HHS Public Access

Author manuscript

Annu Rev Genet. Author manuscript; available in PMC 2022 October 10.

Published in final edited form as:

Annu Rev Genet. 2021 November 23; 55: 583–602. doi:10.1146/annurev-genet-071719-020519.

Variation and Evolution of Human Centromeres: A Field Guide and Perspective

Karen H. Miga[1,2], **Ivan A. Alexandrov**[3,4,5]

[1]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California 95064, USA

[2]Department of Biomolecular Engineering, University of California, Santa Cruz, California 95064, USA

[3]Department of Genomics and Human Genetics, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119991, Russia

[4]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University, Saint Petersburg 199004, Russia

[5]Research Center of Biotechnology of the Russian Academy of Sciences, Moscow 119071, Russia

Abstract

We are entering a new era in genomics where entire centromeric regions are accurately represented in human reference assemblies. Access to these high-resolution maps will enable new surveys of sequence and epigenetic variation in the population and offer new insight into satellite array genomics and centromere function. Here, we focus on the sequence organization and evolution of alpha satellites, which are credited as the genetic and genomic definition of human centromeres due to their interaction with inner kinetochore proteins and their importance in the development of human artificial chromosome assays. We provide an overview of alpha satellite repeat structure and array organization in the context of these high-quality reference data sets; discuss the emergence of variation-based surveys; and provide perspective on the role of this new source of genetic and epigenetic variation in the context of chromosome biology, genome instability, and human disease.

Keywords

centromere; satellite DNA; repeat; genome; variation; epigenetics

1. INTRODUCTION

Centromeres are essential chromosomal structures that mark sites of spindle attachment and ultimately ensure proper chromosome segregation during both mitosis and meiosis.

khmiga@ucsc.edu .

Errors in centromere establishment, inheritance, and maintenance through cell division can result in unequal partitioning of chromosomes and genome instability. Notwithstanding their important cellular function, the precise sequence organization of human centromeres was excluded from initial genome reference assemblies (39, 47, 105) and largely ignored by contemporary genetic and genomic studies over the past two decades. Centromeric regions, and associated pericentromeric heterochromatin, are commonly marked by the enrichment of long arrays of near-identical tandem repeats, or satellite DNAs (91). These highly repetitive sequences have been historically underrepresented due to inherent cloning and sequencing biases: instability in *Escherichia coli* during bacterial artificial chromosome (BAC)-based cloning, the regular occurrence of restriction sites used for cloning in the tandem repeats, or the potential toxicity of the cloned DNA (16, 75). Further, genome assembly methods failed to reliably represent centromeric regions in the past due to the inability to confidently span unique sites in the array that are necessary to predict the linear ordering of thousands of tandem repeats (80, 88). As a result, all human centromeric regions were marked as large gaps, representing megabase-sized placeholders, in our original human reference genomes (39, 47, 105).

Although missing from our initial reference maps, sequences in these regions were not unknown. Focused experimental studies across human centromeric sequences revealed that all normal human centromeric regions are defined at the sequence level by long arrays of alpha satellite DNA, formed by a diverse class of AT-rich tandem DNA repeats, or monomers (58, 59). Individual monomers are commonly arranged into larger, multimonomeric units, or higher-order repeats (HORs) (113), and are organized into one or more highly homogenized arrays at every human centromere. Focused experimental efforts to sample, clone, and directly sequence representative HORs from each centromeric region provided important insight into chromosome-specific subsets of alpha satellite (reviewed in 111), their phylogenetic classification into distinct suprachromosomal families (reviewed in 2), and initial expectations for long-range organization (55, 85, 109, 110). Further, focused studies of the small number of assembled satellites on the chromosome arms adjacent to the centromere gaps of human and nonhuman primate genomes revealed discrete and chronologically ordered alpha satellite layers (81, 85, 89). Emerging databases of alpha satellite–containing reads in whole-genome sequencing data released our first assessments of the frequency of repeat variation within chromosome-assigned arrays, along with early estimates of array length differences between individuals in the population (48, 68, 101). Linear representation of these observed repeat variants and their estimated copy number in the HuRef genome (51) led to the initial release of modeled alpha satellite arrays in the human reference assembly (GRCh38) (68, 104). Although these modeled alpha satellites were inadequate for long-range studies of array structure, they enabled short-read mapping to predict sequence variation (68), detected off-target mapping (66), and offered a more comprehensive study of sequences bound to inner kinetochore proteins (19, 72, 73). Collectively, these extensive studies in the human genome led to the development of the first conceptual representation of centromere genomic organization and sequence evolution across complex genomes.

Advancements in long-read sequencing technologies and recent improvements in repeat assembly methods can now generate complete and accurate assemblies of human

centromeric HOR arrays (15, 41, 52, 67, 74). This progress credits the availability of long reads (~15–20 kb) with extremely high consensus base quality (QV30, 99.9%), or high-fidelity (HiFi) sequencing data from Pacific Biosciences (108), and reads that routinely reach hundreds of kilobases in length (40), or ultralong (UL) data from nanopore sequencing from Oxford Nanopore Technologies. In parallel, we have seen tremendous gains in automated satellite array assembly and quality evaluation protocols (15, 69, 74), coupled with standard validation methods using pulsed-field gel electrophoresis (PFGE) and Southern blotting (52, 67).Notably, our current centromeric reference assemblies are derived from an effectively haploid human cell line derived from a complete hydatidiform mole [CHM13hTERT (94)], in which cells have two nearly identical pairs of chromosomes, greatly simplifying the challenge of repeat assembly compared with typical diploid cell lines. The recent release of the first complete assemblies of two human chromosomes end-to-end, or telomere-to-telomere (T2T) [T2T-ChrX (67) and T2T-Chr8 (52)], offered our first opportunity to evaluate these new alpha satellite assemblies in light of the expectations based on previous experimental studies (31, 55, 85) (discussed in more detail later in this review). Further, with the release of additional human centromeric regions (6), we are now met with an opportunity to blend the old with the new: confirming expectations in our original models and highlighting new discoveries with access to complete and accurate maps.

Centromeric satellite repeat copy number and sequence variants within each array are expected to vary considerably (54, 68, 109) due to unequal crossover and conversion. Therefore, a single haploid representation of each human centromeric region is inadequate to comprehensively study the extent of sequence and epigenetic variation. Indeed, satellite repeat copy number estimates across human diversity cohorts, such as the 1000 Genomes Collection (98), have shown that haploid X (DXZ1 or S3CXH1L) and Y (DYZ3 or S4CYH1L) alpha satellite array lengths can differ by a factor of 5–10 between individuals (48, 64, 68) and can be different in two homologous chromosomes from the same person (117). Previous cytogenetic studies have indicated that such variation may contribute to predispositions to cancer, infertility, and chromosomal aneuploidies (28, 116). In addition to centromere sequence diversity, inner kinetochore proteins that bind alpha satellite also show signs of evolving rapidly across primates (86). Further, a scan of the human genome for signatures of positive selection found evidence of recent selective sweeps at 8 of 17 studied centromeres (114), motivating future studies to explore evidence of centromere strength or drive in human population data, as previously documented in other species (10, 25, 45, 46). Ultimately, extensive analysis of the alpha satellite array variation in humans and nonhuman primates will offer new insights into how these regions evolve over time and how such changes influence the localization and inheritance patterns of inner kinetochore proteins.

This is an exciting time for centromere research. Tools are now available to do in-depth analyses of the intersection of genomics and epigenetics to explore variation in centromere structure and models of evolution. Comprehensive studies of the molecular mechanisms that ensure centromere activity will likely provide new insights into human health, and they have the potential to lead to new diagnostic tools and treatments. Additionally, a more complete understanding of centromeres at the genomic level will likely motivate the development of a new era in synthetic genome biology and gene therapy vectors for use in humans. Here, in

light of this great progress and promise, we discuss the structure of alpha satellite sequences and our current model of alpha satellite evolution, and we provide a perspective on new studies aimed to improve our understanding of centromere biology and human disease.

2. EVOLUTIONARY HIERARCHIES IN ALPHA SATELLITE ARRAY STRUCTURE

2.1. Genomic Model of Human Centromeric Regions

Alpha satellite DNAs are credited as the genetic substrate of endogenous centromeres in primates, starting with the new-world monkeys. No alpha satellites have been found in tarsiers and lemurs (49, 89). In humans, arrays of alpha satellites are organized in discrete layers expanding out from a multimegabase-sized homogeneous core that is composed of chromosome-specific HORs (live or active arrays). Additional subsets of alpha satellites (36, 65, 100) are often observed on one or both sides of the core in a near symmetrical formation. This includes a zone of smaller homogeneous HOR arrays (pseudocentromeres or inactive arrays) followed by an outermost layer of progressively more divergent and smaller (center-to-periphery gradient) HOR and monomeric arrays (relic centromeres). Both inactive HOR arrays and divergent arrays are often in the range of a few to hundreds of kilobases. Other distinct satellite classes, such as the classical human satellites (human satellites 1–3, or HSat1–HSat3), are of variable size (up to several megabases) and positioned in the adjacent pericentromeric regions. Segmental duplications are often observed directly flanking the satellite arrays or in centromeric transition regions extending out to the p-arm or q-arm (greater than a megabase) or between adjacent satellite arrays. The entire centromeric region can be defined by those sequences in linkage or sharing a common centromere-spanning haplotype (cenhap), which is characterized by repressed meiotic recombination (48). All alpha satellite arrays except for the active or live HOR arrays may by opposition also be called inactive or dead centromeres or dead centromeric layers.

The general symmetrical disposition of alpha satellite layers around the homogeneous core, which we noted above (see also Figure 1a), reflects the mode of alpha satellite evolution that may be called expanding centromere. It suggests the periodical emergence of a new centromere within the old one (Figure 1c).Analysis of sequence relationships between different HORs within suprachromosomal families (SFs) (see Section 2.4.1) and between dead monomeric layers has shown that centromere expansion likely goes in waves of interchromosomal transfer and amplification, where the HORs (or monomeric sequences) of the newly formed novel centromere jump from one live centromere to another and amplify in the new location to form the next generation of live centromeres (a centromeric layer) in all chromosomes or in a group of chromosomes (2, 4, 85, 89, 112). The remnants of the old centromere are displaced sideways, shrink, diverge, and structurally degrade (see Sections 2.2 to 2.9) (24, 84, 89).

The sequences in the alpha satellite part of a centromere can be characterized by their monomer composition (SF-specific monomer classes; see Section 2.2), their HOR versus monomeric construction, and an average divergence of neighboring copies of a repeat in an array, as there is a gradient of intra-array divergence from the center to the periphery that

reflects the age of alpha satellite arrays (2, 42, 84, 85, 89). Also crucial is the functional distinction between active (or live) arrays, which host the kinetochore, and inactive (or dead) arrays, which do not. None of these differences are absolute, and there are many exceptions and borderline cases, but they provide a reasonable way to navigate the centromere landscape. It is also important to note that active centromeres of primates in the human lineage pre-dating the apes were the same in all chromosomes and did not have HORs longer than dimers (panchromosomal organization; the centromere array on chromosome Y is often an exception). Chromosome-specific HORs (i.e., chromosome-specific organization) emerged in the great apes (5, 89). Gibbons are a border case, with evidence for both panchromosomal,or genome-wide, organization and chromosome-specific organization in different taxa (9, 18, 44). Hence, in humans, there are dead relic divergent layers that are the remnants of ancestral primate centromeres (2, 89). Their organization is panchromosomal in older monomeric layers and mostly chromosome-specific in younger divergent HOR layers and even younger homogeneous HOR layers (104).

2.2. Alpha Satellite Monomer Classification

Classification of alpha satellite monomers has been summarized in several reviews (2, 36, 63). There are five major alpha satellite SFs. An SF is a group of HOR or monomeric arrays more closely related to each other than to the other groups. Each SF is built of its own set of monomeric classes (see Supplemental Table 1; Section 2.8). The new SFs (SF1–SF3) exist only in African apes (89).They form active (live) centromeres on all human autosomes and the X chromosome (2, 78), most pseudocentromeres (or inactive arrays), and most divergent HORs. The old SF4+ and SF5 unite the dead monomeric layers as well as pseudocentromeres and divergent HOR arrays derived from them by more recent amplifications. SF5 represents centromeres that have been active at the time of the human-orangutan split (89).SF4+ is an umbrella group that unites a large number of old and ancient SFs, such as SF4 proper, SF6, SF7, and more, which correspond to the older primate groups (90). As a notable exception, the active centromere of chromosome Y also belongs to SF4 proper. SF5 is the immediate ancestor of the new SFs (78, 104). It consists of the two monomeric classes R1 and R2, which represent two progenitor types (B and A, respectively) to which all monomeric classes of the new families belong (3, 78). Importantly, the A- and B-type consensus monomers mostly differ from each other in a narrow 17-bp region (the AB-box), which corresponds to the binding site of a well-studied centromeric protein, CENP-B (the B-box, type B monomers) (60), or to the presumed binding site of a very-little-studied (30) pJα protein (the A-box, type A monomers). The relationship of types A and B and SF-specific monomeric classes is shown in Supplemental Table 1.

2.3. Archaic Suprachromosomal Families 01 and 02

Recently, Shepelev et al. (90) and Uralsky et al. (104) analyzed a group of the less abundant alpha satellite sequences detected as atypical or archaic representatives of SF1 and SF2. They were shown to be the interim stages of evolution from ancestral SF5 to typical or modern SF1 and SF2. We propose considering them full-fledged SFs, but for the time being, assigning them the incremental numbers 01 and 02, respectively, to avoid renumbering the other SFs. Their monomer classes should be processed in a standard way and included in the SF table (Supplemental Table 1). In the human genome, SF01 and 02 are represented

by both divergent (see Section 2.8) and homogeneous HORs, including one live centromere in chromosome 6 (D6Z1 or S01C6H1L), the 3-monomer archaic segment in the live HOR of chromosome 3 (D3Z1 or S01/1C3H1L), relatively large pseudocentromeric arrays in centromeres 3 (S01C3H2) and 20 [S02C20H3; was named HOR20–2 by Shepelev et al. (90)], and also large divergent HOR arrays in chromosomes 3 and 6 [S1C3/6H1d; name changed from S1C3H4 used by Uralsky et al. (104)]. Overall, archaic SF arrays are usually found between the new SFs and SF5 arrays (90). Sequence relationships within SF01 were studied in detail by Uralsky et al. (104).

2.4. The Hierarchies in Higher-Order Repeat Domains

Alpha satellite HORs present a complex hierarchy of sequences with different levels of identity between different HORs (coming from different chromosomes or within one chromosome) and different levels of intra-array divergence (2, 4, 6, 15, 36, 63, 65, 81, 90, 104, 111, 112). These levels include SFs, sub-SFs, sister HORs, homogeneous HORs, haplotypes of the same HOR, and divergent HORs, described in the sections that follow.

2.4.1. Suprachromosomal families.—SFs are groups of related HORs that share the same broad classes of monomers (Supplemental Table 1, Section 2.2) and reside on a number of chromosomes (Supplemental Table 2). The divergence between different HORs in one SF is ~12–15% and 20–50% between different SFs (2, 4, 89, 90).

2.4.2. Subsuprachromosomal families.—Sub-SFs are groups of even more closely related HORs within an SF (2, 104).Sub-SFs are known in all new SFs (shown in Supplemental Table 2; e.g., S1C1/5/19H1L, S1C5H2, and S1C16H1L in SF1). Divergence in such groups is ~7–10%.

2.4.3. Sister higher-order repeats.—Sister HORs (Supplemental Table 2) are distinct chromosome-specific sequence variants (major SqVs) within the same HOR that form smaller arrays adjacent to the live HOR [e.g., S3C17H1L (D17Z1), S3C17H1-B (D17Z1-B), and S3C17H1-C (D17Z1-C) (81, 89)] or pseudocentromeric subdomains in the pericentromere [e.g., S3C1H2-A, -B, -C, and -D (104) and S2C18H2-A, -B, -D, and -E (6)]. They are formed by monomers that differ only moderately from respective monomers of the other sister HORs (~3–7%) and may have the same or a somewhat different order of monomers.

2.4.4. Homogeneous higher-order repeats.—Homogeneous HORs (reviewed in 2,36,63) usually have an overall average divergence across the whole array of about 1–2% and are chromosome-specific with a few notable exceptions among the live HOR arrays (double and triple domains) (see Supplemental Table 2; Section 2.5).

2.4.5. Haplotypes of the same higher-order repeat.—Haplotypes of the same HOR (slight SqVs) occupy different regions in the live HOR arrays (6, 15, 52, 65) (see Figure 1b; Section 2.7). A haplotypic HOR region may be formed by one haplotype or by several alternating varieties. Divergence between HORs of different haplotypes may be ~1–3%, and divergence within one haplotype may be as low as 0.5% (see Figure 1b).

2.4.6. Divergent higher-order repeats.—Divergent HORs represent a separate entity that unites HORs that have passed completely or partially through the alleged hypermutability stage and accumulated more divergence than would be possible during their documented or estimated lifespan given the normal mutation rate (see Section 2.8). These are often partially ruined small arrays on the edges of larger homogeneous arrays, some chromosome-specific and some residing on two or several chromosomes. Intra-array divergence is typically over 10% (Figure 1a).

2.5. Homogeneous arrays of Higher-Order Repeats

Typical alpha satellite homogeneous HOR arrays consist of chromosome-specific HORs ~4–40 monomers long. However, some nonhomologous pairs of chromosomes share almost identical or very similar live HORs [the so-called paired domains 13/21 and 14/22 and triple domain 1/5/19; reviewed by Alexandrov et al. (2)]. It is not known if these chromosomes have recently shared the centromeres and did not have enough time to diverge or if there is a continuous homogenizing exchange between these chromosomes. This issue could be addressed using the T2T assembly. Some pseudocentromeric HORs are also shared between two or more chromosomes (e.g., S5C5/19H4 is shared by chromosomes 5 and 19; see other examples in Supplemental Table 2).

The traditional naming system for alpha satellite HORs was a part of the more general human gene mapping (HGM) system. It was not very specific or convenient, and many newly discovered HORs did not have HGM names (see discussion in 104). We therefore propose to use the new naming system designed especially for the alpha satellite HORs described by Uralsky et al. (104), which covers all currently known HORs (see proposed names in Supplemental Table 2) and is easier to operate. In this system, each HOR received a name that included its SF, chromosomal location and index number (e.g., S1C13/21H1 for SF1, chromosomes 13 and 21, and HOR#1). Divergent HORs are marked with the d index after the name (e.g., S1C3/6H1d). Live HORs are always H1 and are additionally marked with index L (e.g., S2C9H1L). This new system should be evaluated by satellite and bioinformatic communities to be modified and/or changed as needed. For the time being, we use the old names (whenever they are available) and new names in parallel. Note that no SFs older than SF6 have been found in HORs so far (Supplemental Table 2), new SF1–SF3 (01 and 02 included) are exclusively HOR, and SF5 and SF4 (proper) and SF6 have both HOR and non-HOR arrays (90, 104).

2.5.1. Sequences for CENP-B and pJα binding sites in alpha satellite.—The new SF1–SF3 form all of the live centromeres except for the Y and form most of the pseudocentromeric and relic inactive, or dead, HOR arrays (Supplemental Table 2). In SF1 and SF2 HORs, the J1 and J2 or D1 and D2 class monomers appear as internal J1J2 or D1D2 dimers, respectively, with perfect (SF1) or near-perfect (SF2) AB periodicity across arrays (Supplemental Table 1) (35, 78, 79, 90). In SF3 and SF5, the monomer classes (W1–W5 and R1R2, respectively) alternate in a more complex manner, and the AB pattern also does not have that simple regularity. Note that the presence of the A- or B-box in a monomer does not mean the presence of the actual pJα- or CENP-B-binding site. Boxes are just alternative configurations of the AB region that are permissive to respective sites

[35–51 bp of the monomer in our cyclic shift (see 78)]. For this review, we have examined the distribution of the actual sites in the T2T assembly (Figure 2). The actual pJα sites first appear in the Na (*green*) monomers of the dimeric OaNa (*olive-green*) dead layer, but not in the Oa (*olive*) ones (Supplemental Table 1). All of the later successive layers have originated from the green monomer layer only (6, 89), and the sites persist there. In SF5 (R1R2), the B-boxes and actual CENP-B sites first appear, and live centromeres start being formed by the AB satellite (Supplemental Table 1). All new SFs have both the A- and the B-type monomers, but in the human genome, only in SF2 do the pJα sites appear in significant numbers, while CENP-B sites are frequent and regular in all three SFs. Moreover, in SF2, the actual pJα sites are frequent only in some live HORs [e.g., S2C2H1L (D2Z1) and S2C8H1L (D8Z2)], and are virtually absent in many others [e.g., S2C9H1L (D9Z4), S2C14/22H1L, and S2C18H1L (D18Z1)]. Thus, there is possibly an evolutionary trend toward loss of the pJα sites, which may have been in effect since the appearance of the CENP-B sites. If true, it would suggest that both proteins have the same or overlapping functions in centromeres.

2.5.2. Structural variants of a higher-order repeat.—

All HORs have structural variants (StVs) that usually can be explained as in/dels of the whole monomers in the primary HOR. Monomer-by-monomer annotation of SF1 reference models in hg38 by Uralsky et al. (104) visualized StVs in HuRef HOR reference models and collected related statistics. Such annotation can now be performed in the T2T CHM13 assembly to collect actual genomic data. Also, the abundant presence of hybrid monomers where a part of one monomer of an HOR was fused to a part of the other was revealed in hg38. The approximate monomer length of ~171 bp is usually conserved in such hybrids. The presence of a hybrid in an StV is a variable feature that depends on a cyclic shift (a monomer start site) used for analysis (15, 21, 22, 104). Therefore, we advocate the use of one universal monomer start site and propose the use of the traditional first nucleotide in the *Bam*HI site of the chromosome X–specific live HOR, which was the first completely sequenced human HOR (106). This cyclic shift was used in about half of the alpha satellite papers over the years and in recently published annotation tools like PERCON, HumAS-HMMER HOR, and CentromereArchitect (22, 90, 104) and is also being used by the T2T consortium for centromere annotation.

Data from Uralsky et al. (104) obtained in hg38 alpha satellite reference models and studies of the first two assembled large centromeres (52, 67) suggest that different live centromeres vary greatly with respect to the abundance of StVs. Some chromosomes, such as X and 11, have non-polymorphic centromeres mainly composed of full-length HOR copies and have only dozens of copies of StVs per 1500–2000 HORs in a centromere. Other centromeres, like 8 and 10, are very polymorphic and have some very high-copy StVs that may exceed the full-length HOR in frequency. It is known that different individual chromosomes (and different persons) may also differ in content and distribution of StVs (1).

2.6. Pseudocentromeres and Centromeric Epialleles

Live HOR arrays organize the kinetochore in most individuals, and they are usually the largest HOR arrays in a given chromosome. However, in some individual chromosomes,

a smaller, technically pseudocentromeric HOR array may assume the role of kinetochore organizer instead of the main array and form a centromeric epiallele (1, 57; reviewed in 63). We propose that such occasionally functional HORs may be called half-alive or epi arrays, as opposed to the dead ones that are never functional. Then there are two slightly different theoretical possibilities (104): (*a*) half-alive centromeres that had once been live but have surrendered the main centromere status to a more efficient competitor and retained only occasional activity; and (*b*) half-alive pseudocentromeres, which are the HORs that have never been live centromeres but are recent amplifications of some dead alpha satellite sequences that occasionally assume centromeric activity.

2.7. Higher-Order Repeat Haplotypes

It has been known for a long time (e.g., 20) that the vast homogeneous core of a centromere formed by nearly identical HORs has some domains made by arrays of even more identical HORs, which share a number of characteristic mutations (a haplotype). Mutations in this case were defined as differences from the overall consensus HOR. Such haplotypes should be considered slight SqVs of a HOR (as opposed to major SqVs, which are sister HORs). Often, they differ not only in sequence but in structure as well, and in those cases they are also StVs. One example of these SqtVs that has been much studied recently is a 13-mer D17Z1 (S3C17H1L) HOR, which is a deleted variant of the complete 16-mer HOR and also differs from it by a number of characteristic mutations (1). In this work, the abundance of this variant HOR in the live arrays of some individual chromosomes 17 apparently prompted the kinetochore to choose an alternative location in the D17Z1-B sister array (1). However, before the complete assemblies of the whole centromeres became available, the data on haplotype patterns within the live arrays were limited. The first two large centromeres assembled by the T2T consortium revealed a considerable heterogeneity of HORs within the live arrays (15, 52, 65). Careful analysis of this heterogeneity (Figure 1b) reveals a phylogenetic tree of haplotypes, a semisymmetric pattern of layers, and a gradient of homogeneity reminiscent of the pattern of pericentromeric dead layers around the live centromere (84, 89). This suggests that the forces and mechanisms operating to create both patterns may partially be the same.

2.8. Divergent Alpha Satellite Arrays

Besides live homogeneous centromeres and pseudocentromeres, alpha satellites are found in two kinds of dead (inactive) divergent arrays (HOR and non-HOR), which may be called dead relic centromeres because they represent the actual remains of formerly active centromeres, once large and homogeneous but now small, divergent, and disorderly. These are dead monomeric layers (the remnants of panchromosomal organization of SF4+ and SF5 centromeres) and divergent HOR arrays (the remnants of chromosome-specific SF1–SF3 and SF5 centromeres and pseudocentromeres). SF5 is present in both divisions because it has both HOR and non-HOR components, and some HORs are divergent. In SF4 proper and SF6, both HOR and non-HOR arrays are observed as well, but all HORs found there so far are homogeneous (90). Both divergent compartments share the signatures of a hypermutability phenomenon that has been proposed as a theoretical explanation of their divergence patterns. It has been demonstrated that the intra-array divergence in dead monomeric layers (89) and in divergent HORs (104) far exceeds what could have been

accumulated during their lifespan with the normal mutation rate. For instance, in dead monomeric layers of centromere X from Ga [*yellow* (Figure 2; Supplemental Table 1)] to Aa [*gray* (Figure 2; Supplemental Table 1)], the divergence goes from 16% to 30%. Shepelev et al. (89) have speculated that hypermutability in freshly dead arrays is caused by replication problems like fork stalling, which induces error-prone DNA polymerases. The above hypermutability hypothesis is based on the assumption that these arrays were once homogeneous with a divergence not exceeding 1–2%, but a burst of mutations occurred to yield a divergence of >10%. Indeed, there is typically a large gap in intra-array identity between homogeneous (divergence 1–2%) and divergent (>10%) compartments. This could be explained in a traditional way by a special recombination process called homogenization, which is supposed to maintain the large size and high identity in the live arrays. The presumed mechanisms of homogenization are gene conversion (83) and mitotic unequal crossover (92), as meiotic crossover is repressed at centromeres (48, 93). It is obvious, however, that when a new centromere appears in the middle of the old one, as stipulated by the expanding centromere scenario, homogenization stops in the now freshly dead domains, which are displaced to the flanks, and they gradually progress to typical dead centromeres, shrunken, divergent, and disorderly (89). If all of this is true, the time since the centromere has died and homogenization has stopped is the interval in which the array has to accumulate its current intra-array divergence (in excess of ~2% or less, which it had as a live array). However, it is known that the long-dead arrays accumulate mutations at a normal rate (81, 89). It follows that the accumulation is nonlinear, and the freshly dead arrays must get many more mutations than normal. Shepelev et al. (89) calculated that the excessive divergence gained by freshly dead arrays during the hypermutability period is about 10%, after which the mutation rate subsides. Note that the age of the arrays could be graded by two alpha satellite–dependent (phylogeny of monomers and divergence) and by two alpha satellite–independent ways (89). One of the latter is the presence of the orthologs or paralogs of a given array in extant primate taxa, the age of which is known (34); another is the age of L1 repeats [also known (43)], which often insert into the dead arrays and are very rarely found in the live ones (42).

2.9. Conclusions and Evolutionary Models

Sequence mapping shows symmetrical layers of distinct alpha satellite families around and within a homogeneous core, centered at the youngest haplotype(s) in the live array, with the age of layers increasing from the center to periphery (2, 15, 52, 65, 84, 85, 89). Divergence data (Figure 1b; see Sections 2.7 and 2.8) suggest the discontinuous gradient of divergence throughout all the layers, with a minimum at the same youngest haplotype(s) and a steep increase at the transition from homogeneous to divergent compartments (Figure 1a), which can be explained by hypermutability in the freshly dead arrays, presumably caused by induction of error-prone DNA synthesis. Additionally, the degree of structural disorder (a number of deletions, inversions, transposable element (TE) insertions, and HSat expansions) is minimal in homogeneous arrays and much higher in divergent arrays. All of this makes up the signature pattern of an expanding centromere. This pattern suggests a stochastic generation of a new centromeric array inside an existing centromere and lateral displacement of the dead remnants of the old centromere. Through interchromosomal exchange (meant as singular one-way events here), the new repeat spreads to all (or a group of) chromosomes

within a short period of time. Such waves of change occur in a regular manner throughout phylogenetic history and create a multilayer centromere, which records its own and its species' history, similar to archeological layers under a city (89).

Two models may be used to interpret this layout. The neutral homogenization model that dominates so far features stochastic homogenization of neutral mutations, some of which may achieve fixation in all repeats of an array and thus provide for the evolution of an array or the concerted evolution of a number of arrays given sufficient exchange (meant as a continuous two-way process here) between them (92, 95, 96). The next-generation model, which we here term kinetochore selection (Figure 1c), would provide for much faster evolution. This model proposes that (*a*) the evolution of centromeric repeats is not entirely neutral, and they are selected by the affinity to a kinetochore, which is free to move and chooses the most favorable place to reside within the live array; (*b*) this selection operates through the ability of a kinetochore to amplify and possibly homogenize the repeats on which it resides (2, 89); and (*c*) the old centromere abandoned by the kinetochore degrades (deletions, inversions, TE insertions, HSat expansions, and hypermutability). It implies that intense amplification/homogenization is not an intrinsic property of any large array of homogeneous tandem repeats but is dependent on the presence of special machinery, which, in the case of centromeric repeats, is associated with a kinetochore. Hence, the term kinetochore-associated recombination machine or KARM was previously proposed (89). The models are not mutually exclusive, as a mutation or a haplotype favored by a kinetochore probably needs to rise to a certain copy number to compete as a centromere, which may be achieved in a stochastic manner. It also seems that kinetochore selection is entirely compatible with the centromere drive model (56), as both assign a major role to some kind of selfish selection (kinetochore selection or meiotic drive). Presumably, the kinds of selfish selection may be more than one and may be combined easily to better explain the coevolution of centromeric DNA repeats and proteins (56). A somewhat different model for selfish selection in the centromeres was recently proposed by Rice (77). We conclude that the whole process of homogenization has to be rethought as not entirely neutral but as a combination of neutral and selective forces.

3. SURVEYS OF GENETIC AND EPIGENETIC VARIATION AT HUMAN CENTROMERES

Centromeric alpha satellite arrays are rich in genetic and epigenetic diversity and present a new and uncharted genomic landscape to catalog structural variation in the human population. Variation in array structure could broaden studies aimed at understanding missing heritability and provide new insight into the genetic basis of complex and rare disorders. Although genome-wide association studies omit centromeric satellite sequences, studies of variants directly adjacent to human centromere arrays, or within centromere-spanning haplotypes, have been observed in a broad number of clinical studies (reviewed in 64), with notable examples in studies of mosaic chromosomal alterations in clonal hematopoiesis (53) and increased risk of multiple sclerosis (76). Efforts to expand our variant maps in centromeric regions are challenging, even with the release of high-quality reference maps, and will require new method development to confidently identify, describe,

and test candidate disease causal variants predicted in satellite DNAs. Further, our understanding of disease-associated variants will need to be evaluated in the context of background sampling estimates across the population. We currently do not understand how quickly these sites evolve in the human population, across multigenerational pedigree data, or across a population of single cells. Such fundamental baselines of satellite variation in healthy populations will be critical to confidently identify genetic features associated with disease.

Efforts to measure and report centromere sequencing variation will need to monitor more than the nonamplified mutations that are present in just one copy or few copies. Much of the variation within satellite arrays will be represented as copy number variations, or expansions and contractions of repeat variants, which can give rise to a haplotype (large-scale amplification) or subhaplotype (small-scale amplification). The emergence of more complete T2T genomes will present a new opportunity for method development to predict comprehensive satellite sequence variation by mapping short- and long-read data sets. Previous analyses have demonstrated the use of array assignment of short-read data sets to monitor repeat expansions and contractions through k-mer-assigned frequencies of select satellites (7, 26, 68, 107). These evaluations often report repeat variant information from pooled diploid chromosomes, in which it is not possible to determine copy number variation of the same k-mer present in unequal copies across the two homologous chromosomes without the use of pedigree information or orthogonal phased data sets. High-quality long-read data have been useful in predicting variation in repeat structure (e.g., HOR rearrangements, inversions, transposition, and single-nucleotide variants) as well as in copy number estimates (21, 67, 87). The use of long-read data in satellite DNA variant prediction and discovery is often challenged by inherent biases in sequencing coverage and, in extreme cases, sequencing of only one strand (17, 27), which can influence downstream variant prediction and interpretation.

Ideally, as we reach larger cohorts of completely assembled and properly phased T2T diploid genomes, direct array-to-array comparisons will be possible, allowing direct comparisons of the total length of the array, shared haplotype and subhaplotype repeat expansions, and rare repeat sequences that are not shared between individuals. Such assembly-based comparisons rely on the use of highly accurate sequences to ensure that conclusions are not influenced by introduced assembly error. Ultimately, it is likely that efforts to characterize satellite array variation will need to make a comprehensive assessment of variants to test if features within each array structure (notably, this is a broader definition of a locus, rather than one single-nucleotide polymorphism) are associated with disease.

4. NEW PERSPECTIVE ON CENTROMERE GENOMIC STRUCTURE AND FUNCTION

Access to highly accurate reference maps will offer new insight into the range of genomic structure compatible with centromere function. These tools are useful to ensure that confident and precise mapping of short- and long-read functional data sets will promote studies of the positioning of inner centromere proteins and, ultimately, of how variation at

the epigenetic level influences chromosome stability during cell division. These emerging mechanistic studies will require a broad, multidimensional view of epigenetics, replication, transcription, and recombination across human centromeric regions.

Although meiotic recombination is suppressed in centromeric regions (12, 55, 62), aligned with the observation of large cenhaps (48), other types of recombination are prolific, leading to repeat amplification, deletions, and inversions. This introduced genetic variation within and between alpha satellite arrays has been shown to influence centromere activity (11, 32). Notably, this has been demonstrated in studies of centromeric epialleles on chromosome 17 (D17Z1 or S3C17H1L; D17Z1B or S3C17H1-B), where HOR size and sequence variants were important features in establishing whether HOR arrays are competent for centromere formation (1, 57). Further, studies of chromosome-specific aneuploidies provide evidence that array composition or particular HOR sequence features [such as the frequency and abundance of CENP-B motifs (61)], rather than the overall array length, influence chromosome segregation fidelity during cell division (19). Studies of the assembled centromeric regions of chromosomes X [DXZ1 or S3CXH1L (65, 67)] and 8 [D8Z2 or S2C8H1L (52)] revealed distinct haplotype blocks where a set of shared HOR variants are localized within the array (52, 67). Careful annotation of entire assembled arrays reveals an uneven distribution of CENP-B motifs across a given array and perhaps indicates collections of repeat units within the array that are less competent for the maintenance of human centromeres (37, 65). Notably, these T2T studies present a snapshot of the precise linear arrangement of HORs within a single individual, and additional studies are critically needed to ensure that we have a more comprehensive understanding of centromere array haplotype blocks within the human population. These data may provide a better genomic context for future centromere genomic studies than general estimates of total array size. That is, the expansion and contraction of variants within specific haplotype arrays provide new insight into segregation fidelity and centromere sequence competency. Indeed, we may find that epialleles exist within a single multimegabase-sized HOR array, and that perhaps the distance between these CENP-A-bound sequences, as shown for other dicentric chromosome models (97, 99), may contribute to our understanding of centromere strength (46).

Focusing exclusively on the enrichment patterns of inner kinetochore proteins with the underlying alpha satellite DNAs may provide an incomplete picture of the epigenetic determinants of centromere identity and function. Kinetochore assembly is observed over a small proportion of the array, with flanking regions enriched in pericentric heterochromatin and CpG methylation (29, 82). The dynamics between centrochromatin domains (13) and pericentromeric heterochromatin may have broad influence, from spatial localization in the interphase nucleus (8), the formation of three-dimensional structure during mitosis (13), and low transcription and increase in chromosomal passenger complex occupancy (71). Therefore, the use of the new alpha satellite assemblies will provide a unique opportunity for the comprehensive analysis of centromere biology and cellular function throughout different stages of cell division and in early development, where the sites and size of the kinetochore are first established (18, 65, 66). Variation at the sequence level could influence the rates and fidelity of alpha satellite replication (23). Further, epigenetic variation in centromere protein deposition could influence array stability (33) and improve our understanding of

inner kinetochore protein inheritance and maintenance over time (73). Alpha satellite RNA transcripts have been associated with centromere function (20, 59, 60), proximity of the nucleolus (14), and genome instability (102, 118). We now have the ability to precisely map these transcripts to the genome and study the association of nearby transcription factors and bound polymerases (38). The structure and function of these highly repetitive regions of our genome present a large, unexplored genomic and epigenomic landscape. We are now faced with the challenge of closing the gaps in not only our genetic maps but also our epigenetic maps of these regions. Doing so will rely on new innovations in a number of long-read technologies to ensure the comprehensive assessment of methylation (67), replication (70), open chromatin (50), spatial maps (103), and long-read transcriptional data (115) from human centromeric satellite arrays. Future studies of epigenetic regulation of alpha satellites in early development, aging, and disease are expected to lead to a new era of discovery in centromere biology and function. Ultimately, access to alpha satellite assemblies will drive new high-resolution studies of basic cellular processes and regulation at human centromeres and has the potential to improve our understanding of human disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

I.A.A. was partially supported by Saint Petersburg State University (grant ID PURE 73023573) and by the Ministry of Science and Higher Education of the Russian Federation. K.H.M. is supported by National Institutes of Health/National Human Genome Research Institute grants U01HG010971, 1R21HG010548-01, and 1R01HG011274.

DISCLOSURE STATEMENT

K.H.M. has received travel funds to speak at symposia organized by Oxford Nanopore, and contributes to the scientific advisory board of Centaura.

LITERATURE CITED

1. Aldrup-MacDonald ME, Kuo ME, Sullivan LL, Chew K, Sullivan BA. 2016. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. Genome Res. 26(10):1301–11 [PubMed: 27510565]
2. Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y. 2001. Alpha-satellite DNA of primates: old and new families. Chromosoma 110(4):253–66 [PubMed: 11534817]
3. Alexandrov IA, Medvedev LI, Mashkova TD, Kisselev LL, Romanova LY, Yurov YB.1993.Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. Nucleic Acids Res. 21(9):2209–15 [PubMed: 8502563]
4. Alexandrov IA, Mitkevich SP, Yurov YB. 1988. The phylogeny of human chromosome specific alpha satellites. Chromosoma 96(6):443–53 [PubMed: 3219915]
5. Alkan C, Ventura M, Archidiacono N, Rocchi M, Sahinalp SC, Eichler EE. 2007. Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. PLOS Comput. Biol 3(9):1807–18 [PubMed: 17907796]
6. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, et al. 2021. Complete genomic and epigenetic maps of human centromeres. bioRxiv 2021.07.12.452052. 10.1101/2021.07.12.452052
7. Altemose N, Miga KH, Maggioni M, Willard HF. 2014. Genomic characterization of large heterochromatic gaps in the human genome assembly. PLOS Comput. Biol 10(5):e1003628 [PubMed: 24831296]

8. Andronov L, Ouararhni K, Stoll I, Klaholz BP, Hamiche A. 2019. CENP-A nucleosome clusters form rosette-like structures around HJURP during G1. Nat. Commun 10(1):4436 [PubMed: 31570711]
9. Baicharoen S, Arsaithamkul V, Hirai Y, Hara T, Koga A, Hirai H. 2012. In situ hybridization analysis of gibbon chromosomes suggests that amplification of alpha satellite DNA in the telomere region is confined to two of the four genera. Genome 55(11):809–12 [PubMed: 23199575]
10. Balzano E, Giunta S. 2020. Centromeres under pressure: evolutionary innovation in conflict with conserved function. Genes 11(8):912
11. Barra V, Fachinetti D. 2018. The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. Nat. Commun 9(1):4340 [PubMed: 30337534]
12. Beadle GW. 1932. A possible influence of the spindle fibre on crossing-over in Drosophila. PNAS 18(2):160–65 [PubMed: 16577442]
13. Blower MD, Sullivan BA, Karpen GH. 2002. Conserved organization of centromeric chromatin in flies and humans. Dev. Cell 2(3):319–30 [PubMed: 11879637]
14. Bury L, Moodie B, Ly J, McKay LS, Miga KHH, Cheeseman IM. 2020. Alpha-satellite RNA transcripts are repressed by centromere-nucleolus associations. eLife 9:e59770 [PubMed: 33174837]
15. Bzikadze AV,Pevzner PA.2020.Automated assembly of centromeres from ultra-long error-prone reads. Nat. Biotechnol 38(11):1309–16 [PubMed: 32665660]
16. Carlson M, Brutlag D. 1977. Cloning and characterization of a complex satellite DNA from *Drosophila melanogaster*. Cell 11(2):371–81 [PubMed: 408008]
17. Cechova M. 2020. Probably correct: rescuing repeats with short and long reads. Genes 12(1):48 [PubMed: 33396198]
18. Cellamare A, Catacchio CR, Alkan C, Giannuzzi G, Antonacci F, et al. 2009. New insights into centromere organization and evolution from the white-cheeked gibbon and marmoset. Mol. Biol. Evol 26(8):1889–900 [PubMed: 19429672]
19. Dumont M, Gamba R, Gestraud P, Klaasen S, Worrall JT, et al. 2020. Human chromosome-specific aneuploidy is influenced by DNA-dependent centromeric features. EMBO J. 39(2):e102924 [PubMed: 31750958]
20. Durfy SJ, Willard HF. 1989. 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(4):810–21 [PubMed: 2591964]
21. Dvorkina T, Bzikadze AV, Pevzner PA. 2020. The string decomposition problem and its applications to centromere analysis and assembly. Bioinformatics 36(Suppl. 1):i93–101 [PubMed: 32657390]
22. Dvorkina T, Kunyavskaya O, Bzikadze AV, Alexandrov I, Pevzner PA. 2021. CentromereArchitect: inference and analysis of the architecture of centromeres. Bioinformatics 37(Suppl. 1):i196–204 [PubMed: 34252949]
23. Erliandri I, Fu H, Nakano M, Kim J-H, Miga KH, et al. 2014. Replication of alpha-satellite DNA arrays in endogenous human centromeric regions and in human artificial chromosome. Nucleic Acids Res. 42(18):11502–16 [PubMed: 25228468]
24. Finelli P, Antonacci R, Marzella R, Lonoce A, Archidiacono N, Rocchi M.1996.Structural organization of multiple alphoid subsets coexisting on human chromosomes 1, 4, 5, 7, 9, 15, 18, and 19. Genomics 38(3):325–30 [PubMed: 8975709]
25. Fishman L, Saunders A. 2008. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. Science 322(5907):1559–62 [PubMed: 19056989]
26. Flynn JM, Brown EJ, Clark AG. 2021. Copy number evolution in simple and complex tandem repeats across the C57BL/6 and C57BL/10 inbred mouse lines.G3 Genes Genomes Genet.11(8):jkab184.10.1093/g3journal/jkab184
27. Flynn JM, Long M, Wing RA, Clark AG. 2020. Evolutionary dynamics of abundant 7-bp satellites in the genome of *Drosophila virilis*. Mol. Biol. Evol 37(5):1362–75 [PubMed: 31960929]
28. Ford JH, Lester P. 1978. Chromosomal variants and nondisjunction. Cytogenet. Cell Genet 21(5):300–3 [PubMed: 679729]

29. Fukagawa T, Earnshaw WC.2014. The centromere: chromatin foundation for the kinetochore machinery. Dev. Cell 30(5):496–508 [PubMed: 25203206]
30. Gaff C, du Sart D, Kalitsis P, Iannello R, Nagy A, Choo KHA. 1994. A novel nuclear protein binds centromeric alpha satellite DNA. Hum. Mol. Genet 3(5):711–16 [PubMed: 8081356]
31. Ge Y, Wagner MJ, Siciliano M, Wells DE.1992.Sequence,higher order repeat structure, and long-range organization of alpha satellite DNA specific to human chromosome 8. Genomics 13(3):585–93 [PubMed: 1639387]
32. Giunta S, Funabiki H. 2017. Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. PNAS 114(8):1928–33 [PubMed: 28167779]
33. Giunta S, Hervé S, White RR, Wilhelm T, Dumont M, et al. 2021. CENP-A chromatin prevents replication stress at centromeres to avoid structural aneuploidy. PNAS 118(10):e2015634118 [PubMed: 33653953]
34. Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, et al. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol 9(3):585–98 [PubMed: 9668008]
35. Haaf T, Mater AG, Wienberg J, Ward DC. 1995. Presence and abundance of CENP-B box sequencesin great ape subsets of primate-specific α-satellite DNA. J. Mol. Evol 41(4):487–91 [PubMed: 7563136]
36. Hartley G, O'Neill RJ. 2019. Centromere repeats: hidden gems of the genome. Genes 10(3):223
37. Hoffmann S, Izquierdo HM, Gamba R, Chardon F, Dumont M, et al. 2020. A genetic memory initiates the epigenetic loop necessary to preserve centromere position. EMBO J. 39(20):e105505 [PubMed: 32945564]
38. Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, et al. 2021. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. bioRxiv 2021.07.12.451456. 10.1101/2021.07.12.451456
39. Int. Hum. Genome Seq. Consort. 2004. Finishing the euchromatic sequence of the human genome. Nature 431(7011):931–45 [PubMed: 15496913]
40. Jain M, Koren S, Miga KH, Quick J, Rand AC, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol 36(4):338–45 [PubMed: 29431738]
41. Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, et al. 2018. Linear assembly of a human centromere on the Y chromosome. Nat. Biotechnol 36(4):321–23 [PubMed: 29553574]
42. Kazakov AE, Shepelev VA, Tumeneva IG, Alexandrov AA, Yurov YB, Alexandrov IA. 2003. Interspersed repeats are found predominantly in the "old" α satellite families. Genomics 82(6):619–27 [PubMed: 14611803]
43. Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16(1):78–87 [PubMed: 16344559]
44. Koga A, Hirai Y, Hara T, Hirai H. 2012. Repetitive sequences originating from the centromere constitute large-scale heterochromatin in the telomere region in the siamang, a small ape. Heredity 109(3):180–87 [PubMed: 22669075]
45. Kursel LE, Malik HS.2018.The cellular mechanisms and consequences of centromere drive. Curr. Opin. Cell Biol 52:58–65 [PubMed: 29454259]
46. Lampson MA, Black BE. 2017. Cellular and molecular mechanisms of centromere drive. Cold Spring Harb. Symp. Quant. Biol 82:249–57 [PubMed: 29440567]
47. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921 [PubMed: 11237011]
48. Langley SA, Miga KH, Karpen GH, Langley CH. 2019. Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. eLife 8:e42989 [PubMed: 31237235]
49. Lee H-R, Hayden KE, Willard HF. 2011. Organization and molecular evolution of CENP-A–associated satellite DNA families in a basal primate genome. Genome Biol. Evol 3:1136–49 [PubMed: 21828373]
50. Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, et al. 2020. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat. Methods 17(12):1191–99 [PubMed: 33230324]

51. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. 2007. The diploid genome sequence of an individual human. PLOS Biol. 5(10):e254 [PubMed: 17803354]
52. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, et al. 2021. The structure, function and evolution of a complete human chromosome 8. Nature 593(7857):101–7 [PubMed: 33828295]
53. Loh P-R, Genovese G, Handsaker RE, Finucane HK, Reshef YA, et al. 2018. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature 559(7714):350–55 [PubMed: 29995854]
54. Mahtani MM, Willard HF. 1990. Pulsed-field gel analysis of α-satellite DNA at the human X chromosome centromere: high-frequency polymorphisms and array size estimate. Genomics 7(4):607–13 [PubMed: 1974881]
55. Mahtani MM, Willard HF. 1998. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. Genome Res. 8(2):100–10 [PubMed: 9477338]
56. Malik HS, Henikoff S. 2001. Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics 157(3):1293–98 [PubMed: 11238413]
57. Maloney KA, Sullivan LL, Matheny JE, Strome ED, Merrett SL, et al. 2012. Functional epialleles at an endogenous human centromere. PNAS 109(34):13704–9 [PubMed: 22847449]
58. Manuelidis L. 1976. Repeating restriction fragments of human DNA. Nucleic Acids Res. 3(11):3063–76 [PubMed: 794832]
59. Manuelidis L, Wu JC. 1978. Homology between human and simian repeated DNA. Nature 276(5683):92–94 [PubMed: 105293]
60. Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T.1989.A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J. Cell Biol 109(5):1963–73 [PubMed: 2808515]
61. Masumoto H, Yoda K, Ikeno M, Kitagawa K, Muro Y, Okazaki T. 1993. Properties of CENP-B and its target sequence in a satellite DNA. In Chromosome Segregation and Aneuploidy, ed. Vig BK, pp. 31–43. Berlin: Springer
62. Mather K. 1939. Crossing over and heterochromatin in the X chromosome of *Drosophila melanogaster*. Genetics 24(3):413–35 [PubMed: 17246931]
63. McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the genome. Chromosome Res. 26(3):115–38 [PubMed: 29974361]
64. Miga KH. 2019. Centromeric satellite DNAs: hidden sequence variation in the human population. Genes 10(5):352
65. Miga KH. 2020. Centromere studies in the era of “telomere-to-telomere” genomics. Exp. Cell Res 394(2):112127
66. Miga KH, Eisenhart C, Kent WJ. 2015.Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. Nucleic Acids Res. 43(20):e133 [PubMed: 26163063]
67. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. Nature 585(7823):79–84 [PubMed: 32663838]
68. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res. 24(4):697–707 [PubMed: 24501022]
69. Mikheenko A, Bzikadze AV, Gurevich A, Miga KH, Pevzner PA. 2020. TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. Bioinformatics 36(Suppl. 1):i75–83 [PubMed: 32657355]
70. Müller CA, Boemo MA, Spingardi P, Kessler BM, Kriaucionis S, et al. 2019. Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. Nat. Methods 16(5):429–36 [PubMed: 31011185]
71. Murillo-Pineda M, Valente LP, Dumont M, Mata JF, Fachinetti D, Jansen LET. 2021. Induction of spontaneous human neocentromere formation and long-term maturation. J. Cell Biol 220(3):e202007210 [PubMed: 33443568]
72. Nechemia-Arbely Y, Fachinetti D, Miga KH, Sekulic N, Soni GV, et al. 2017. Human centromeric CENP-A chromatin is a homotypic, octameric nucleosome at all cell cycle points. J. Cell Biol216(3):607–21 [PubMed: 28235947]

73. Nechemia-Arbely Y, Miga KH, Shoshani O, Aslanian A, McMahon MA, et al. 2019. DNA replication acts as an error correction mechanism to maintain centromere identity by restricting CENP-A to centromeres. Nat. Cell Biol 21(6):743–54 [PubMed: 31160708]
74. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, et al. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res.30(9):1291–305 [PubMed: 32801147]
75. Osoegawa K, Vessere GM, Li Shu C, Hoskins RA, Abad JP, et al. 2007. BAC clones generated from sheared DNA. Genomics 89(2):291–99 [PubMed: 17098394]
76. Reich D, Patterson N, De Jager PL, McDonald GJ, Waliszewska A, et al. 2005. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. Nat. Genet 37(10):1113–18 [PubMed: 16186815]
77. Rice WR. 2019. A game of thrones at human centromeres II. A new molecular/evolutionary model. bioRxiv 731471. 10.1101/731471
78. Romanova LY, Deriagin GV, Mashkova TD, Tumeneva IG, Mushegian AR, et al. 1996. Evidence for selection in evolution of alpha satellite DNA: the central role of CENP-B/pJα binding region. J. Mol. Biol 261(3):334–40 [PubMed: 8780776]
79. Rosandić M, Paar V, Basar I, Glunčić M, Pavin N, Pilaš I. 2006. CENP-B box and pJα sequence distribution in human alpha satellite higher-order repeats (HOR). Chromosome Res. 14(7):735–53 [PubMed: 17115329]
80. Rudd MK, Willard HF. 2004.Analysis of the centromeric regions of the human genome assembly. Trends Genet. 20(11):529–33 [PubMed: 15475110]
81. Rudd MK, Wray GA, Willard HF.2006. The evolutionary dynamics of α-satellite. Genome Res. 16(1):88–96 [PubMed: 16344556]
82. Scelfo A, Fachinetti D. 2019. Keeping the centromere under control: a promising role for DNA methylation. Cells 8(8):e202007210
83. Schindelhauer D, Schwarz T. 2002. Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous α-satellite DNA array. Genome Res. 12(12):1815–26 [PubMed: 12466285]
84. Schueler MG, Dunn JM, Bird CP, Ross MT, Viggiano L, et al. 2005. Progressive proximal expansion of the primate X chromosome centromere. PNAS 102(30):10563–68 [PubMed: 16030148]
85. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF.2001.Genomicandgenetic definition of a functional human centromere. Science 294(5540):109–15 [PubMed: 11588252]
86. Schueler MG, Swanson W, Thomas PJ, NISC Comp. Seq. Program, Green ED. 2010. Adaptive evolution of foundation kinetochore proteins in primates. Mol. Biol. Evol 27(7):1585–97 [PubMed: 20142441]
87. Sevim V, Bashir A, Chin C-S, Miga KH. 2016. Alpha-CENTAURI: assessing novel centromeric repeat sequence variation with long read sequencing. Bioinformatics 32(13):1921–24 [PubMed: 27153570]
88. She X, Horvath JE, Jiang Z, Liu G, Furey TS, et al. 2004. The structure and evolution of centromeric transition regions within the human genome. Nature 430(7002):857–64 [PubMed: 15318213]
89. Shepelev VA, Alexandrov AA, Yurov YB, Alexandrov IA. 2009. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. PLOS Genet. 5(9):e1000641 [PubMed: 19749981]
90. Shepelev VA, Uralsky LI, Alexandrov AA, Yurov YB, Rogaev EI, Alexandrov IA. 2015. Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. Genom. Data 5:139–46 [PubMed: 26167452]
91. Singer MF. 1982. Highly repeated sequences in mammalian genomes. Int. Rev. Cytol 76:67–112 [PubMed: 6749748]
92. Smith GP. 1976. Evolution of repeated DNA sequences by unequal crossover. Science 191(4227):528–35 [PubMed: 1251186]

93. Smith GR, Nambiar M. 2020. New solutions to old problems: molecular mechanisms of meiotic crossover control. Trends Genet. 36(5):337–46 [PubMed: 32294414]
94. Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res. 24(12):2066–76 [PubMed: 25373144]
95. Stephan W. 1989. Tandem-repetitive noncoding DNA: forms and forces. Mol. Biol. Evol 6(2):198–212 [PubMed: 2716519]
96. Stephan W, Cho S. 1994. Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. Genetics 136(1):333–41 [PubMed: 8138169]
97. Stimpson KM, Matheny JE, Sullivan BA. 2012. Dicentric chromosomes: unique models to study centromere function and inactivation. Chromosome Res. 20(5):595–605 [PubMed: 22801777]
98. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526(7571):75–81 [PubMed: 26432246]
99. Sullivan BA, Willard HF. 1998. Stable dicentric X chromosomes with two functional centromeres. Nat. Genet 20(3):227–28 [PubMed: 9806536]
100. Sullivan LL, Sullivan BA. 2020. Genomic and functional variation of human centromeres. Exp. Cell Res 389(2):111896
101. Suzuki Y, Myers EW, Morishita S. 2020. Rapid and ongoing evolution of repetitive sequence structures in human centromeres. Sci. Adv 6(50):eabd9230
102. Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, et al. 2011. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science 331(6017):593–96 [PubMed: 21233348]
103. Ulahannan N, Pendleton M, Deshpande A, Schwenk S, Behr JM, et al. 2019. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. bioRxiv 833590. 10.1101/833590
104. Uralsky LI, Shepelev VA, Alexandrov AA, Yurov YB, Rogaev EI, Alexandrov IA. 2019. Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome assembly. Data Brief 24:103708
105. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. Science 291(5507):1304–51 [PubMed: 11181995]
106. Waye JS, Willard HF. 1985. Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. Nucleic Acids Res. 13(8):2731–43 [PubMed: 2987865]
107. Wei KH-C, Grenier JK, Barbash DA, Clark AG. 2014. Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. PNAS 111(52):18793–98 [PubMed: 25512552]
108. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol 37(10):1155–62 [PubMed: 31406327]
109. Wevrick R, Willard HF. 1989. Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. PNAS 86(23):9394–98 [PubMed: 2594775]
110. Wevrick R, Willard HF. 1991. Physical map of the centromeric region of human chromosome 7: relationship between two distinct alpha satellite arrays. Nucleic Acids Res. 19(9):2295–2301 [PubMed: 2041770]
111. Willard HF. 1985. Chromosome-specific organization of human alpha satellite DNA. Am. J. Hum. Genet 37(3):524–32 [PubMed: 2988334]
112. Willard HF, Waye JS. 1987. 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(3):207–14 [PubMed: 2822935]
113. Willard HF, Waye JS. 1987. Hierarchical order in chromosome-specific human alpha satellite DNA. Trends Genet. 3:192–98
114. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. PLOS Genet. 3(6):e90 [PubMed: 17542651]

115. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, et al. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods 16(12):1297–305 [PubMed: 31740818]

116. Wyandt HE, Tonk VS. 2012. Human Chromosome Variation: Heteromorphism and Polymorphism. Dordrecht, Neth.: Springer

117. Yurov YB, Mitkevich SP, Alexandrov IA. 1987. Application of cloned satellite DNA sequences to molecular-cytogenetic analysis of constitutive heterochromatin heteromorphisms in man. Hum. Genet 76(2):157–64 [PubMed: 3475246]

118. Zhu Q, Pao GM, Huynh AM, Suh H, Tonnu N, et al. 2011. BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. Nature 477(7363):179–84 [PubMed: 21901007]

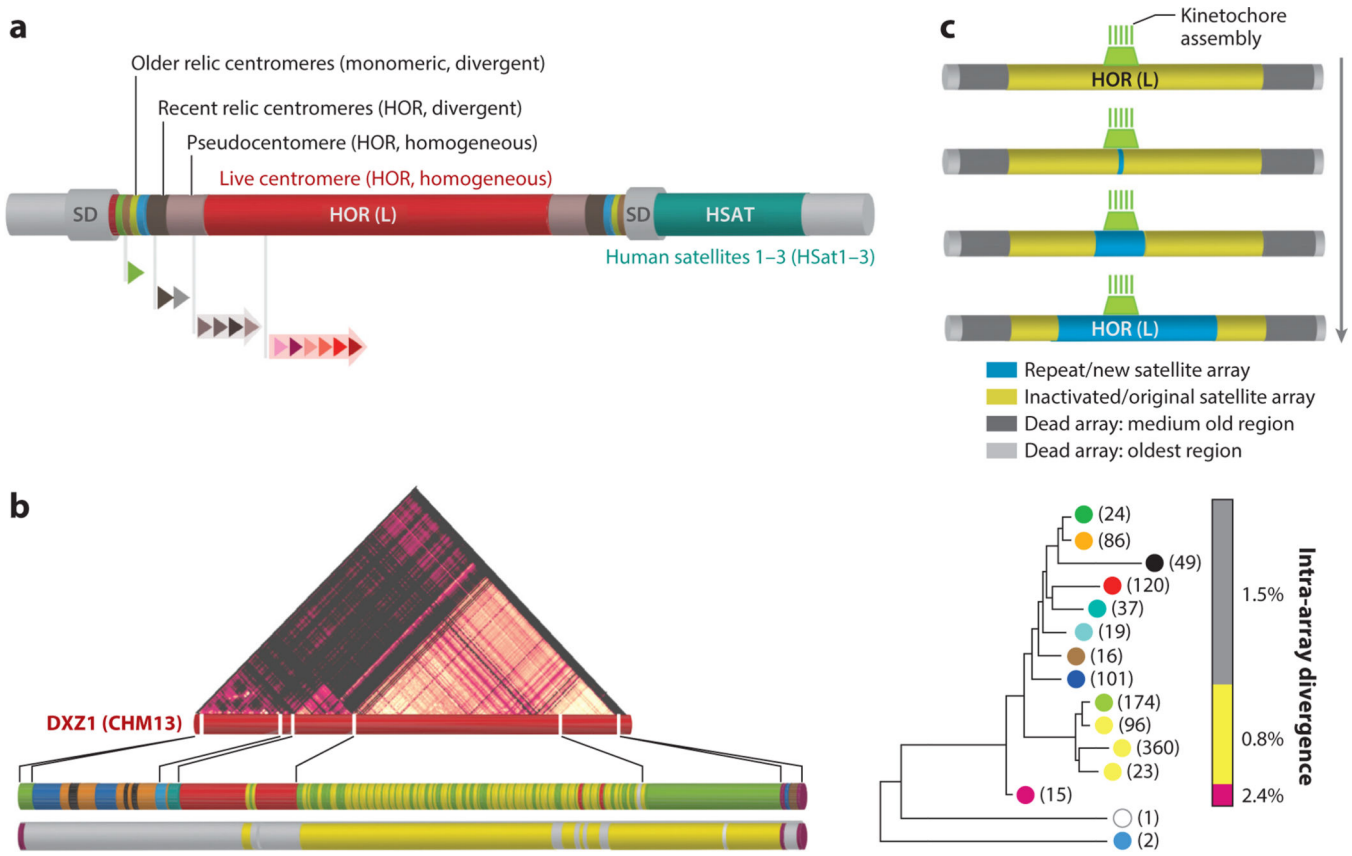

Figure 1.

Structure and evolution of alpha satellite arrays. (*a*) Illustration of the general genomic organization of a human centromeric region, which includes one homogeneous core made of chromosome-specific HORs (*red*) and the imperfect symmetrical organization of smaller arrays of various other homogeneous HORs [pseudocentromeres or inactive HOR arrays (*light gray*)], divergent HORs [recent relic centromeres (*dark gray*)], and multiple distinct divergent monomeric arrays (older relic centromeres, with blocks indicating colors describing phylogenetic assignments listed in Supplemental Table 1). These regions typically include other pericentromeric satellite classes [e.g., HSat1–HSat3 (*teal*)] and SDs. The entire centromeric region is defined by those sequences in the cenhap (48), presented as gray flanking regions extending into the p-arm and q-arm. Arrayed triangles indicate alpha satellite monomers and HORs of various length and structures composed of several different monomers. (*b*) Centromere X array haplotype maps, as determined from DXZ1 (S3CXH1L) HOR clustering and divergence data, provide evidence for block organization and gradient of divergence throughout all the layers. Classification of haplotypes is determined by phylogenetic relationships of the DXZ1 HOR repeats, revealing three distinct larger haplotypes (*gray*, *yellow*, and *light purple*). The larger haplotype structure (three major branches on the phylogenetic tree of haplotype consensus HORs) can be further characterized into 14 DXZ1-HOR subgroupings representing individual haplotypes (6, 65). One subbranch (*white*) represented by one HOR is a hybrid between two other haplotypes. The numbers in parentheses indicate the number of HORs in each clade. The dot plot for the self-aligned DXZ1 array (lighter areas have higher homogeneity) and StV map with few

variant HORs (*white*) are also shown. (*c*) Kinetochore selection model for satellite array evolution. This model (see Section 2.9) proposes that selfish selection operates on the array through the amplification of the repeat (*light blue*) due to the association with kinetochore (*green*) assembly, which locates itself on repeats to which it happens to have maximal affinity. Over time, the new satellite array (*light blue*) replaces the original satellite array (*yellow*), which shrinks progressively due to the ongoing deletion process. Centromeric arrays that are no longer associated with the kinetochore are considered dead and are arranged symmetrically, flanking the live arrays. Dead arrays are depicted as light gray (oldest region), dark gray (medium old), and adjacent yellow (newly inactivated dead alpha satellite array). Abbreviations: cenhap, centromere-spanning haplotype; HOR, higher-order repeat; HOR (L): live, or HOR array associated with kinetochore assembly; HSat, classical human satellites; SD, segmental duplication; StV, structural variant of a HOR. Figure adapted from data presented in Reference 6.

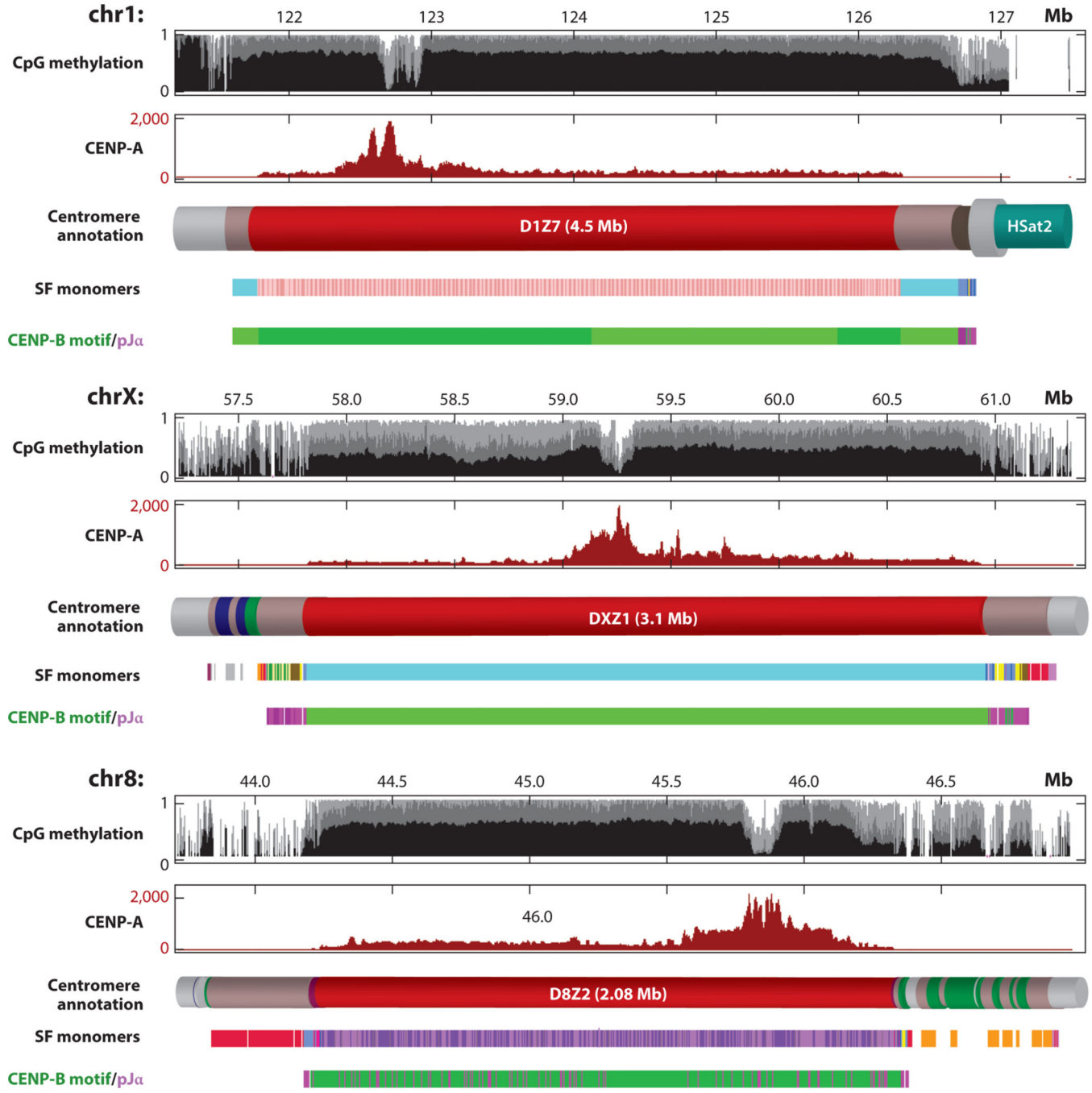

Figure 2.

Epigenetic characterization of three complete centromeric arrays from T2T assemblies of chr1, chrX, and chr8. Access to complete and accurate assemblies of human centromeric regions provides a new opportunity to characterize all live alpha satellite HOR arrays [shown for D1Z7, chr1-SF1 (*pink*); DXZ1, chrX-SF3 (*blue*); and D8Z2, chr8-SF2 (*purple*)] and adjunct dead arrays. Further, these maps offer a high-resolution study of CENP-B-binding motifs (*dark green* represents repeats where the motif is in forward orientation and *light green* represents those with a motif in reverse orientation), and pJα-binding

site sequences (*light purple*). Note that the regions enriched in reverse motifs indicate an inversion in centromere 1, the single unique event in all of the live centromeres. With the exception of centromere 8 (where CENP-B boxes and pJα are intermixed in the live array), live arrays within centromeric regions on chromosomes 1 and X contain CENP-B boxes, and flanking divergent monomeric regions contain pJα. The map of CpG methylation in ultralong Nanopore data obtained using long-read mapping protocols (previously described in 67) reveals dips in methylation that are coincident with sites of kinetochore assembly [illustrated with enrichment of CENP-A in native ChIP-seq data (52)]. Abbreviations: CENP-A, centromere protein A; CENP-B, centromere protein B; ChIP-seq, chromatin immunoprecipitation sequencing; chr, chromosome; HOR, higher-order repeat; SF, suprachromosomal family; T2T, telomere-to-telomere.