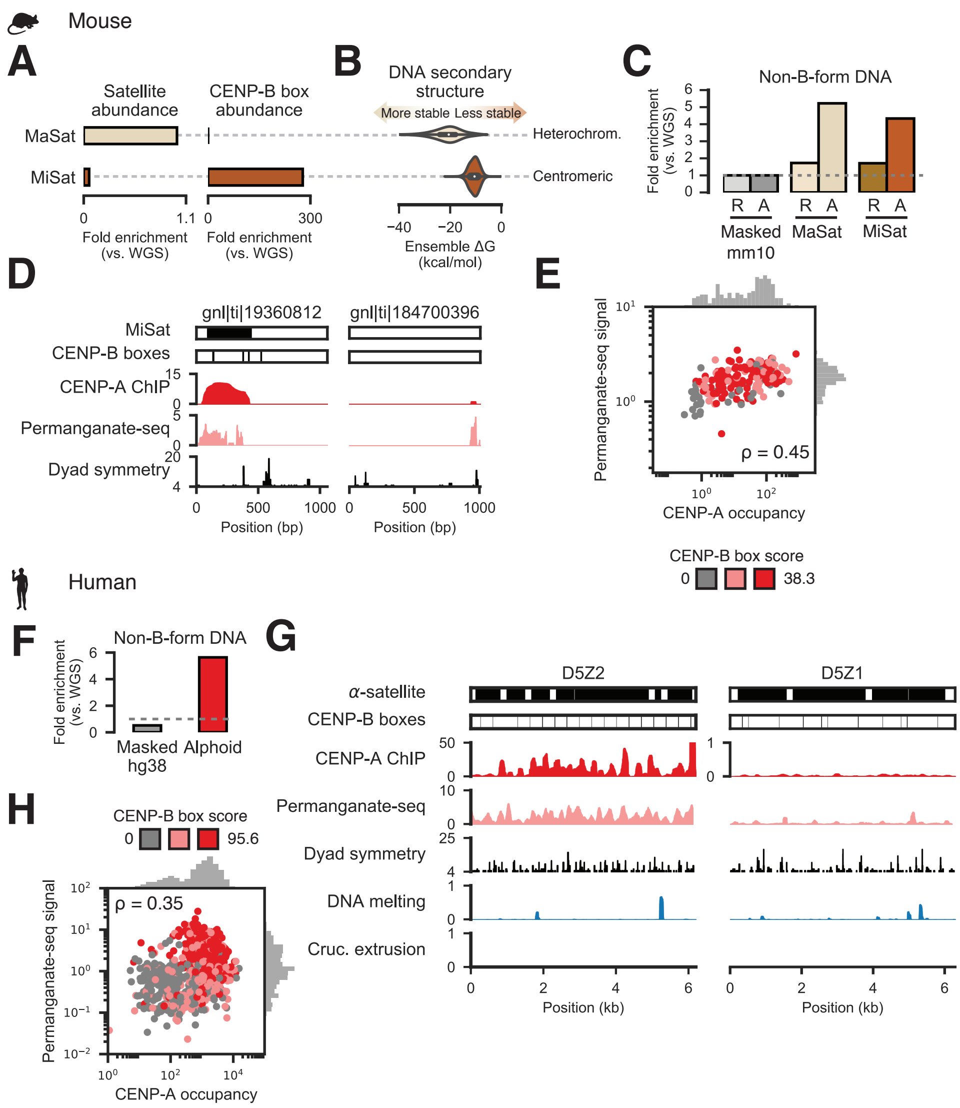
Kasinathan & Henikoff – Figure 1



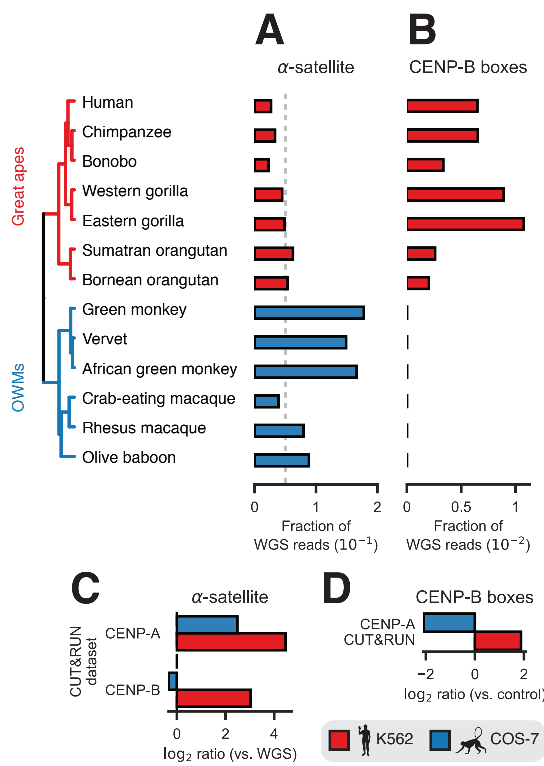
Kasinathan & Henikoff – Figure 2



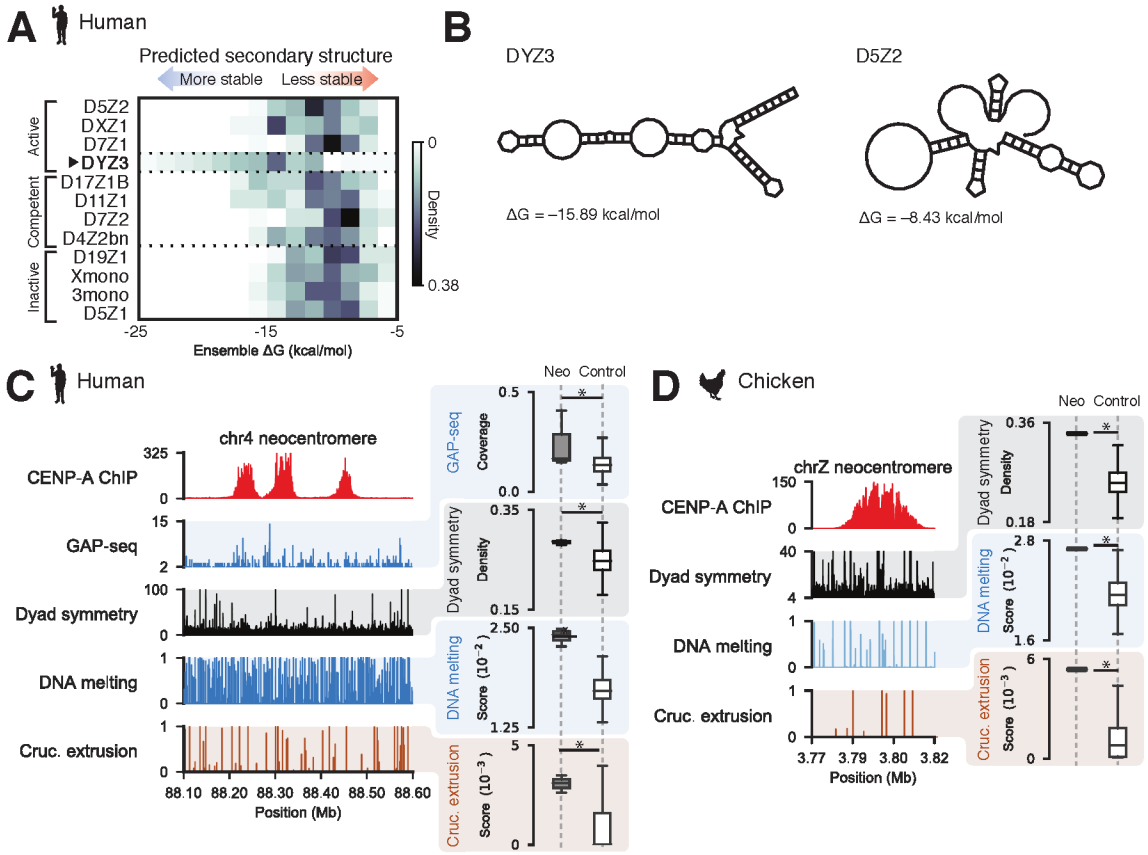
Kasinathan & Henikoff – Figure 3



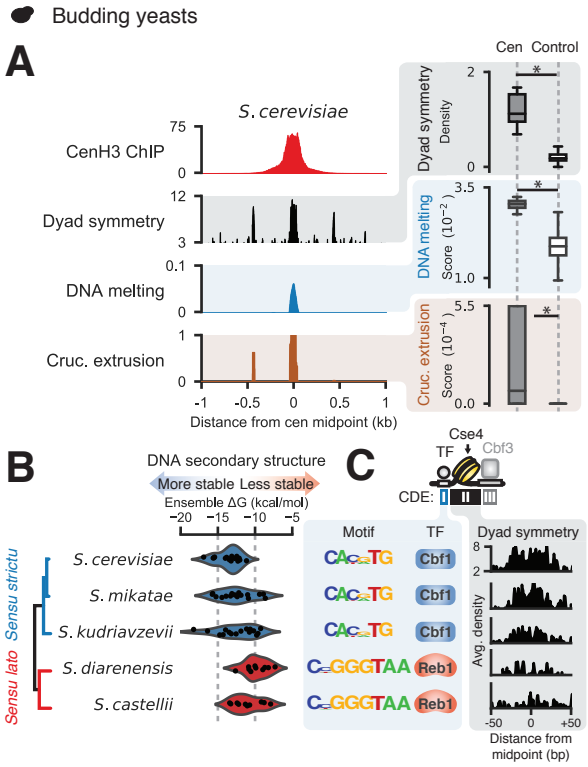
Kasinathan & Henikoff – Figure 4



Kasinathan & Henikoff – Figure 5



Kasinathan & Henikoff – Figure 6



Kasinathan & Henikoff – Figure 7