

Genomic and functional variation of human centromeres

Lori L. Sullivan^a, Beth A. Sullivan^{a,b,*}

^a Department of Molecular Genetics and Microbiology, USA

^b Division of Human Genetics, Duke University School of Medicine, Durham, NC, 27710, USA



ARTICLE INFO

Keywords:

Satellite DNA
Repetitive DNA
Epiallele
Haplotype
Dicentric chromosome
Kinetochore

ABSTRACT

Centromeres are central to chromosome segregation and genome stability, and thus their molecular foundations are important for understanding their function and the ways in which they go awry. Human centromeres typically form at large megabase-sized arrays of alpha satellite DNA for which there is little genomic understanding due to its repetitive nature. Consequently, it has been difficult to achieve genome assemblies at centromeres using traditional next generation sequencing approaches, so that centromeres represent gaps in the current human genome assembly. The role of alpha satellite DNA has been debated since centromeres can form, albeit rarely, on non-alpha satellite DNA. Conversely, the simple presence of alpha satellite DNA is not sufficient for centromere function since chromosomes with multiple alpha satellite arrays only exhibit a single location of centromere assembly. Here, we discuss the organization of human centromeres as well as genomic and functional variation in human centromere location, and current understanding of the genomic and epigenetic mechanisms that underlie centromere flexibility in humans.

The centromere is the chromosomal locus that ensures chromosome inheritance through cell division. The kinetochore, the multi-protein structure that attaches chromosomes to spindle microtubules, is assembled at the centromere, locally coordinating chromosome movement during cell division. Each chromosome must have a centromere; without one, the chromosome will be lost, leading to an aneuploid karyotype. A locus with such an essential role in genome stability in all organisms would be expected to exhibit similar genomic characteristics among organisms. Surprisingly, centromeric DNAs differ among organisms and even between different chromosomes of the same organism. Centromeres range in size from small point centromeres (~125bp) in budding yeasts to large regional centromeres (100 kb – 5 Mb) in humans and plants. Despite these sequence disparities, the proteins of eukaryotic centromeres are related, emphasizing the functional importance of the locus. Centromeres are defined by specialized nucleosomes containing the histone H3 variant CENP-A. Clusters of CENP-A nucleosomes are interspersed with groups of canonical H3 nucleosomes to create a unique type of centromeric (CEN) chromatin, also sometimes referred to as centrochromatin, that differentiates the centromere from the rest of the chromosome. CEN chromatin serves as the foundation of the kinetochore, interacting with CENP-C and other members of the constitutive centromere associated network (CCAN) to assemble the protein network between the DNA and the microtubules. CEN chromatin assembly occurs on DNA sequences that differ among

organisms and even within the same organism, suggesting a general lack of sequence specificity for CENP-A and other centromere proteins and pointing to CENP-A and/or CEN chromatin as an important epigenetic determinant of centromere identity and maintenance. The lack of sequence similarities at eukaryotic centromeres led to the current view of centromere identity as an epigenetic process, with little contribution from the underlying DNA.

1. The genomics of human centromeres

Native human centromeres are formed at regions of alpha satellite, a repetitive DNA that is defined by a 171bp monomeric sequence unit. Individual monomers that are 50–70% identical are arranged tandemly (Fig. 1a). A defined number of monomers create a higher order repeat (HOR) unit that is repeated hundreds to thousands of times to produce a large homogenous array in which the HOR units are 97–100% identical. The number and order of monomers within a HOR unit confers chromosome specificity, so that even though the same monomers are present at every centromere in the genome, their placement within a HOR unit contributes to individual alpha satellite arrays that can be molecularly distinguished [1]. For instance, the human X chromosome (HSAX) centromere is defined by a HOR unit of 12 monomers (DXZ1, 12-mer), whereas the human Y (HSAY) chromosome centromere is defined by a 34-mer HOR unit (DYZ3) [2,3]. Since the order and

* Corresponding author. Department of Molecular Genetics and Microbiology, USA.

E-mail address: beth.sullivan@duke.edu (B.A. Sullivan).

<https://doi.org/10.1016/j.yexcr.2020.111896>

Received 19 December 2019; Received in revised form 29 January 2020; Accepted 5 February 2020

Available online 06 February 2020

0014-4827/ © 2020 Elsevier Inc. All rights reserved.

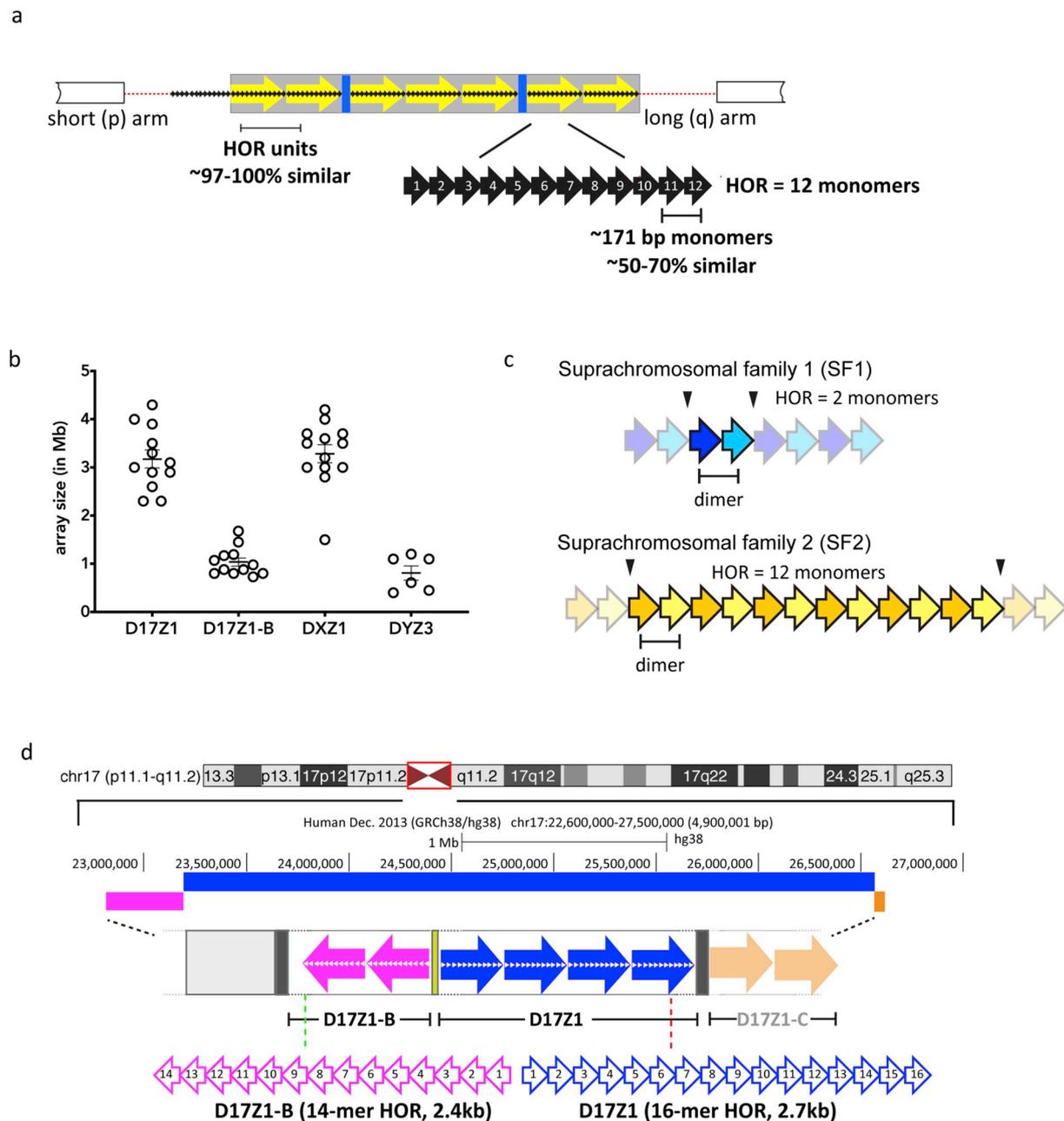


Fig. 1. Genomic organization and variation of alpha satellite DNA at human centromeres. **a.** Schematic of the general organization of an alpha satellite DNA array at human centromere regions. Alpha satellite is based on ~171bp monomeric repeat units (black arrows with white numbers) that are 50–70% identical in sequence and arranged tandemly to form a HOR unit; shown here as a 12 monomer HOR (yellow array). Monomers are numbered by their position within the HOR and not based on their homology between two distinct HORs. The HORs are repeated hundreds to thousands of times to create homogenous arrays in which HORs within a given array are 97–100% identical. The HOR array is flanked by degenerate alpha satellite DNA monomers (small black arrows) that lack hierarchical structure and separate the HOR array from the chromosome arms. HOR arrays are interrupted by other repetitive elements, such as transposable elements (TEs, blue), but the extent of TE distribution across specific alpha satellite arrays is currently unclear due to the lack of linear, contiguous alpha satellite assemblies. **b.** The number of times a chromosome-specific HOR unit is repeated to create an extensive homogenous array varies between homologs and individuals. The distribution of total array sizes from a subset of the population is shown for four distinct alpha satellite arrays, including D17Z1 and D17Z1-B, two arrays from *Homo sapiens* chromosome 17 (HSA17), as well as DXZ1 and DYZ3, the alpha satellite arrays from HSA18 and HSA19. Each open circle represents an independent chromosome/individual. **c.** Individual alpha satellite monomers are distributed into five *suprachromosomal groups* or *families*, based on the concentration of specific monomer types (based on sequence homology) on distinct chromosomes. Suprachromosomal families 1 and 2 (SF1, SF2) represent the most abundant HOR configuration within human centromere regions and are defined by distinct dimeric configurations (shown as blue or yellow). HOR units on some chromosomes like HSA1 (SF1) are defined by the two dimers themselves while larger HOR units on chromosomes like HSA18 (SF2) are comprised of multiple alternating copies of the same two monomers. The HOR units are then repeated hundreds to thousands of times to produce a large homogenous satellite array. **d.** Over half of native human chromosomes contain more than one distinctive alpha satellite array. For example, HSA17 has three arrays that are ~92% identical and defined by different HOR units. D17Z1, the largest array (blue) is defined by a canonical 16 monomer (16-mer) HOR that is repeated thousands of times to produce total array sizes that range from 2.3 to 4.2 Mb in the population. D17Z1-B (pink) and D17Z1-C (light orange) flank D17Z1 and are each defined by different 14-mer HOR units. Centromere assembly can occur at either D17Z1 or D17Z1-B. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

sequence of the monomers vary between chromosome-specific HORs, alpha satellite arrays on each chromosome can be distinguished molecularly by PCR, Southern blotting, and fluorescence in situ hybridization (FISH) under stringent conditions [4–6]. Despite basic structural and organizational features to differentiate alpha satellite arrays, the human genome assembly lacks contiguous alpha satellite sequences at centromere regions on every chromosome. Although monomers are present within short-read sequence datasets, they are present in hundreds to thousands of copies within the genome and at individual centromeres, so it is difficult to accurately and confidently assign a monomeric read to a HOR unit or multi-megabase array (reviewed in Ref. [7]). Thus, we do not have a complete understanding of alpha satellite organization and structure on individual chromosomes within the population, nor of genotype-function correlations, owing to the lack of contiguous centromeric reference genome assemblies.

2. Genomic variation within human centromeres

Sequence variation is a source of functional diversity that is typically studied in the context of gene expression and non-coding regulatory regions. It has been difficult to define the amounts and types of variation that are present within alpha satellite without genome assemblies. However, early molecular studies of HOR unit organization of some alpha satellite arrays have revealed several types of variation at human centromeres: 1) total size of the array, 2) structure of the HOR units within a given array, 3) sequence variation within the same HOR unit of a distinct array, and 4) number of independent arrays on a single chromosome [8–10].

2.1. Total size variation of chromosome-specific alpha satellite arrays

Alpha satellite is frequently thought to be identical at all human centromeres, but although each array is built from the same pool of monomers, it has its own distinctive organization within the human genome. Additionally, degrees of variation exist within an array. On a given chromosome, the number of times a HOR unit is repeated differs within the population, giving rise to a range of total array lengths between homologs in the same individual and among homologs within different individuals. For example, on individual HSA17s in the population, the 2.0 kb DXZ1 HOR is repeated 750–2100 times, yielding total array size lengths that range from 1.5 Mb to 4.2 Mb (mean of 3.0 Mb) (Fig. 1b) [3,11,12]. Alpha satellite arrays on autosomes also exhibit similar inter-homolog and inter-individual array size polymorphisms, such that population array lengths vary 10- to 20-fold. However, within a given family, alpha satellite array sizes are heritable and stable through meiosis, to the extent that specific chromosome homologs can be tracked molecularly through families based solely on alpha satellite array sizes [13–15]. Alpha satellite array sizes of the same chromosome have not been routinely compared among more than two tissues, but appear to be largely stable in phenotypically normal individuals [16]. Whether this remains true for all cells, including undifferentiated or aging cells, remains to be explored.

2.2. Structure and sequence variation of HOR units within specific arrays

Alpha satellite monomers differ in sequence by 10–40%. Two adjacent monomers may differ in sequence, but similarities in monomer sequence and their order in the HOR unit are shared among different chromosomes. Individual monomers are derived from twelve consensus alpha satellite monomers that are distributed among five *suprachromosomal groups or families*, based on the combination and concentration of these monomers on specific chromosomes. Suprachromosomal families 1 through 3 (SF1-3) represent the majority of “functional” alpha satellite HORs and are typically enriched for centromere proteins [17,18]. SF1 and SF2 are defined by distinct types of dimeric (two monomer) configurations (Fig. 1c) and are present on all chromosomes

with the exception of HSA1, HSA11, HSA17, and HSAX that are instead defined by the pentameric (five monomer) HOR configuration of SF3 [18–20]. While both monomers of dimeric HOR unit arrays are enriched for centromere proteins, not all monomers within pentameric HOR arrays are equally associated with centromere proteins [21].

The binding of Centromere Protein B (CENP-B) to alternate monomers in dimeric HOR subfamilies SF1 and SF2 confers enhanced binding and integrity of the constitutive centromere associated-network complex (CCAN). CENP-B is a DNA binding protein that recognizes a 17-bp sequence motif called the CENP-B box that is present in a subset of alpha satellite monomers [22,23]. The density of CENP-B boxes within HOR alpha satellite arrays has been correlated with stronger CENP-A enrichment, and led to the model that dimeric arrays exhibit the highest CENP-B box density [18]. Within pentameric HOR units of SF3, CENP-B boxes are irregularly spaced. Furthermore, the density of CENP-B boxes is also influenced by total array size, such that a 3 Mb dimeric HOR array will have the same number of CENP-B boxes as a 3 Mb pentameric HOR array that has irregularly spaced CENP-B boxes. A lower density of CENP-B boxes within an alpha satellite array may not entirely disqualify it for centromere assembly. Indeed, centromere assembly occurs on minor arrays of HOR alpha satellite even when a nearby major HOR array on the same chromosome has twice or thrice the number of CENP-B boxes [15,17,24].

3. Genomic variation within HORs of specific alpha satellite arrays

The suprachromosomal family classifications illustrate that variation within the alpha satellite DNA is common and complex, due to monomeric differences and chromosome-specific differences in HOR unit size and monomer order and organization. Although HOR units are largely thought to be highly identical within a given chromosome-specific array [25], some HOR units within the array contain structural variants such as monomer insertions or deletions, monomer or entire HOR unit inversions, and insertion of non-satellite elements such as transposons [26–28]. Thus, alpha satellite arrays actually exist as complex arrays of variant and canonical (wild-type) HORs punctuated with other repetitive elements [1,29].

HSA17 is one chromosomal example of HOR unit polymorphisms within an alpha satellite array. The predominant HOR unit on D17Z1 is a 16-monomer (16-mer) (Fig. 2a), and makes up most of the HOR units in D17Z1 within the population. However, single or multiple monomeric deletions within the canonical/wild-type 16-mer HOR unit have produced variant D17Z1 HOR units including 15-mers, 14-mers, 13-mers, 12-mers, and very rare 11-mers. These monomer deletions presumably arose by unequal recombination or gene conversion during meiosis [30,31]. Within humans, 15-, 14- and 12-mer HOR units represent a small fraction of the total HOR units within a given D17Z1 array. However, the 13-mer HOR unit is the most abundant variant HOR unit within D17Z1 arrays. As such, D17Z1 arrays in the population are defined by two haplotypes: 1) Haplotype I: the wild-type haplotype defined by arrays consisting of 16-mers, with occasional 15- and 14-mers, that are found on 65% of HSA17s, and 2) Haplotype II, the variant haplotype that is comprised of arrays that contain 13-mer HOR units and are present on 35% of HSA17s (Fig. 2a). The proportion of variant 13-mer HOR units within D17Z1 arrays ranges from as little as 1% to over 75% of the array, depending on the chromosome and the individual [15].

In addition to HOR unit variation, single nucleotide changes have also been mapped to specific monomers within some of the HOR units of alpha satellite arrays [3,4,25,32]. The functional significance of size and sequence variants in alpha satellite is unclear. Centromere-specific/centromere-proximal haplotypes (cenhaps) are present within human populations [33], and might be associated with different long-term chromosome stability outcomes. Ultra-long (UL) read sequence assemblies of DXZ1 (HSAX) and DYZ3 (HSAY) have revealed multiple types of

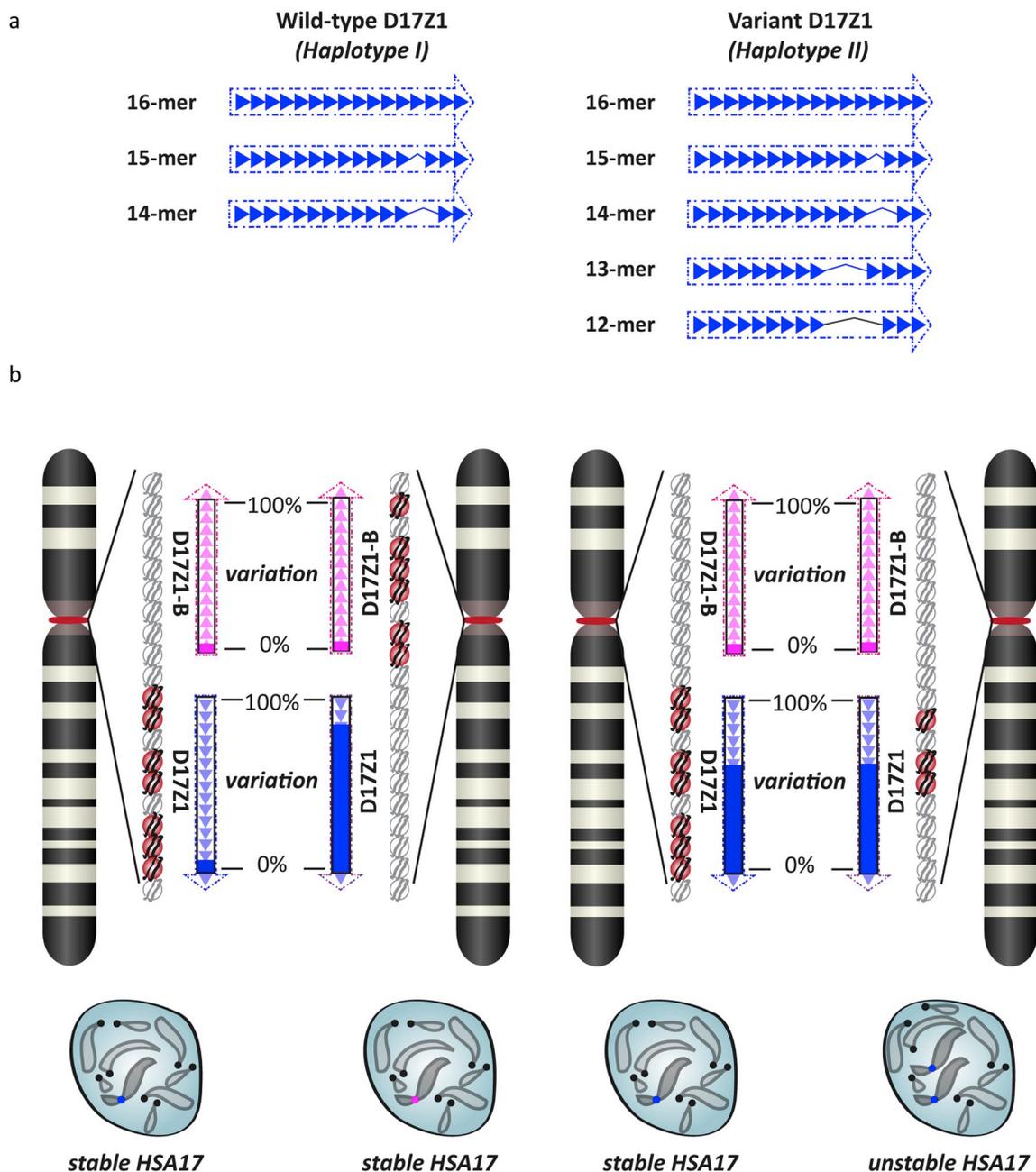


Fig. 2. Functional centromeric epialleles on *Homo sapiens* chromosome 17 (HSA17) are influenced by genomic variation within alpha satellite DNA. **a.** The large array D17Z1 on HSA17 is defined by a canonical (wild-type) 16-mer HOR unit. However, D17Z1 is highly polymorphic such that single and multiple monomer deletions produce HOR variants that differ in length by an integral number of monomers. In the general population, HOR variants range from 15-mers to 12-mers. Two major haplotypes (Haplotype I - wildtype; Haplotype II - variant) exist in the population, distinguished by the presence or absence of the 13-mer HOR unit. **b.** HSA17 exhibits centromeric epialleles (i.e. multiple sites of centromere assembly) based on the amount of variation within D17Z1. Arrays containing predominantly wild-type 16-mer HORs are preferred locations for centromere assembly, denoted by CENP-A (red circles), and are associated with stable HSA17s. When D17Z1 arrays are composed of more than 70% variant 13-mer HORs (higher variation denoted by greater solid blue shading), centromere assembly occurs on neighboring D17Z1-B (pink arrows) and chromosome stability is comparable to HSA17s with wild-type arrays. However, centromere assembly occurs on D17Z1 arrays exhibiting intermediate levels of variation (40–60%) but leads to two distinct chromosome phenotypes: stable and unstable. Variant arrays exhibiting the unstable HSA17 phenotype are associated with reduced numbers of centromere proteins. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

structural variation within alpha satellite, including single and multiple monomeric insertions and deletions that changed HOR size and in the case of DXZ1, a single LINE-1 (L1H) insertion [3,26].

3.1. Multi-array chromosomes exhibit variation in the number of independent alpha satellite arrays within the human karyotype

Although it was noted in the 1990s that the acrocentric

chromosomes (HSA13, 14, 15, 21, and 22) had more than one distinct alpha satellite, it has been more recently appreciated that over half of human chromosomes, such as HSA1, HSA7, HSA17, HSA18, HSA19, and HSA20, contain two or even three HOR arrays [17,20]. HSA17 for example contains three distinct arrays, D17Z1, D17Z1-B, and D17Z1-C (Fig. 1d). The presence of multiple arrays on the same chromosome has led to two models: 1) one array is “active/live” and the other is “inactive/dead”; or 2) native chromosomes are arranged as structurally

dicentric or trivalent chromosomes with arrays that retain functional potential. Studies using human artificial chromosomes (HACs) to test the functional potential of “live” or “dead” arrays argue against the first model for chromosomes such as HSA7 and HSA17 [17,34] (see section on “*Functional variation of centromeres*”). Arrays lacking HOR unit structure are incapable of supporting *de novo* centromere function, so there appears to be a clear distinction between the functional potential of HOR arrays and monomeric arrays [17,35].

4. Epigenomic variation of human centromeres

4.1. Human centromeres have an epigenomic RNA component

Centromeres and heterochromatin were long thought to be transcriptionally inert because they can effectively silence transgenes or genes that are inserted in or nearby the centromere [36]. However, studies in the fission yeast *Schizosaccharomyces pombe* (*S. pombe*) first revealed that small RNAs produced from centromere regions are required to establish and reinforce heterochromatin at the centromere and pericentromere [37–40]. Mammalian pericentromeres also have an RNA component that appears to be linked to the cell cycle [38,39,41].

Alpha satellite DNA is incorporated into both CEN chromatin and pericentric heterochromatin [5,42]. CEN chromatin covers 30–45% of a single array of alpha satellite at native chromosomes or centromeres of human artificial chromosomes (HACs) [5,43–45]. Within CEN chromatin, the H3 nucleosomes within CEN chromatin are marked by dimethylation on lysine 4 (K4me2) and lysine 36 (K36me2), post-translational histone modifications that are typically associated with euchromatin. The remaining portion of the alpha satellite array is assembled into heterochromatin marked by histone modifications such as H3K9me2, H3K9me3, and H3K27me3 [5,43–46]. Each chromosome is associated with sequence-specific, array-specific non-coding transcripts. The alpha satellite non-coding transcripts are produced *in cis* and are complexed with centromere proteins CENP-A and CENP-C, as well as the alpha satellite DNA binding protein CENP-B [24,47]. Although CENP-C is a known RNA-binding protein, the specific binding sites for alpha satellite RNAs on CENP-C and other centromere proteins have not yet been identified. However, non-coding alpha satellite RNAs also recruit the histone lysine methyltransferases SUV39H1/2 that add methyl groups to H3 at lysine 9 to create a binding site for heterochromatin protein 1 (HP1) [40,48]. How a single alpha satellite array is apportioned into multiple chromatin domains is not clear, although the cell cycle timing, structure, or long-term stability of alpha satellite transcripts may be a factor in their ultimate function. Alpha satellite transcripts have been proposed to exist as RNA:DNA hybrids or as single-stranded RNAs [40,49], so it is possible that distinct sets of transcripts or their local concentrations at various regions of the alpha satellite array feed into a CEN chromatin versus heterochromatin assembly pathway.

4.2. DNA methylation within centromere regions

Centromere regions are also enriched for DNA methylation that can be linked to histone methylation. For instance, methylation of mouse minor (centromere) and major (pericentromere) satellite DNAs are important for chromosome and satellite array stability [50–52]. Alpha satellite DNA is hyper-methylated, presumably through the action of DNMT3B that interacts with CENP-C [53,54]. Variation in enrichment of DNA methylation between different alpha satellite arrays in the same or different individuals is not well understood, but technological advancements may help resolve this gap in knowledge. Nanopore and PacBio sequencing technologies can detect modifications to DNA bases, such as cytosine methylation, through either alternation in normal electrolytic current signals or polymerase synthesis rates, respectively, when a methylated cytosine is encountered. The use of nanopore UL-reads to fully assemble the human X chromosome has provided

information on DNA methylation within the DXZ1 alpha satellite array of a single individual [26]. Although DXZ1 was generally methylated throughout the array, a 60 kb region of hypomethylation was detected. A comparatively sized (75 kb) region of hypomethylation was also detected within the manually assembled D8Z2 array on HSA8. Similar hypomethylation of the satellite repeats in *Arabidopsis thaliana* and *Zea mays* has been reported previously [55]. The functional relevance of hypermethylated and hypomethylated regions within human centromeres remains to be fully determined, although it possible that hypomethylation demarcates distinct chromatin domains within centromere regions or ensures the faithful binding of centromere proteins [56].

5. Functional variation of human centromeres

Human centromere function varies in a few notable ways. The most drastic example is the occurrence of neocentromeres, atypical centromeres that arise spontaneously at non-canonical (i.e. non alpha satellite) sequences [57–59], either on broken chromosomes or when the native centromere loses function or is functionally mutated. The basis of such functional variation in centromere location remains a topic of intense interest, and approaches to engineer neocentromeres are aimed at exploring (epi)genomic features that promote neocentromere formation. Here, we focus on variability in native centromere function.

5.1. Stability of native chromosomes as a measure of centromere fitness

One alpha satellite array on each chromosome has the highest local CENP-A concentration and is where new CENP-A is deposited each cell cycle [44,60]. Thus, once established by CENP-A incorporation, a centromere is maintained at the same location on a portion of alpha satellite DNA [44]. Approximately 35% of any given array is assembled into CEN chromatin (i.e. CENP-A plus H3K4me2 nucleosomes combined). Considering the range of alpha satellite array lengths in the population, CEN chromatin domain lengths on a chromosome-specific array can range from 500 kb to 1.5 Mb [6,43]. Although differences in the amount of total centromere proteins between large and small chromosomes and large and small alpha satellite arrays have been reported [61,62], quantitative analyses of CENP-A nucleosomes at native centromeres have concluded that the number of CENP-A molecules does not vary largely among all centromeres [44,60]. These results suggest that the sizes and/or number of CENP-A-containing and H3-containing subdomains may differ on individual chromosome-specific arrays. Although patterns of subdomain organization remain to be validated experimentally at native human centromeres, ChIP studies of a neocentromere formed on HSA10 (mardel10) showed a range of CENP-A domain sizes across the CEN chromatin domain [63]. Thus, not only could subdomain sizes differ between individual arrays, they might also differ on the same array.

Variation in CENP-A and H3 subdomain sizes may reflect inherent natural flexibility in centromere organization and normal homeostasis of CENP-A enrichment on a range of alpha satellite array lengths. The establishment and maintenance of CENP-A domain sizes may be dictated by sequence and organization of the HOR units within an array. Several recent studies suggest that under perturbing conditions that challenge chromosome stability, the genomic composition of centromeres influences the sensitivity of the centromere [64,65]. A subset of alpha satellite monomers contain the 17bp CENP-B box motif where CENP-B binds [66]. The location of CENP-B boxes varies based on chromosome-specific HORs and is linked to the HOR structure [22]. Some HORs have multiple CENP-B boxes, while others like the 16-mer HOR D7Z2 on HSA7, have only a single CENP-B box (reviewed in Refs. [20,67]). The DY3Z array on HSA10 completely lacks CENP-B boxes [22,68]. Overall, the range of total array lengths in the population also reflect differences in the total number of functionally significant CENP-B boxes. A recent study reported that alpha satellite arrays with more

CENP-B boxes and thus greater CENP-B binding were less likely to mis-segregate [65]. The increased number of CENP-B boxes correlated with increased binding of CENP-C, indicating that centromeres that are less sensitized to destabilizing conditions contain more functional CENP-B boxes that recruit CENP-B and CENP-C [69]. Longer alpha satellite arrays might be predicted to build stronger centromeres because they theoretically have more CENP-B boxes, but the correlation between centromere length and centromere strength was less significant. Not all CENP-B boxes in alpha satellite arrays are incorporated into kinetochore chromatin, since a portion of the array is also assembled into heterochromatin [5,70]. Regulation of the binding of CENP-B and CENP-C across the alpha satellite array, perhaps through methylation of CENP-B boxes that inhibits CENP-B binding and/or controls satellite transcription [54,56], may influence or control centromere strength.

5.2. Centromeric epialleles: variation in alpha satellite array location of centromere assembly

As mentioned above, HSA17 is unique in that it has three alpha satellite arrays D17Z1, D17Z1-B and D17Z1-C. HAC technology showed that D17Z1 and D17Z1-B arrays can independently support *de novo* centromere formation (Fig. 2) [17,34], suggesting that in vivo, centromere location might vary on native chromosomes. Indeed, on HSA17, either D17Z1 or D17Z1-B can be the site of centromere assembly [15,17,34]. These *centromeric epialleles* are mitotically and meiotically stable [15]. They also are not exclusive to HSA17, since recent studies have identified centromeric epialleles on multiple chromosomes [20,24,71,72], indicating that flexibility in centromere location is an inherent, and previously unappreciated, property of native human chromosomes.

An intriguing question is how centromere location is determined on multi-array chromosomes. On HSA17, D17Z1 is a large array that ranges in length from 2 to 4 Mb and contains many CENP-B boxes, compared to D17Z1-B that is a smaller array (0.5–1.5 Mb) and contains fewer (but perhaps a proportionately equal number of) CENP-B boxes. Although D17Z1 is predominantly the site of centromere assembly [15,71], even when it is not the site of centromere assembly, its array length exceeds that of D17Z1-B. Like recent studies, this suggests that centromere length does not completely predict centromere location. Instead, the variation within D17Z1 arrays (described above in the section *Structure and sequence variation of HOR units*) appears to influence centromere location on HSA17. On HSA17s in which D17Z1 contains little to no variation, the centromere forms at D17Z1; conversely when HSA17 has a D17Z1 array with greater than 70% variation (i.e. more 13-mer HORs versus wild-type 16-mers), centromere assembly occurs on D17Z1-B (Fig. 2b). Centromeres formed on wild-type (16-mer HOR) D17Z1 arrays or on D17Z1-B arrays when D17Z1 harbors extensive variation appear to be highly efficient and are rarely associated with HSA17 instability or aneuploidy.

Nevertheless, the presence of variation within D17Z1 does not absolutely correlate with centromeric epialleles and formation of the centromere at D17Z1-B. In fact, variant D17Z1 arrays can be the site of centromere assembly, but many of these HSA17s exhibit chromosome instability and increased aneuploidy long-term [15]. Variant, active D17Z1 arrays suggest that the *amount of variation* within an alpha satellite array is functionally important. In other words, centromere assembly occurs on D17Z1 arrays with moderate variation (40–60% of the array contains variant HOR units), but there is a functional cost (Fig. 2b). Unstable HSA17s that build the centromere on variant D17Z1 arrays are associated with reduced amounts of CENP-A and CENP-C, suggesting that variant alpha satellite DNA assembles a kinetochore that is defective in architecture and/or composition [15]. However, not every variant D17Z1 array is associated with a completely defective kinetochore and/or chromosome mis-segregation, raising the possibility that differences in long-range organization (i.e. where specific HOR units are located with respect to one another within an array) and

perhaps even transcription of distinct variant arrays are linked to a range of centromere strength and kinetochore quality outcomes.

The molecular basis for reduced association of CENPs and inefficient centromere function on variant D17Z1 alpha satellite arrays remains to be specifically determined. However, studies of HSA17 and other human centromeres point toward DNA-dependent alpha satellite factors, including alpha satellite sequence variation, that strongly affect centromere assembly and chromosome stability. The opportunity now exists to combine the newest ultra-long read sequencing technologies and optical mapping approaches with molecular and functional assays to comprehensively capture alpha satellite organization and variation within the population and test how specific (epi)genomic configurations of alpha satellite impact centromere assembly and chromosome stability.

CRedit authorship contribution statement

Lori L. Sullivan: Conceptualization, Writing - review & editing.
Beth A. Sullivan: Conceptualization, Writing - original draft, Visualization, Supervision, Funding acquisition.

Acknowledgements

This work was supported by NIH grants R01 GM124041, R01 GM129263, and R21 CA238758 to B.A.S. We are especially grateful to Rachel J. O'Neill (University of Connecticut, Storrs) for conceptualization and design ideas for Fig. 2.

References

- [1] H.F. Willard, J.S. Wayne, Hierarchical order in chromosome-specific human alpha satellite DNA, *Trends Genet.* (1987), [https://doi.org/10.1016/0168-9525\(87\)90232-0](https://doi.org/10.1016/0168-9525(87)90232-0).
- [2] J.S. Wayne, H.F. Willard, Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome, *Nucleic Acids Res.* 13 (1985) 2731–2743, <https://doi.org/10.1093/nar/13.8.2731>.
- [3] K.H. Miga, Y. Newton, M. Jain, N. Altomare, H.F. Willard, W.J. Kent, Centromere reference models for human chromosomes X and Y satellite arrays, *Genome Res.* 24 (2014) 697–707, <https://doi.org/10.1101/gr.159624.113>.
- [4] P.E. Warburton, G.M. Greig, T. Haaf, H.F. Willard, PCR amplification of chromosome-specific alpha satellite DNA: definition of centromeric STS markers and polymorphic analysis, *Genomics* 11 (1991) 324–333.
- [5] A.L. Lam, C.D. Boivin, C.F. Bonney, M.K. Rudd, B.A. Sullivan, Human centromeric chromatin is a dynamic chromosomal domain that can spread over noncentromeric DNA, *Proc. Natl. Acad. Sci. U.S.A.* 103 (2006) 4186–4191.
- [6] L.L. Sullivan, C.D. Boivin, B. Mravinac, I.Y. Song, B.A. Sullivan, Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells, *Chromosome Res.* 19 (2011) 457–470, <https://doi.org/10.1007/s10577-011-9208-5>.
- [7] K.H. Miga, Completing the human genome: the progress and challenge of satellite DNA assembly, *Chromosome Res.* 23 (2015) 421–426, <https://doi.org/10.1007/s10577-015-9488-2>.
- [8] H.F. Willard, J.S. Wayne, Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat, *J. Mol. Evol.* 25 (1987) 207–214.
- [9] M.K. Rudd, H.F. Willard, Analysis of the centromeric regions of the human genome assembly, *Trends Genet.* 20 (2004) 529–533.
- [10] H.F. Willard, Centromeres of mammalian chromosomes, *Trends Genet.* 6 (1990) 410–416.
- [11] M.M. Mahtani, H.F. Willard, Pulsed-field gel analysis of alpha-satellite DNA at the human X chromosome centromere: high-frequency polymorphisms and array size estimate, *Genomics* 7 (1990) 607–613.
- [12] K.H. Miga, Chromosome-specific centromere sequences provide an estimate of the ancestral chromosome 2 fusion event in hominin genomes, *J. Hered.* (2017), <https://doi.org/10.1093/jhered/esw039>.
- [13] R. Wevrick, H.F. Willard, Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability, *Proc. Natl. Acad. Sci. U.S.A.* 86 (1989) 9394–9398.
- [14] H.F. Willard, G.M. Greig, V.E. Powers, J.S. Wayne, Molecular organization and haplotype analysis of centromeric DNA from human chromosome 17: implications for linkage in neurofibromatosis, *Genomics* 1 (1987) 368–373.
- [15] M.E. Aldrup-MacDonald, M.E. Kuo, L.L. Sullivan, K. Chew, B.A. Sullivan, Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles, *Genome Res.* 26 (2016) 1301–1311, <https://doi.org/10.1101/gr.206706.116>.

- [16] R. Wevrick, W.C. Earnshaw, P.N. Howard-Peebles, H.F. Willard, Partial deletion of alpha satellite DNA associated with reduced amounts of the centromere protein CENP-B in a mitotically stable human chromosome rearrangement, *Mol. Cell Biol.* 10 (1990) 6374–6380.
- [17] K.E. Hayden, E.D. Strome, S.L. Merrett, H.R. Lee, M.K. Rudd, H.F. Willard, Sequences associated with centromere competency in the human genome, *Mol. Cell Biol.* 23 (2003) 763–772, <https://doi.org/10.1128/MCB.01198-12>.
- [18] J.G. Henikoff, J. Thakur, S. Kasinathan, S. Henikoff, A unique chromatin complex occupies young a-satellite arrays of human centromeres, *Sci. Adv.* 1 (2015) e1400234, <https://doi.org/10.1126/sciadv.1400234>.
- [19] T.D. Mashkova, T.A. Akopian, L.Y. Romanova, S.P. Mitkevich, Y.B. Yurov, L.L. Kisselev, I.A. Alexandrov, Genomic organization, sequence and polymorphism of the human chromosome 4-specific alpha-satellite DNA, *Gene* 140 (1994) 211–217.
- [20] S.M. McNulty, B.A. Sullivan, Alpha satellite DNA biology: finding function in the recesses of the genome, *Chromosome Res.* 26 (2018) 115–138, <https://doi.org/10.1007/s10577-018-9582-3>.
- [21] D. Hasson, T. Panchenko, K.J. Salimian, M.U. Salman, N. Sekulic, A. Alonso, P.E. Warburton, B.E. Black, The octamer is the major form of CENP-A nucleosomes at human centromeres, *Nat. Struct. Mol. Biol.* 20 (2013) 687–695, <https://doi.org/10.1038/nsmb.2562>.
- [22] Y. Muro, H. Masumoto, K. Yoda, N. Nozaki, M. Ohashi, T. Okazaki, Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box, *J. Cell Biol.* 116 (1992) 585–596.
- [23] H. Masumoto, H. Masukata, Y. Muro, N. Nozaki, T. Okazaki, A human centromere antigen (CENP-B) interacts with a short specific sequence in aliphoid DNA, a human centromeric satellite, *J. Cell Biol.* 109 (1989) 1963–1973, <https://doi.org/10.1083/jcb.109.5.1963>.
- [24] S.M. McNulty, L.L. Sullivan, B.A. Sullivan, Human centromeres produce chromosome-specific and array-specific alpha satellite transcripts that are complexed with CENP-A and CENP-C, *Dev. Cell* 42 (2017) 226–240, <https://doi.org/10.1016/j.devcel.2017.07.001>.
- [25] S.J. Durfy, H.F. Willard, Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: evidence for short-range homogenization of tandemly repeated DNA sequences, *Genomics* 5 (1989) 810–821, [https://doi.org/10.1016/0888-7543\(89\)90123-7](https://doi.org/10.1016/0888-7543(89)90123-7).
- [26] K.H. Miga, S. Koren, A. Rhie, M.R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G.A. Logsdon, V.A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G.G. Bouffard, A.M. Chang, N.F. Hansen, F. Thibaud-Nissen, A.D. Schmitt, J.-M. Belton, S. Selvaraj, M.Y. Dennis, D.C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N.J. Loman, N. Holmes, M. Loose, U. Surti, R. ana Risques, T.A.G. Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J.C. Mullikin, P.A. Pevzner, J.L. Gerton, B.A. Sullivan, E.E. Eichler, A.M. Phillippy, Telomere-to-telomere Assembly of a Complete Human X Chromosome, *BioRxiv*, 2019, <https://doi.org/10.1101/735928>.
- [27] K.H. Miga, Centromeric satellite DNAs: hidden sequence variation in the human population, *Genes* 10 (2019) E352, <https://doi.org/10.3390/genes10050352>.
- [28] J. Zahn, M.H. Kaplan, S. Fischer, M. Dai, F. Meng, A.K. Saha, P. Cervantes, S.M. Chan, D. Dube, G.S. Omenn, D.M. Markovitz, R. Contreras-Galindo, Expansion of a novel endogenous retrovirus throughout the pericentromeres of modern humans, *Genome Biol.* 16 (2015) 74, <https://doi.org/10.1186/s13059-015-0641-1>.
- [29] J.S. Wayne, H.F. Willard, Molecular analysis of a deletion polymorphism in alpha satellite of human chromosome 17: evidence for homologous unequal crossing-over and subsequent fixation, *Nucleic Acids Res.* 14 (1986) 6915–6927.
- [30] P.E. Warburton, J.S. Wayne, H.F. Willard, Nonrandom localization of recombination events in human alpha satellite repeat unit variants: implications for higher-order structural characteristics within centromeric heterochromatin, *Mol. Cell Biol.* 13 (1993) 6520–6529, <https://doi.org/10.1128/mcb.13.10.6520>.
- [31] P.E. Warburton, H.F. Willard, Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: evidence for concerted evolution along haplotypic lineages, *J. Mol. Evol.* 41 (1995) 1006–1015, <https://doi.org/10.1007/BF00173182>.
- [32] S.J. Durfy, H.F. Willard, Molecular analysis of a polymorphic domain of alpha satellite from the human X chromosome, *Am. J. Hum. Genet.* 41 (1987) 391–401.
- [33] S.A. Langley, K.H. Miga, G.H. Karpen, C.H. Langley, Haplotypes Spanning Centromeric Regions Reveal Persistence of Large Blocks of Archaic DNA, (2019), <https://doi.org/10.7554/eLife.42989.001> Elife.
- [34] K.A. Maloney, L.L. Sullivan, J.E. Matheny, E.D. Strome, S.L. Merrett, A. Ferris, B.A. Sullivan, Functional epialleles at an endogenous human centromere, *Proc. Natl. Acad. Sci. U.S.A.* 109 (2012) 13704–13709, <https://doi.org/10.1073/pnas.1203126109>.
- [35] M. Ikeno, B. Grimes, T. Okazaki, M. Nakano, K. Saitoh, H. Hoshino, N.I. McGill, H. Cooke, H. Masumoto, Construction of YAC-based mammalian artificial chromosomes, *Nat. Biotechnol.* 16 (1998) 431–439.
- [36] K. Ekwall, T. Olsson, B.M. Turner, G. Cranston, R.C. Allshire, Transient inhibition of histone deacetylation alters the structural and functional imprint at fission yeast centromeres, *Cell* 91 (1997) 1021–1032.
- [37] T. Fukagawa, M. Nogami, M. Yoshikawa, M. Ikeno, T. Okazaki, Y. Takami, T. Nakayama, M. Oshimura, Dicer is essential for formation of the heterochromatin structure in vertebrate cells, *Nat. Cell Biol.* 6 (2004) 784–791.
- [38] M. Guenatri, D. Bailly, C. Maison, G. Almouzni, Mouse centric and pericentric satellite repeats form distinct functional heterochromatin, *J. Cell Biol.* 166 (2004) 493–505.
- [39] J. Lu, D.M. Gilbert, Proliferation-dependent and cell cycle regulated transcription of mouse pericentric heterochromatin, *J. Cell Biol.* 179 (2007) 411–421.
- [40] W.L. Johnson, W.T. Yewdell, J.C. Bell, S.M. McNulty, Z. Duda, R.J. O'Neill, B.A. Sullivan, A.F. Straight, RNA-dependent stabilization of SUV39H1 at constitutive heterochromatin, *Elife* 6 (2017), <https://doi.org/10.7554/eLife.25299>.
- [41] C. Maison, D. Bailly, A.H.F.M. Peters, J.P. Quivy, D. Roche, A. Taddei, M. Lachner, T. Jenwein, G. Almouzni, Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component, *Nat. Genet.* 30 (2002) 329–334, <https://doi.org/10.1038/ng843>.
- [42] B.A. Sullivan, G.H. Karpen, Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin, *Nat. Struct. Mol. Biol.* 11 (2004) 1076–1083.
- [43] B. Mravinac, L.L. Sullivan, J.W. Reeves, C.M. Yan, K.S. Kopf, C.J. Farr, M.G. Schueler, B.A. Sullivan, Histone modifications within the human X centromere region, *PLoS One* 4 (2009) e6602, <https://doi.org/10.1371/journal.pone.0006602>.
- [44] J.E. Ross, K.S. Woodlief, B.A. Sullivan, Inheritance of the CENP-A centromere domain is spatially and temporally constrained at human centromeres, *Epigenet. Chromatin* 9 (2016) 20, <https://doi.org/10.1186/s13072-016-0071-7>.
- [45] J.M. Spence, R. Critcher, T.A. Ebersole, M.M. Valdivia, W.C. Earnshaw, T. Fukagawa, C.J. Farr, Co-localization of centromere activity, proteins and topoisomerase II within a subdomain of the major human X alpha-satellite array, *EMBO J.* 21 (2002) 5269–5280.
- [46] J.H. Bergmann, N.M. Martins, V. Larionov, H. Masumoto, W.C. Earnshaw, HACKING the centromere chromatin code: insights from human artificial chromosomes, *Chromosome Res.* 20 (2012) 505–519, <https://doi.org/10.1007/s10577-012-9293-0>.
- [47] D. Quenet, Y. Dalal, A long non-coding RNA is required for targeting centromeric protein A to the human centromere, *Elife* 3 (2014) e03254, <https://doi.org/10.7554/eLife.03254>.
- [48] M. Lachner, D. O'Carroll, S. Rea, K. Mechtler, T. Jenwein, Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins, *Nature* 410 (2001) 116–120, <https://doi.org/10.1038/35065132>.
- [49] L. Kabeche, H.D. Nguyen, R. Buisson, L. Zou, A mitosis-specific and R loop-driven ATR pathway promotes faithful chromosome segregation, *Science* 359 (2018) 108–114, <https://doi.org/10.1126/science.aan6490>.
- [50] B. Lehnertz, Y. Ueda, A.A. Derijck, U. Braunschweig, L. Perez-Burgos, S. Kubicek, T. Chen, E. Li, T. Jenwein, A.H. Peters, Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin, *Curr. Biol.* 13 (2003) 1192–1200.
- [51] J.H.A. Martens, R.J. O'Sullivan, U. Braunschweig, S. Opravil, M. Radolf, P. Steinlein, T. Jenwein, The profile of repeat-associated histone lysine methylation states in the mouse epigenome, *EMBO J.* 24 (2005) 800–812, <https://doi.org/10.1038/sj.emboj.7600545>.
- [52] I. Jaco, A. Canela, E. Vera, M.A. Blasco, Centromere mitotic recombination in mammalian cells, *J. Cell Biol.* 181 (2008) 885–892, <https://doi.org/10.1083/jcb.200803042>.
- [53] Fachinetti Scelfo, Keeping the centromere under control: a promising role for DNA methylation, *Cells* 8 (2019) E912, <https://doi.org/10.3390/cells8080912>.
- [54] S. Gopalakrishnan, B.A. Sullivan, S. Trazzi, G. Della Valle, K.D. Robertson, DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions, *Hum. Mol. Genet.* 18 (2009) 3178–3193.
- [55] W. Zhang, H.R. Lee, D.H. Koo, J. Jiang, Epigenetic modification of centromeric chromatin: hypomethylation of DNA sequences in the CENH3-associated chromatin in Arabidopsis thaliana and maize, *Plant Cell* 20 (2008) 25–34, <https://doi.org/10.1105/tpc.107.057083>.
- [56] Y. Tanaka, H. Kurumizaka, S. Yokoyama, CpG methylation of the CENP-B box reduces human CENP-B binding, *FEBS J.* 272 (2005) 282–289, <https://doi.org/10.1111/j.1432-1033.2004.04406.x>.
- [57] D. du Sart, M.R. Cancilla, E. Earle, J.I. Mao, R. Saffery, K.M. Tainton, P. Kalitsis, J. Martyn, A.E. Barry, K.H. Choo, A functional neo-centromere formed through activation of a latent human centromere and consisting of non-alpha-satellite DNA, *Nat. Genet.* 16 (1997) 144–153.
- [58] T.W. Depinet, J.L. Zackowski, W.C. Earnshaw, S. Kaffe, G.S. Sekhon, R. Stallard, B.A. Sullivan, G.H. Vance, D.L. Van Dyke, H.F. Willard, A.B. Zinn, S. Schwartz, Characterization of neo-centromeres in marker chromosomes lacking detectable alpha-satellite DNA, *Hum. Mol. Genet.* 6 (1997) 1195–1204.
- [59] L.E. Voullaire, H.R. Slater, V. Petrovic, K.H. Choo, A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? *Am. J. Hum. Genet.* 52 (1993) 1153–1163.
- [60] D.L. Bodor, J.F. Mata, M. Sergeev, A.F. David, K.J. Salimian, T. Panchenko, D.W. Cleveland, B.E. Black, J. V. Shah, L.E. Jensen, The quantitative architecture of centromeric chromatin, *Elife* 3 (2014) e02137, <https://doi.org/10.7554/eLife.02137>.
- [61] D. V. Irvine, D.J. Amor, J. Perry, N. Sirvent, F. Pedetour, K.H. Choo, R. Saffery, Chromosome size and origin as determinants of the level of CENP-A incorporation into human centromeres, *Chromosome Res.* 12 (2004) 805–815.
- [62] A.W. Lo, G.C. Liao, M. Rocchi, K.H. Choo, Extreme reduction of chromosome-specific alpha-satellite array is unusually common in human chromosome 21, *Genome Res.* 9 (1999) 895–908.
- [63] A.W. Lo, D.J. Magliano, M.C. Sibson, P. Kalitsis, J.M. Craig, K.H. Choo, A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA, *Genome Res.* 11 (2001) 448–457.
- [64] J.T. Worrall, N. Tamura, A. Mazzagatti, N. Shaikh, T. van Ling, B. Bakker, D.C.J. Spierings, E. Vladimirov, F. Fojter, S.E. McClelland, Non-random mis-segregation of human chromosomes, *Cell Rep.* 23 (2018) 3366–3380, <https://doi.org/10.1016/j.celrep.2018.05.047>.
- [65] M. Dumont, R. Gamba, P. Gestraud, S. Klaasen, J.T. Worrall, S.G. De Vries,

- V. Boudreau, C. Salinas-Luypaert, P.S. Maddox, S.M. Lens, G.J. Kops, S.E. McClelland, K.H. Miga, D. Fachinetti, Human chromosome-specific aneuploidy is influenced by DNA-dependent centromeric features, *EMBO J.* 39 (2019) e102924, <https://doi.org/10.15252/embj.2019102924>.
- [66] H. Masumoto, H. Masukata, Y. Muro, N. Nozaki, T. Okazaki, A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite, *J. Cell Biol.* 109 (1989) 1963–1973.
- [67] M. Dumont, D. Fachinetti, DNA sequences in centromere formation and function, *Prog. Mol. Subcell. Biol.* 56 (2017) 305–336, https://doi.org/10.1007/978-3-319-58592-5_13.
- [68] T. Haaf, A.G. Mater, J. Wienberg, D.C. Ward, Presence and abundance of CENP-B box sequences in great ape subsets of primate-specific alpha-satellite DNA, *J. Mol. Evol.* 41 (1995) 487–491.
- [69] D. Fachinetti, J.S. Han, M.A. McMahon, P. Ly, A. Abdullah, A.J. Wong, D.W. Cleveland, DNA sequence-specific binding of CENP-B enhances the fidelity of human centromere function, *Dev. Cell* 33 (2015) 314–327, <https://doi.org/10.1016/j.devcel.2015.03.020>.
- [70] T. Okada, J. Ohzeki, M. Nakano, K. Yoda, W.R. Brinkley, V. Larionov, H. Masumoto, CENP-B controls centromere formation depending on the chromatin context, *Cell* 131 (2007) 1287–1300.
- [71] R. Contreras-Galindo, S. Fischer, A.K. Saha, J.D. Lundy, P.W. Cervantes, M. Mourad, C. Wang, B. Qian, M. Dai, F. Meng, A. Chinnaiyan, G.S. Omenn, M.H. Kaplan, D.M. Markovitz, Rapid molecular assays to study human centromere genomics, *Genome Res.* 27 (2017) 2040–2049, <https://doi.org/10.1101/gr.219709.116>.
- [72] N. Pironon, J. Puechberty, G. Roizès, Molecular and evolutionary characteristics of the fraction of human alpha satellite DNA associated with CENP-A at the centromeres of chromosomes 1, 5, 19, and 21, *BMC Genom.* 11 (2010) 195, <https://doi.org/10.1186/1471-2164-11-195>.