Title: The genetics and epigenetics of satellite centromeres

Authors: Paul B. Talbert and Steven Henikoff

Howard Hughes Medical Institute

Fred Hutchinson Cancer Research Center

1100 Fairview Ave N

Seattle, Washington, USA 98109

Emails: ptalbert@fredhutch.org, steveh@fredhutch.org

Abstract

Centromeres, the chromosomal loci where spindle fibers attach during cell division to segregate chromosomes, are typically found within satellite arrays in plants and animals. Satellite arrays have been difficult to analyze because they comprise megabases of tandem head-to-tail highly repeated DNA sequences. Much evidence suggests that centromeres are epigenetically defined by the location of nucleosomes containing the centromere-specific histone H3 variant cenH3, independently of the DNA sequences where they are located; however, the reason that cenH3 nucleosomes are generally found on rapidly evolving satellite arrays has remained unclear. Recently, long read sequencing technology has clarified the structures of satellite arrays and sparked rethinking of how they evolve, while new experiments and analyses have helped bring both understanding and further speculation about the role these highly repeated sequences play in centromere identification.

Keywords: Higher Order Repeats; Non-B DNA; Centromere Protein B; Break-Induced Replication; Molecular Drive

Introduction

Centromeres are the genomic loci where the proteinaceous kinetochores are assembled to attach to spindle microtubules in order to orchestrate chromosome segregation during mitosis and meiosis. In most organisms, a single centromere is found on each chromosome at a specific location. The location of centromeres is widely viewed as an epigenetic phenomenon (Karpen and Allshire 1997) based on the location of the centromere-specific variant of histone H3 (cenH3), known as CENP-A in animals (Earnshaw and Rothfield 1985) or CENH3 in plants (Zhong et al. 2002), which specifies kinetochore assembly. Yet centromeres in animals and plants are usually located in species-specific satellite arrays, those very highly repeated tandem head-to-tail arrays of non-coding sequences that typically occupy both the centromere and the flanking pericentromere of animal and plant chromosomes (Plohl et al. 2014). Satellite arrays have been called the "dark matter" of the genome (Ahmad et al. 2020) because of the difficulty of assembling blocks of sequences that are identical or nearly so, leaving large gaps in chromosome assemblies. In recent years, however, long and ultra-long sequencing reads from Pacific Biosciences (Pac-Bio) SMRT technology and Oxford Nanopore Technologies have cast illumination on previously dark matter, allowing assembly of previously intractable arrays. While short sequencing reads have defined the point centromeres of budding yeast (Fitzgerald-Hayes et al. 1982) and short regional centromeres of unicellular eukaryotes (Sanyal et al. 2004; Kanesaki et al. 2015), long reads have helped to assemble the transposon-rich centromeres of fungi (Sonnenberg et al. 2020) and satellite centromeres in maize (Wolfgruber et al. 2016; Liu et al. 2020). Though numerous challenges in assembly of satellites remain (Miga 2020), the use of long read sequencing technologies recently allowed the completion of the telomere to telomere (T2T) assembly of the human X chromosome (Miga et al. 2020) and chromosome 8 (Logsdon et al. 2021), the first human chromosomes to be completely sequenced, more than 17 years after the human genome project was declared to be complete. Such T2T assemblies allow us to see how satellite families and subfamilies are arranged, and offer insight into their evolution and functions.

The very first sequenced centromeres from budding yeast (Fitzgerald-Hayes et al. 1982) are occupied by a single CENP-A-like nucleosome (Furuyama and Biggins 2007; Henikoff and Henikoff 2012) and are generally regarded as genetic centromeres, since they contain binding sites for specific DNA-binding proteins that can self-assemble the kinetochore, including the kinetochore-specifying cenH3. However, the view that other centromeres are predominantly epigenetic is supported by much evidence, notably the occurrence of human neocentromeres, in which CENP-A nucleosomes are found on sequences that

lack satellite arrays, that were not previously centromeres, and that can nevertheless function as centromeres that can be inherited across generations (Amor et al. 2004). In addition, centromeric sequences are known to evolve rapidly and differ dramatically between sibling species (Henikoff et al. 2001), suggesting that sequence conservation does not exist for this essential function in every cell cycle. More recently, insect holocentromeres, centromeres that occupy large chromosomal regions instead of a specific locus, have been found to lack CENP-A (Drinnenberg et al. 2014) and to occupy large domains of inactive chromatin covering half of the genome (Senaratne et al. 2021). These domains can be lost or gained in response to nearby gene activation or silencing. These observations and others have been interpreted to mean that DNA sequence does not matter for most centromeres. Yet this leaves unexplained why the vast majority of natural animal and plant centromeres occupy large satellite arrays, and why satellite centromeres seem to be restricted to animals and plants and are not found in fungi or other eukaryotes (Talbert and Henikoff 2020).

Why satellite arrays?

A potential explanation for the existence of satellite arrays was proposed in the Unequal Exchange model (Smith 1976). In this model, once a tandem duplication is established through periodicities generated by random mutation followed by unequal exchange between sister chromatids that does not require extensive homology, the resulting duplication can undergo unequal out-of-register exchange with its copy on the sister chromatid (or homolog), generating further reciprocal duplications and deletions. As mutation alters the sequence of individual monomers, they can become encompassed within Higher Order Repeats (HORs), in which sets of distinct monomers are duplicated together to form larger repeats. With an exchange rate high enough, homogeneity can be maintained in the face of mutation. This model is neutral, in that there is no preference for preserving duplications rather than deletions, and if an array is deleted down to one monomer the process is extinguished, suggesting a need for some mutational or selective force to maintain or expand the array in order to generate the natural arrays of megabases of repeats.

Dover (Dover 1982) emphasized the importance of gene conversion in homogenizing families of repetitive sequences, particularly when they are physically close, as in tandem arrays. Dover viewed gene conversion, unequal exchange, and transposition as processes that turn over DNA and can be stochastic or directional, which he termed molecular drive. He proposed that the accumulation of such

changes and homogenization within populations could lead to “accidental speciation” due to incompatibility between separate populations, offering a rationale for the long-held suspicion that satellite arrays have a role in speciation (Yunis and Yasmineh 1971; Ferree and Prasad 2012).

A Darwinian process that favors expansion of centromeric arrays was provided in the centromere drive model (Henikoff et al. 2001). In this model, the observed rapid evolution of both centromeres and kinetochore proteins such as CENP-A and CENP-C (another foundational kinetochore protein) was proposed to result from a genetic conflict between satellite DNA variants acting selfishly to favor their own transmission through female meiosis and kinetochore proteins evolving to suppress this biased transmission. Because female animals and plants have asymmetric meiosis in which only one of the four meiotic products is transmitted to the next generation, centromere variants will compete for inclusion in the egg or megaspore, and variants that can attract more kinetochore proteins will have ‘stronger’ centromeres that can favor their orientation at the first meiotic division so that they end up in the egg rather than in a polar body. Such lack of parity between centromeres may cause problems from unequal tension in male meiosis, where all four meiotic products contribute to fertility, and so it is hypothesized that kinetochore proteins evolve to suppress centromere drive by restoring parity between chromosomes. The ensuing rapid divergence of centromeres and kinetochore proteins was proposed as a possible mode of generating incompatibilities that result in speciation (Henikoff et al. 2001). In contrast to a purely epigenetic view of centromere specification, this model implies that variant satellite arrays differ genetically in their ability to recruit kinetochore proteins. Centromere drive therefore can be viewed as favoring genetic control by a satellite variant over the assembly of kinetochore proteins, especially cenH3, while suppression can be viewed as a disruption of variant-specific interactions to make kinetochore assembly insensitive to driving genetic variants and restore a more epigenetic or DNA-sequence-independent mode of kinetochore assembly (Dawe and Henikoff 2006). Strong supporting evidence for centromere drive has been found in monkeyflowers (*Mimulus sp*.), where the large satellite duplication D can be transmitted to 90% of offspring through female meiosis but male meiosis follows Mendel’s rules (Fishman and Willis 2005; Finseth et al. 2015), and in mice, where chromosomes with more centromeric repeats load more CENP-A and are preferentially segregated to the egg. (Chmatal et al. 2014; Iwata-Otsubo et al. 2017).

The centromere drive model predicts that variant centromeres that acquire more kinetochore proteins will be favored in female meiosis in plants and animals, but it does not tell us what features of satellites in particular are favored. Although satellites come in a range of lengths, sizes approximating the length

of one or two nucleosomes predominate (Melters et al. 2013). Satellites impose translational and rotational phasing on nucleosomes (Hasson et al. 2013; Zhang et al. 2013; Henikoff et al. 2015), generating regular nucleosome arrays and potentially a regular kinetochore structure. Many satellites have a 10-bp periodicity of WW (W = A or T) dinucleotides (Talbert and Henikoff 2020), which favors rotational phasing and minimizes the bending energy of wrapping DNA around nucleosomes (Prytkova et al. 2011; Struhl and Segal 2013), making nucleosomes more stable. This greater stability may be important to form a strong kinetochore under the tension exerted by microtubules during anaphase I, so that selection would favor the expansion of structurally suitable sequences.

Insights from long sequencing reads

Can the fully assembled centromere sequences available from long read sequencing technologies tell us more about how satellites evolve or why they are favored in evolution? One of the first satellite centromeres to be assembled to near completion using Pac-Bio long reads was the 1.85 Mb *centromere 10* (*CEN 10*) from maize (Wolfgruber et al. 2016). This study uncovered evidence of frequent recombination events mediated by microhomology, including presumed intrastrand events such as a hemicentric inversion that split the original array of *CentC* (the maize centromeric satellite), internal deletions in CRM centromeric retrotransposons, recombination between nearby retroelements of different subtypes mediated by 5 bp of identity, a segmental duplication, insertion of mitochondrial sequences, and a HOR recently formed by adjacent duplication. The authors argued that these events are better explained by microhomology-mediated end-joining, a mode of error-prone double-strand break (DSB) repair, than by unequal exchange or gene conversion.

More recently, long-read sequencing has allowed assembly of seven maize centromeres, including T2T assemblies of chromosomes 3 and 9 (Liu et al. 2020). Three centromeres lack *CentC* entirely, being composed of the CRM transposons that target centromeres and other transposons. In these maize centromeres there does not appear to be any preference by CENH3 for *CentC* versus other sequences. This lack of correlation may be explained because inbreeding and selection for centromere-linked genes during domestication greatly reduced *CentC* and the number of surviving haplotypes, while simultaneously selecting for the fixation of at least 57 distinct neocentromeres (Schneider et al. 2016).

Human centromeres are made up of α-satellite, with monomers of ~171 bp. Most monomers fall into two types, A and B (Alexandrov et al. 2001). A monomers have a 19-bp motif called an n box (Rice

2020b) that overlaps the binding site for a protein of unknown function pJα (Gaff et al. 1994), and B monomers have in the corresponding location a 17-bp binding site called the CENPB-box or simply the b-box which is bound by CENP-B, the only known sequence-specific human kinetochore protein (Masumoto et al. 1989). A and B monomers usually occur in alternation (Alexandrov et al. 2001). A and B monomers are arranged in HORs such that while individual pairs of monomers within a HOR may be only 50-70% identical, copies of a particular multimeric HOR are usually nearly identical. The edges of satellite arrays have disordered monomeric satellites (Schueler et al. 2001; Miga et al. 2020; Logsdon et al. 2021), while the middle of arrays comprise HORs, of which the simplest is a dimer of A and B monomers. In an analysis of PacBio reads and consistent with earlier results, b-boxes were most frequently found in every other monomer, *i.e.* as part of n/b dimers, and only rarely were found in adjacent monomers (Rice 2020b). The reason that b-boxes seem to be disfavored in adjacent monomers is not clear. The simple n/b dimer structure forms the basis of other HORs, which may have 2, 3, or 4 n/b dimers, or in longer HORs the dimer structure may be interrupted by additional monomers. In some dimers, especially in longer HORs, a b-box may be mutated so that it no longer binds CENP-B, and there may be non-canonical monomers. A notable exception to this dimeric structure is the Y chromosome, which has A monomers with n-boxes but lacks B monomers and b-boxes, and has the longest human HOR at 34 monomers (Jain et al. 2018). Some chromosomes have more than one HOR, though usually only one can form the centromere (McNulty and Sullivan 2018). n/b dimers can be further subdivided into families by their degree of sequence similarity, with Suprachromosomal Family 1 (SF1) on chromosomes 1, 3, 5, 6, 7, 10, 12, 16, and 19, and SF2 on chromosomes 2, 4, 8, 9, 13, 14, 15, 18, 20, 21, and 22, while other families are found on chromosomes 11, 17, and X, and as additional HORs on other chromosomes (Alexandrov et al. 2001; Henikoff et al. 2015).

The Break-Induced Replication (BIR) model

While unequal exchange or gene conversion may contribute to homogenizing repeat sequences in satellite arrays, the higher order structures of repeats have been more difficult to explain. In analyzing long reads from human HORs enriched in CENP-A, Rice (Rice 2020b; Rice 2020c) concluded that the complex nested structures of HORs could be created by break-induced replication (BIR). In this model, the Constitutive Centromere Associated Network (CCAN), the persistent core of the kinetochore present throughout the cell cycle in most animals, presents a barrier to replication, as it does in yeasts (Greenfeder and Newlon 1992; Mitra et al. 2014) which results in fork pausing and collapse, creating a

one-ended DSB. Indeed, human α-satellites are enriched for aphidocolin-sensitive DSBs, indicative of replication stalling (Crosetto et al. 2013). Re-starting collapsed forks is carried out by the BIR pathway (Sakofsky and Malkova 2017). Resection of the one-ended break allows the free 3' strand of the truncated sister chromatid to re-initiate replication on its sister chromatid by BIR or microhomology-mediated BIR (Kockler et al. 2021). Initiation will frequently be out-of-register in a tandem array, with initiation behind the fork leading to expansion of the array and initiation ahead of the fork leading to deletion (Figure 1). In yeast rDNA arrays, expansion by BIR is favored over deletion (Kobayashi 2014). Newly replicated chromatin is hyperacetylated and depleted for histone H1 relative to bulk chromatin (Perry and Annunziato 1989), while parental chromatin in front of replication forks is positively supercoiled (overtwisted). The relatively increased accessibility behind the replication fork would facilitate strand invasion and favor repeat expansion.

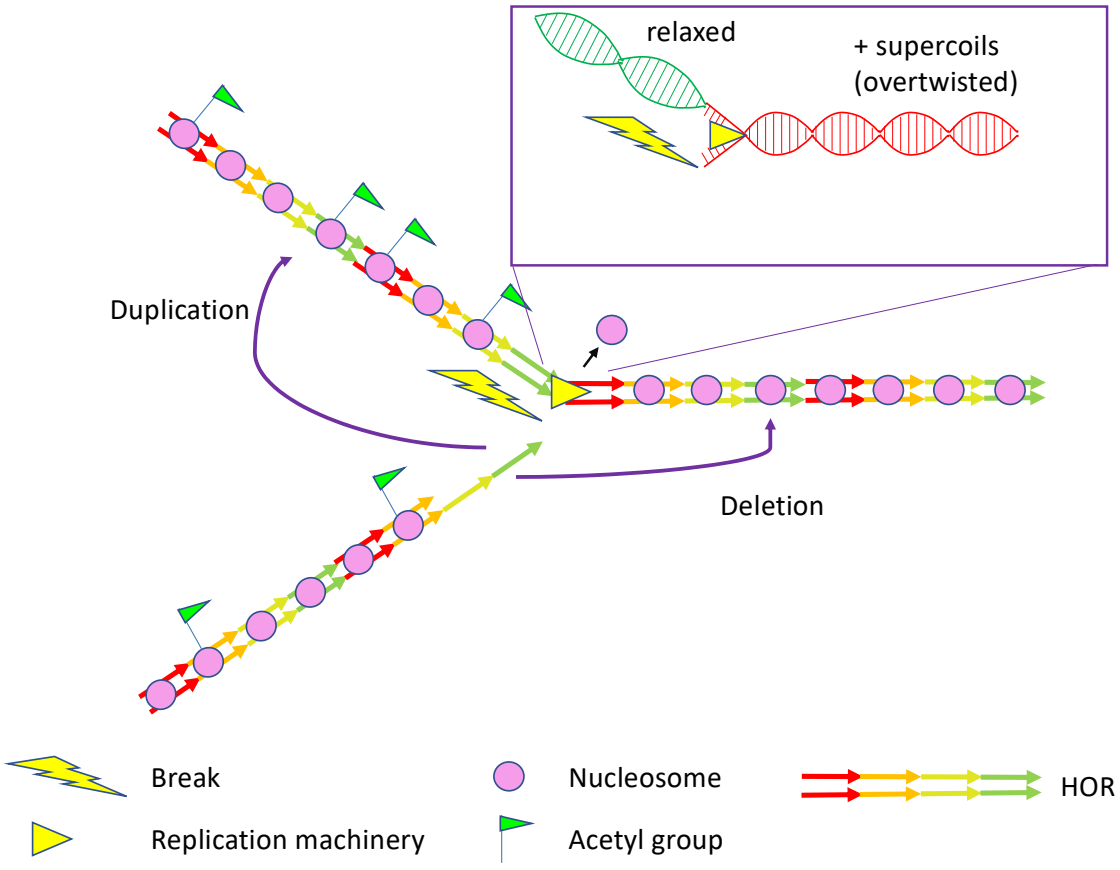

Figure 1. Amplification of Higher Order Repeats (HORs) through Break-Induced Replication (BIR). Replication fork stalling can lead to one-ended double-strand breaks (DSBs). Resection yields a free single-strand 3' end that can invade a homologous sequence and re-initiate replication. Re-initiating at an out-of-register repeated sequence ahead of the fork will lead to deletion while re-initiating at one behind

the fork will lead to duplication. Duplication appears to be favored, perhaps because the chromatin behind the fork is more accessible to strand invasion owing to the new acetylated histones and/or the relaxed torsional state behind the overtwisted DNA ahead of the fork (inset).

In the BIR model (Rice 2020c), HORs are hypothesized to go through a 'life cycle' starting with n/b-box dimers, which are favored by centromere drive because CENP-B enhances recruitment of CENP-C, making a stronger centromere and increasing the fidelity of centromere function (Fachinetti et al. 2015). Centromere drive acting on b-boxes has also been proposed as the explanation for why the Y chromosome lacks b-boxes, since it never experiences centromere drive in female meiosis (Marshall and Choo 2012). From a n/b-box dimer, additional dimers can be added to make longer HORs, which are favored because they can expand laterally more quickly and occupy the central core of the satellite array more easily, pushing out older HORs to the sides of the kinetochore, where they decay over time because they are no longer subject to frequent BIR (Rice 2020c). However as HORs increase in length they are also more likely to acquire b-box mutations, additional n-box monomers, or other divergences that make them susceptible to replacement by a young n/b-box dimer HOR, perhaps inserted from a different chromosome by template switching. Besides potentially accounting for the expansion of highly identical HORs, BIR is also mutagenic, with elevated levels of frameshifts and base substitutions that are 500-fold or more greater than in normal S-phase replication (Sakofsky and Malkova 2017). Error-prone BIR therefore may account for the rapid divergence of centromeric HORs at the nucleotide level, which is estimated to be greater than 10 times the divergence on chromosome arms between humans and chimps (Rice 2020a). This is consistent with the view that elevated mutation rates at the point or short regional centromeres of yeasts (Padmanabhan et al. 2008; Bensasson 2011) may be the result of fork stalling (Greenfeder and Newlon 1992; Mitra et al. 2014) followed by BIR repair.

Human centromeres

The BIR model is supported by the recently completed T2T assembly of human chromosome 8 (Logsdon et al. 2021). The authors of this study compared the 2 Mb centromeric alpha satellite array with Centromere 8 assemblies from chimpanzees, orangutans and rhesus macaques and found that each of these primate centromeres showed a largely symmetrical satellite array with four or five layers of evolutionary structure, with each layer similar on the p and q arms (Figure 2). The α-satellite monomers in the flanking pericentromeres of humans and chimps (layer 1) fall into two clades, one of which is

present only in the q arm and which has common ancestors with monomers and dimers from macaque, indicating an ancient stratum of the α array. The second human layer is a short (~60 kb) transitional region between monomers and HORs. The third human layer is composed largely of an 11 monomer HOR. The large fourth layer has the greatest variety of HOR subtypes including HORs of 4, 7, and 8 monomers intermixed with the 11 monomer HOR from which they are derived. The fifth layer is a 177-kb region entirely composed of nearly identical 7 monomer HORs. The HORs of great apes all have a common origin distinct from α monomers, and chimpanzees and gorillas resemble humans in having similar transitional layers from monomers to HORs with different arrangements of blocks of HORs in subsequent layers, while macaques have a large central block of highly uniform dimers flanked by more divergent dimers. An elevated mutational divergence was found between centromeres, 2-4-fold higher than at random loci, consistent with an error-prone repair process such as BIR. The authors proposed a model in which highly identical repeats expand, pushing older repeats out of the centromere. They hypothesized that the more divergent clade of monomers shared between macaques and the q arm of apes represents the remnants of the ancestral centromere.

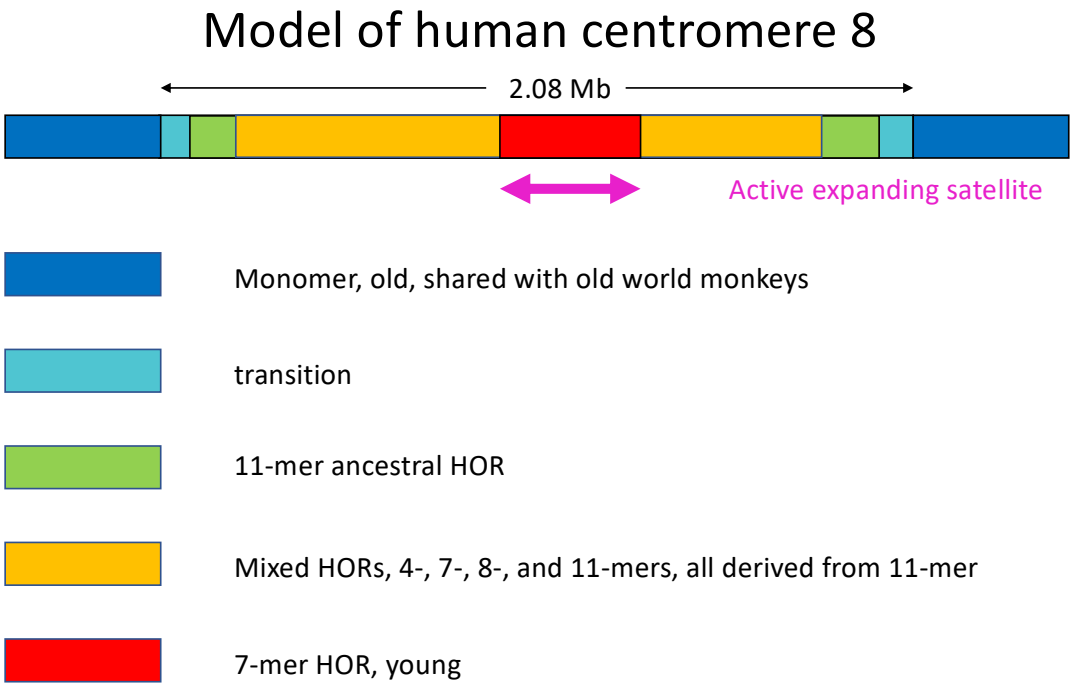

Figure 2. Human centromere 8. Human centromere 8 shows successive evolutionary layers, with the youngest, active, layer in the middle and the oldest monomer layer on the edges of the array.

An intriguing feature of the centromeres of both the X chromosome and chromosome 8 is a ~60-70 kb hypomethylated region, which on chromosome 8 is in the middle of the 632-kb region occupied by CENP-A. Hypomethylation of satellite repeats occupied by CENH3 compared with the same repeats in flanking heterochromatin was also reported in *Arabidopsis*, maize, and *cen11* of rice, though other rice centromeres that have more transposons and less of the satellite *CentO* showed elevated methylation instead (Zhang et al. 2008; Yan et al. 2010). In contrast, the HORs of human X and 8 centromeres are essentially devoid of transposons yet are mostly methylated, so the significance of hypomethylation in centromeres remains unclear.

Replication of α-satellite

While the BIR model is supported by the high mutation rate in centromeres and the structure of HORs and their evolutionary layers in human centromere 8, an apparent conflict exists with the assumption of the model that the CCAN causes replication stalling and breakage in satellite centromeres. Contrary to expectation, depletion of CENP- A greatly increases fork-stalling in human centromeres with increased unequal exchange and formation of R-loops, likely caused by replication-transcription conflicts, followed by unfinished replication and anaphase bridges or by breakage and translocations at centromeres (Giunta and Funabiki 2017; Giunta et al. 2021). While this does not preclude a role for the CCAN in causing fork stalling and BIR, it indicates that satellite centromeres face additional more serious causes of fork stalling in repeated sequences when CENP-A and the CCAN are reduced. Mismatch repair proteins that bind to 4-stranded Holliday junctions and their single-stranded progenitor structures such as DNA hairpins (Snowden et al. 2004) are enriched in replicating α-satellite, suggesting that DNA secondary structures form in single-stranded repetitive DNA behind the replication fork (Aze et al. 2016), with the potential to contribute to fork stalling if they interfere with DNA polymerization. Positively supercoiled DNA and chromatin loops are also enriched in replicating α-satellite, dependent on topoisomerase I, which acts together with condensins to introduce positive supercoils into DNA (Hirano 2012). Positive supercoiling suppresses the accumulation of the single-strand binding protein replication protein A (RPA), which can activate Ataxia telangiectasia and Rad3 related (ATR)-dependent DNA-damage-checkpoint signaling, thereby giving time for secondary structures to be resolved and facilitating replication through α-satellite (Aze et al. 2016). Centromeres are enriched during interphase in the condensin II complex, which is necessary for proper CENP-A loading and retention (Bernad et al. 2011) and is mutually interdependent for centromeric localization with HJURP (Holliday junction recognition protein), the chaperone that assembles CENP-A into centromeres during G1 (Barnhart-Dailey et al. 2017) and that is necessary to retain CENP-A through replication (Zasadzińska et al. 2018). HJURP can interact with the mismatch repair protein MSH5 and can bind to Holliday junctions in vitro (Kato et al. 2007), suggesting a possible role for the secondary structures that form on replicating α-satellite in directing or supporting HJURP's role in retaining CENP-A through replication (Figure 3).

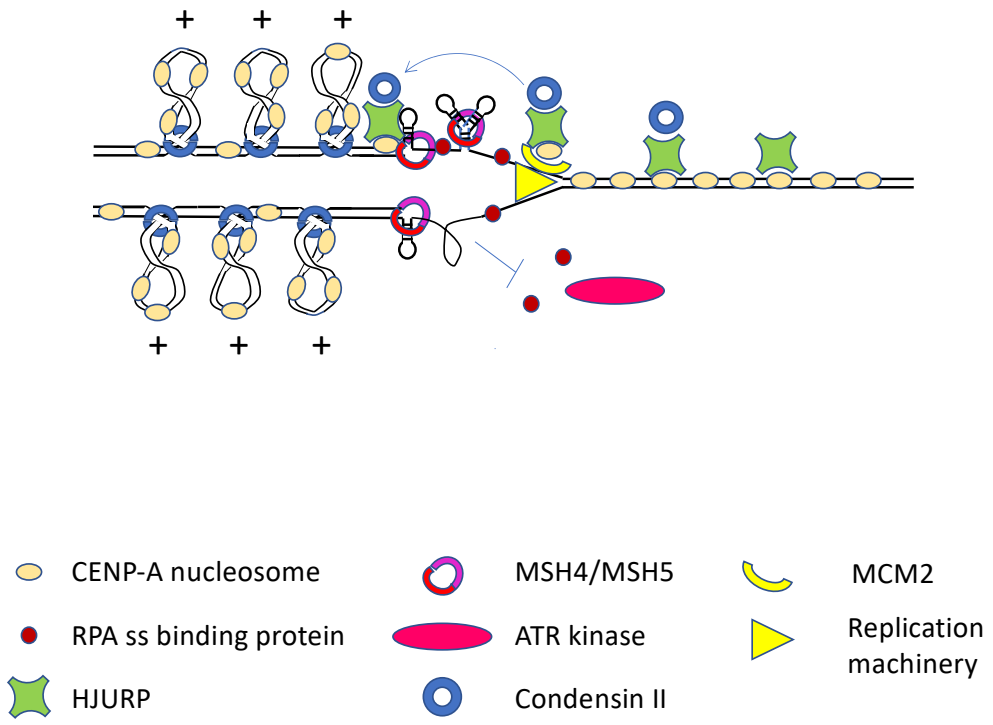

Figure 3. Model of replication through a-satellite. Holliday junction recognition protein (HJURP) associates with CENP-A nucleosomes prior to S-phase and recruits the condensin II complex. At the replication fork, HJURP and the MCM2 subunit of the replication machinery work together to assure that CENP-A nucleosomes re-assemble behind the fork. DNA secondary structures form on single-stranded repetitive DNA behind the fork, and HJURP and mismatch repair proteins (MSH4 and MSH5 are shown) bind to them and resolve them. Condensin II complexes extrude positively supercoiled DNA loops, and the positive torsion inhibits the binding of replication protein A (RPA), which binds single-strand DNA and must accumulate in order for the Ataxia telangiectasia and Rad3 related (ATR) kinase to signal that DNA damage has occurred and arrest replication, thus giving time for secondary structures to be resolved. Condensin II is also needed with HJURP to assemble new CENP-A nucleosomes in G1, and condensin-mediated loops may play a role in the organization of the kinetochore. Modified from (Kato et al. 2007; Bernad et al. 2011; Aze et al. 2016; Barnhart-Dailey et al. 2017; Zasadzińska et al. 2018).

What does CENP-B do?

The BIR model proposes that n/b-box dimers were acquired through centromere drive and are the foundation from which other HORs are built, presumably because CENP-B strengthens the kinetochore. What exactly does CENP-B do? CENP-B is a protein that is a domesticated transposase that has lost

transposase activity (Smit and Riggs 1996; Kipling and Warburton 1997). It is conserved throughout mammals, but b-boxes are present in the centromeric satellite arrays of only some mammalian clades, such as great apes, mice, horses, but not in old world monkeys, rabbits, carnivores, and others. The function of CENP-B in clades that lack b-boxes is unknown, while the nine bases required for CENP-B binding appear to have evolved independently in each lineage that has b-boxes, leading to the observation that CENP-B appears to have evolved to stabilize kinetochore function in pre-existing satellite centromeres (Gamba and Fachinetti 2020).

In vitro CENP-B binds the b-box in the major grooves and is able to kink the DNA with a bend of 59° (Tanaka et al. 2001). It forms antiparallel homodimers which can bind two b-boxes at once and can form loops between b-boxes on the same DNA molecule (Yoda et al. 1998). In cells, the acidic domain of CENP-B has seemingly conflicting functions promoting both kinetochore formation and heterochromatin formation through different interacting partners (Otake et al. 2020). CENP-B binds to both the CENP-A N-terminal tail and to CENP-C and is necessary to maintain proper levels of CENP-C (Fachinetti et al. 2015). Neocentromeres and the Y chromosome centromere, which both lack b-boxes, have reduced levels of CENP-C and have increased levels of chromosome mis-segregation compared to other centromeres, consistent with the view that CENP-B makes stronger centromeres that are favored by centromere drive.

In the prevailing templating model of CENP-A localization and maintenance, CENP-A recruits CENP-C, which recruits the M18BP1 licensing complex and the CENP-A chaperone HJURP to load new CENP-A next to its pre-existing locations in a self-dependent loop (reviewed in (McKinley and Cheeseman 2016)). Using an auxin-inducible degron system that destroys existing CENP-A, the Fachinetti group found that new CENP-A localized back to the same HORs in native centromeres, dependent on the ability of DNA-bound CENP-B to bind to CENP-C, on CENP-C recruitment of the M18BP1 licensing complex and HJURP, and on loading of new CENP-A by HJURP (Hoffmann et al. 2020). Thus, de novo CENP-A deposition did not depend on pre-existing CENP-A at centromeres. Using a lacO system to tether CENP-B to an ectopic site, the authors showed that CENP-B could recruit CENP-C and CENP-A, but CENP-A recruitment was dependent on CENP-C and could not be recruited directly by CENP-B. Although nearly 100% of cells recruited new CENP-A to native centromeres in the presence of CENP-B, in the absence of CENP-B about 40% of centromeres were still able to partially load de novo CENP-A, and de novo CENP-A was loaded onto ~25% of Y chromosomes, suggesting that α-satellite has some ability to recruit CENP-A even without CENP-B, but that pre-existing CENP-A probably also contributes to maintaining CENP-A at the Y

centromere via the M18BP1 licensing complex. These results are consistent with the long-held observation that human artificial chromosomes with functioning centromeres can be made from α-satellite HORs that contain b-boxes (Ohzeki et al. 2002; Ohzeki et al. 2020). These observations indicate that human centromeres are genetic in the same sense as budding yeast centromeres in that a CENP-B is able to bind the centromere and assemble a kinetochore, analogous to the sequence-specific DNA-binding proteins of yeast.

Do other sequences besides the b-box matter?

Prominent phasing of CENP-A nucleosomes occurs on HORs of both the X and Y chromosomes, though phasing is more precise on the X, suggesting that b-boxes are unnecessary for phasing but contribute to its precision (Hasson et al. 2013), probably by direct contact between bound CENP-B and the N-terminus of CENP-A. Mapping CENP-A ChIP-seq reads onto PacBio reads, long arrays of Centromere 1-like dimers (SF1) and Centromere 13-like dimers (SF2) were found to comprise most active centromeres and to precisely position CENP-A and CENP-C on each monomer in the dimer, with a b-box between them (Henikoff et al. 2015). CENP-A and CENP-C occupancy diminished with as little as 2-10% divergence from the consensus sequence of Cen1-like and Cen13-like dimers. In a follow-up study, high salt extraction released intact particles containing CENP-A/B/C that probably represent the intact CCAN (Thakur and Henikoff 2018). Enrichment of these particles correlated with the density of b-boxes in different HORs, though lower enrichment of CENP-A-containing particles was also found on sequences with few or no b-boxes, such as the D7Z2 HOR of chromosome 7. Mapping of fragments onto SF1 dimer arrays revealed a 50-fold difference in occupancy of different dimers, and a diversity of structures. For example, mapping to four adjacent dimers of D7Z1 that are 88-96% identical, particles were found on both monomers or only one monomer of a dimer. In the latter case, the particles could overlap the b-box either from the left or right. These observations suggest that very similar sequences can dramatically affect occupancy by the CCAN, which appears to be flexible in conformation.

Non-B form DNA in centromeric satellites

HORs and b-boxes characterize satellite arrays in great apes, but in other organisms both satellite and non-satellite centromeres are enriched in dyad symmetries that are predicted to form non-B form DNA structures such as cruciforms or hairpins (Koch 2000; Kasinathan and Henikoff 2018). Short (<10 bp)

dyad symmetries that are predicted to extrude cruciform structures are common features in the α-satellite of old world monkeys, in the human Y centromere, in human and chicken neocentromeres, and in the centromeres of horses, chickens, plants, and fission yeast (Kasinathan and Henikoff 2018). In contrast, the b-box-containing α-satellite of great apes and mouse centromeric satellite are predicted to have a low propensity to form cruciforms, but genome-wide mapping using permanganate treatment in the human and mouse genomes (Kouzine et al. 2013; Kouzine et al. 2017) nevertheless revealed non-B form DNA in these centromeres that correlated with CENP-A enrichment (Kasinathan and Henikoff 2018). CENP-B can bend DNA by 59° (Tanaka et al. 2001), and this may enhance cruciform formation by b-box-containing repeats. Such secondary structure may be a defining feature directing CENP-A deposition. The CENP-A chaperone HJURP was originally identified as a protein that could bind four-way DNA junctions in vitro (Kato et al. 2007), and it is possible that it recognizes cruciform structures in centromeres and deposits CENP-A. It is unknown whether its distant fungal homolog Scm3 also binds four-way junctions, but Scm3 homologs in various fungi contain AT hooks, myb domains, and zinc fingers (Aravind et al. 2007) that might impose or stabilize cruciform structures on transposons or other centromeric sequences. In this way, either spontaneous or induced cruciforms would constitute sequence-encoded features targeted by CENP-A chaperones. These structural features could be the raw material on which centromere drive acts.

Perspective

Long-read sequencing has made it possible to know the complete structures of satellite centromeres, and while only a few are known so far, the structures have brought into question the long accepted but seldom carefully examined unequal exchange model for their evolution. Evidence for microhomology-based repair mechanisms has been invoked from both maize and human centromeres and further evaluations of repair and recombination mechanisms in satellites are warranted, as well as better understanding of the elevated mutational rates in centromeres of all types. With tools such as degron and tethering systems, the genetic properties of human centromeres and the role of b-boxes have been clarified, and these and other tools promise further progress in understanding the interactions between centromeres, kinetochores, chaperones, replication, and transcription in mitosis and meiosis.

The development of tools to better predict and map non-B DNA structures and supercoiling in centromeres could possibly change the way we think about centromere specification. The ability to form

non-B DNA from a variety of sequences, including both native centromeres and sequences that become neocentromeres, could unite the genetic and epigenetic views of centromeres. Non-B DNA provides a large sequence space from which centromeric DNA can be selected, and may provide a rationale for why centromeres are usually formed on AT-rich DNA (Talbert and Henikoff 2020), which melts more easily and could aid in forming transient cruciforms or other secondary structures. Such structures might contribute to the fork-stalling, breakage, error-prone repair, expansions, and rearrangements that occur at centromeres, the processes that make centromeres the most evolutionarily dynamic structures in the genome.

Acknowledgments

We thank Kami Ahmad for helpful comments on the manuscript and Howard Hughes Medical Institute for funding.

References

Ahmad SF, Singchat W, Jehangir M, Suntronpong A, Panthum T, Malaivijitnond S, Srikulnath K. 2020. Dark Matter of Primate Genomes: Satellite DNA Repeats and Their Evolutionary Dynamics. *Cells* **9**.

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

Amor DJ, Bentley K, Ryan J, Perry J, Wong L, Slater H, Choo KH. 2004. Human centromere repositioning "in progress". *Proc Natl Acad Sci U S A* **101**: 6542-6547.

Aravind L, Iyer LM, Wu C. 2007. Domain architectures of the Scm3p protein provide insights into centromere function and evolution. *Cell Cycle* **6**: 2511-2515.

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.

Barnhart-Dailey MC, Trivedi P, Stukenberg PT, Foltz DR. 2017. HJURP interaction with the condensin II complex during G1 promotes CENP-A deposition. *Mol Biol Cell* **28**: 54-64.

Bensasson D. 2011. Evidence for a high mutation rate at rapidly evolving yeast centromeres. *BMC Evol Biol* **11**: 211.

Bernad R, Sanchez P, Rivera T, Rodriguez-Corsino M, Boyarchuk E, Vassias I, Ray-Gallet D, Arnaoutov A, Dasso M, Almouzni G et al. 2011. Xenopus HJURP and condensin II are required for CENP-A assembly. *J Cell Biol* **192**: 569-582.

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.

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.

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

Dover G. 1982. Molecular drive: a cohesive mode of species evolution. *Nature* **299**: 111-117.

Drinnenberg IA, deYoung D, Henikoff S, Malik HS. 2014. Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. *Elife* **3**.

Earnshaw WC, Rothfield N. 1985. Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma* **91**: 313-321.

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.

Ferree PM, Prasad S. 2012. How can satellite DNA divergence cause reproductive isolation? Let us count the chromosomal ways. *Genet Res Int* **2012**: 430136.

Finseth FR, Dong Y, Saunders A, Fishman L. 2015. Duplication and Adaptive Evolution of a Key Centromeric Protein in Mimulus, a Genus with Female Meiotic Drive. *Mol Biol Evol* **32**: 2694-2706.

Fishman L, Willis JH. 2005. A novel meiotic drive locus almost completely distorts segregation in mimulus (monkeyflower) hybrids. *Genetics* **169**: 347-353.

Fitzgerald-Hayes M, Clarke L, Carbon J. 1982. Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. *Cell* **29**: 235-244.

Furuyama S, Biggins S. 2007. Centromere identity is specified by a single centromeric nucleosome in budding yeast. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 14706-14711.

Gaff C, du Sart D, Kalitsis P, Iannello R, Nagy A, Choo KH. 1994. A novel nuclear protein binds centromeric alpha satellite DNA. *Hum Mol Genet* **3**: 711-716.

Gamba R, Fachinetti D. 2020. From evolution to function: Two sides of the same CENP-B coin? *Exp Cell Res* **390**: 111959.

Giunta S, Funabiki H. 2017. Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. *Proc Natl Acad Sci U S A* **114**: 1928-1933.

Giunta S, Hervé S, White RR, Wilhelm T, Dumont M, Scelfo A, Gamba R, Wong CK, Rancati G, Smogorzewska A et al. 2021. CENP-A chromatin prevents replication stress at centromeres to avoid structural aneuploidy. *Proc Natl Acad Sci U S A* **118**.

Greenfeder SA, Newlon CS. 1992. Replication forks pause at yeast centromeres. *Mol Cell Biol* **12**: 4056-4066.

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. *Nature structural & molecular biology*.

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 (New York, NY)* **293**: 1098-1102.

Henikoff S, Henikoff JG. 2012. "Point" centromeres of Saccharomyces harbor single centromere-specific nucleosomes. *Genetics* **190**: 1575-1577.

Hirano T. 2012. Condensins: universal organizers of chromosomes with diverse functions. *Genes Dev* **26**: 1659-1678.

Hoffmann S, Izquierdo HM, Gamba R, Chardon F, Dumont M, Keizer V, Hervé S, McNulty SM, Sullivan BA, Manel N et al. 2020. A genetic memory initiates the epigenetic loop necessary to preserve centromere position. *The EMBO journal* **39**: e105505.

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* **27**: 2365-2373 e2368.

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018. Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* **36**: 321-323.

Kanesaki Y, Imamura S, Matsuzaki M, Tanaka K. 2015. Identification of centromere regions in chromosomes of a unicellular red alga, Cyanidioschyzon merolae. *FEBS Lett* **589**: 1219-1224.

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

Kasinathan S, Henikoff S. 2018. Non-B-form DNA is enriched at centromeres. *Mol Biol Evol*.

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 T. 2014. Ribosomal RNA gene repeats, their stability and cellular senescence. *Proc Jpn Acad Ser B Phys Biol Sci* **90**: 119-129.

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

Kockler ZW, Osia B, Lee R, Musmaker K, Malkova A. 2021. Repair of DNA Breaks by Break-Induced Replication. *Annu Rev Biochem*.

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.

Liu J, Seetharam AS, Chougule K, Ou S, Swentowsky KW, Gent JI, Llaca V, Woodhouse MR, Manchanda N, Presting GG et al. 2020. Gapless assembly of maize chromosomes using long-read technologies. *Genome Biol* **21**: 121.

Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A et al. 2021. The structure, function and evolution of a complete human chromosome 8 *Nature*.

Marshall OJ, Choo KH. 2012. Putative CENP-B paralogues are not present at mammalian centromeres. *Chromosoma* **121**: 169-179.

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. *The Journal of cell biology* **109**: 1963-1973.

McKinley KL, Cheeseman IM. 2016. The molecular basis for centromere identity and function. *Nat Rev Mol Cell Biol* **17**: 16-29.

McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res* **26**: 115-138.

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 biology* **14**: R10.

Miga KH. 2020. Centromere studies in the era of 'telomere-to-telomere' genomics. *Exp Cell Res* **394**: 112127.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**: 79-84.

Mitra S, Gómez-Raja J, Larriba G, Dubey DD, Sanyal K. 2014. Rad51-Rad52 mediated maintenance of centromeric chromatin in Candida albicans. *PLoS Genet* **10**: e1004344.

Ohzeki J, Nakano M, Okada T, Masumoto H. 2002. CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *The Journal of cell biology* **159**: 765-775.

Ohzeki JI, Otake K, Masumoto H. 2020. Human artificial chromosome: Chromatin assembly mechanisms and CENP-B. *Exp Cell Res* **389**: 111900.

Otake K, Ohzeki JI, Shono N, Kugou K, Okazaki K, Nagase T, Yamakawa H, Kouprina N, Larionov V, Kimura H et al. 2020. CENP-B creates alternative epigenetic chromatin states permissive for CENP-A or heterochromatin assembly. *J Cell Sci* **133**.

Padmanabhan S, Thakur J, Siddharthan R, Sanyal K. 2008. Rapid evolution of Cse4p-rich centromeric DNA sequences in closely related pathogenic yeasts, Candida albicans and Candida dubliniensis. *Proc Natl Acad Sci U S A* **105**: 19797-19802.

Perry CA, Annunziato AT. 1989. Influence of histone acetylation on the solubility, H1 content and DNase I sensitivity of newly assembled chromatin. *Nucleic Acids Res* **17**: 4275-4291.

Plohl M, Mestrovic N, Mravinac B. 2014. Centromere identity from the DNA point of view. *Chromosoma* **123**: 313-325.

Prytkova TR, Zhu X, Widom J, Schatz GC. 2011. Modeling DNA-bending in the nucleosome: role of AA periodicity. *The journal of physical chemistryB* **115**: 8638-8644.

Rice W. 2020a. Why do centromeres evolve so fast: BIR replication, hypermutation, transposition, and molecular drive. *wwwpreprintsorg*.

Rice WR. 2020b. A Game of Thrones at Human Centromeres I. Multifarious structure necessitates a new molecular/evolutionary model. *bioRxiv*.

Rice WR. 2020c. A Game of Thrones at Human Centromeres II. A new molecular/evolutionary model. **bioRxiv**.

Sakofsky CJ, Malkova A. 2017. Break induced replication in eukaryotes: mechanisms, functions, and consequences. *Crit Rev Biochem Mol Biol* **52**: 395-413.

Sanyal K, Baum M, Carbon J. 2004. Centromeric DNA sequences in the pathogenic yeast Candida albicans are all different and unique. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 11374-11379.

Schneider KL, Xie Z, Wolfgruber TK, Presting GG. 2016. Inbreeding drives maize centromere evolution. *Proc Natl Acad Sci U S A* **113**: E987-996.

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. 2001. Genomic and genetic definition of a functional human centromere. *Science (New York, NY)* **294**: 109-115.

Senaratne AP, Muller H, Fryer KA, Kawamoto M, Katsuma S, Drinnenberg IA. 2021. Formation of the CenH3-Deficient Holocentromere in Lepidoptera Avoids Active Chromatin. *Curr Biol* **31**: 173-181.e177.

Smit AF, Riggs AD. 1996. Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* **93**: 1443-1448.

Smith GP. 1976. Evolution of repeated DNA sequences by unequal crossover. *Science (New York, NY)* **191**: 528-535.

Snowden T, Acharya S, Butz C, Berardini M, Fishel R. 2004. hMSH4-hMSH5 recognizes Holliday Junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. *Mol Cell* **15**: 437-451.

Sonnenberg ASM, Sedaghat-Telgerd N, Lavrijssen B, Ohm RA, Hendrickx PM, Scholtmeijer K, Baars JJP, van Peer A. 2020. Telomere-to-telomere assembled and centromere annotated genomes of the two main subspecies of the button mushroom Agaricus bisporus reveal especially polymorphic chromosome ends. *Sci Rep* **10**: 14653.

Struhl K, Segal E. 2013. Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**: 267-273.

Talbert PB, Henikoff S. 2020. What makes a centromere? *Exp Cell Res* **389**: 111895.

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. *The EMBO journal* **20**: 6612-6618.

Thakur J, Henikoff S. 2018. Unexpected conformational variations of the human centromeric chromatin complex. *Genes Dev* **32**: 20-25.

Wolfgruber TK, Nakashima MM, Schneider KL, Sharma A, Xie Z, Albert PS, Xu R, Bilinski P, Dawe RK, Ross-Ibarra J et al. 2016. High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events. *Front Plant Sci* **7**: 308.

Yan H, Kikuchi S, Neumann P, Zhang W, Wu Y, Chen F, Jiang J. 2010. Genome-wide mapping of cytosine methylation revealed dynamic DNA methylation patterns associated with genes and centromeres in rice. *Plant J* **63**: 353-365.

Yoda K, Ando S, Okuda A, Kikuchi A, Okazaki T. 1998. In vitro assembly of the CENP-B/alpha-satellite DNA/core histone complex: CENP-B causes nucleosome positioning. *Genes to cells : devoted to molecular & cellular mechanisms* **3**: 533-548.

Yunis JJ, Yasmineh WG. 1971. Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation. *Science* **174**: 1200-1209.

Zasadzińska E, Huang J, Bailey AO, Guo LY, Lee NS, Srivastava S, Wong KA, French BT, Black BE, Foltz DR. 2018. Inheritance of CENP-A Nucleosomes during DNA Replication Requires HJURP. *Dev Cell* **47**: 348-362.e347.

Zhang T, Talbert PB, Zhang W, Wu Y, Yang Z, Henikoff JG, Henikoff S, Jiang J. 2013. The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. *Proc Natl Acad Sci U S A* **110**: E4875-4883.

Zhang W, Lee HR, Koo DH, Jiang J. 2008. Epigenetic modification of centromeric chromatin: hypomethylation of DNA sequences in the CENH3-associated chromatin in Arabidopsis thaliana and maize. *Plant Cell* **20**: 25-34.

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, Birchler JA, Jiang J, Dawe RK. 2002. Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell* **14**: 2825-2836.