Chromosome Res (2012) 20:621–633
DOI 10.1007/s10577-012-9295-y

Human centromere genomics: now it's personal

Karen E. Hayden

Published online: 17 July 2012

Abstract Advances in human genomics have accelerated studies in evolution, disease, and cellular regulation. However, centromere sequences, defining the chromosomal interface with spindle microtubules, remain largely absent from ongoing genomic studies and disconnected from functional, genome-wide analyses. This disparity results from the challenge of predicting the linear order of multi-megabase-sized regions that are composed almost entirely of near-identical satellite DNA. Acknowledging these challenges, the field of human centromere genomics possesses the potential to rapidly advance given the availability of individual, or personalized, genome projects matched with the promise of long-read sequencing technologies. Here I review the current genomic model of human centromeres in consideration of those studies involving functional datasets that examine the role of sequence in centromere identity.

Responsible Editors: Rachel O'Neill and Beth Sullivan.

K. E. Hayden
Center for Biomolecular Science and Engineering,
University of California,
Santa Cruz, CA 95064, USA

K. E. Hayden (✉)
501 Engineering 2 Building, Mailstop CBSE/ITI,
UC Santa Cruz, 1156 High Street,
Santa Cruz, CA 95064, USA
e-mail: kehayden@soe.ucsc.edu

Keywords centromere · alpha satellite DNA · kinetochore · genomics

Abbreviations

CCAN	Constitutive centromere-associated network
CENP-B box	Centromere protein B binding motif
CENPs	Centromere proteins
HAC	Human artificial chromosome
HORs	Higher-order repeats
HSAT	Human satellite
LINE	Long interspersed repeat

Introduction

Centromere sequences function through an interaction with a hierarchical protein structure, the kinetochore, that couples the chromosomal locus to the mitotic spindle (Cheeseman and Desai 2008). Despite an essential role in genome inheritance, these specialized genomic regions remain vastly underrepresented in our reference assemblies and, consequently, remain isolated and unaccountable in ongoing studies in the genome sciences (Rudd and Willard 2004; She et al. 2004). Our limited view of centromeric regions stems from the challenges associated with assembling across millions of bases of highly repetitive and near-identical sequences.

These extreme sequence complexities have effectively thwarted standard sequence assembly and mapping algorithms optimized for single-copy DNA. Ongoing efforts to characterize and map across these regions will require new computational and experimental approaches that not only bridge the gap in our understanding of centromere sequence content and organization, but also promote the integration of existing comparative and functional datasets. Such advances are expected to heavily rely on existing experimental studies, as those briefly summarized in this review, that collectively establish guidelines to evaluate the first centromeric reference database and assembly.

Centromeric sequences vary considerably between individuals (Wevrick and Willard 1989; Warburton et al. 1991), introducing new datasets to the study of human genetic diversity, evolution, and disease. As a consequence of this variation, each human genome is expected to contain a personalized inventory of centromere sequence composition and organization. The increasing availability of personalized, or individual, genomes ushers in a new era for centromere genomics and further emphasizes the demand for sequence analysis tools that are poised to evaluate whole-genome datasets rather than select collections of cloned sequence libraries. In this review, I present a generalized view of sequence organization within the human centromeric regions, explore how this organization is anticipated to vary within population-based studies, and query the consequences of this introduced sequence variation with respect to centromere identity or the chromosomal capacity for kinetochore function.

A genomic model of human centromeres

All normal human centromeres are defined by the presence and abundance of an AT-rich satellite family, known as alpha satellite (Manuelidis 1978). This fundamental satellite sequence has been credited as the genetic and genomic definition of a human centromere: it interacts biochemically with inner kinetochore proteins and, at least in the few intensely studied subsets of alpha satellite sequences, is competent for de novo establishment of centromere identity in artificial chromosome assays (Harrington et al. 1997; Schueler et al. 2001). As a result, human centromere genomic models are centered on the sequence characteristics and chromosomal distribution of alpha satellite DNAs (Fig. 1a). Generally, alpha satellite is defined as an ~171-bp repeat unit, or monomer, arranged in a head-to-tail orientation and, often, extending with limited interruptions for millions of bases (Gray et al. 1985; Manuelidis 1976, 1978). Sequence comparisons between individual monomers reveal a highly divergent sequence family, with average pairwise identities of ~60–80 % (Rudd and Willard 2004; Waye and Willard 1987). Previous studies have capitalized on these diagnostic sequence patterns, providing a robust genomic understanding of monomer-based relationships within and between chromosomal subsets, often resulting in chromosome-specific maps of alpha satellite organization.

Within each human centromeric region, collections of alpha satellite can be further characterized by their relative genomic organization and functional correlation into two general subtypes: those that appear to be highly divergent, with infrequent occurrences of local homology, known as monomeric, and those monomers that are organized into multi-monomer repeat units, known as higher-order repeats (HORs), that are involved in expansive arrays of near-identical tandem repeats (Alexandrov et al. 1993; Willard and Waye 1987). Unlike monomeric regions, particular HORs directly interact with proteins involved in kinetochore assembly and are competent for de novo centromere establishment (Harrington et al. 1997). In fact, HOR units are enriched with a 17-bp binding motif for centromere protein B, or CENP-B box, shown in artificial chromosome assays to be relevant for centromere establishment (Masumoto et al. 1989; Ohzeki et al. 2002). All human centromeric regions contain one or more HOR arrays, often organized in megabase-sized arrays, largely distinguishable by the multi-monomer arrangement within each repeating unit (Alexandrov et al. 2001b).

Studies of the arrangement of individual HOR repeat units within each array provide evidence for localized expansion and contraction of variant repeat units into spatially distinct, homogenized domains (Warburton et al. 1992). The limited sequence variations that distinguish repeat units within a single array are insufficient for standard assembly; as a consequence, the linear organization within HOR arrays, often including sequences that are found between two adjacent HOR arrays, is collapsed or omitted from

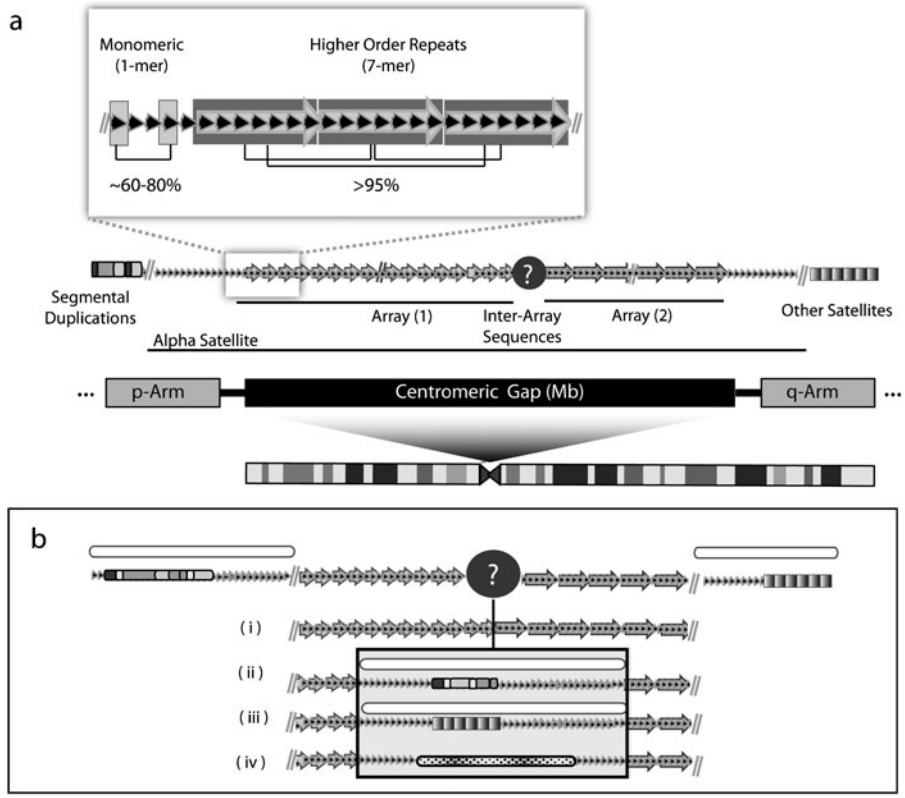

Fig. 1 Human sequence model of centromeric organization. Depicted are the currently known sequence composition and organization of human centromeric domains, **a** illustrating sites of segmental duplication, adjacent array, transposable elements, and alpha satellite sequence organization. These domains have a predominant satellite family, or alpha satellite, that is organized as either monomeric or HORs, representing a repeat consisting of multiple, divergent monomers. Individual alpha satellite monomers are, on average, highly divergent, yet comparisons between adjacent HORs reveal that they are near identical. More than one homogenized array can be described within a single human centromeric region, containing arrays with different monomer organizations of HORs, shown here as array 1 and array 2. Sites between two adjacent arrays represent largely uncharacterized regions in the human genome. **b** The "inner" centromeric transitions: (*i*) first, there is an absence of intermediate sequences, suggesting a model where the two arrays are directly adjacent; (*ii* and *iii*) inner centromeric transitions are predicted to be similar to those observed in the current reference assembly, with monomeric, segmental duplications, and/or other adjacent satellite arrangements; and (*iv*) a model illustrating where nonduplicated sequences could be present within inter-array sites, currently removed from view by the inability to assemble across HOR satellite arrays

genome assemblies. In contrast, monomeric regions provide adequate sequence variation for standard overlap alignment and therefore represent the majority of alpha satellite sequences included in the human reference assembly (Rudd and Willard 2004; She et al. 2004). Genomic studies of the annotated monomeric regions adjacent to centromere-assigned gaps on chromosomes X, 8, and 17 in the human reference assembly provide evidence for phylogenetically defined "blocks" of divergent satellite sequences that appear to gradually shift away from the homogenized array (Schueler et al. 2005, 2001; Shepelev et al. 2009). This specialized form of sequence evolution has also been monitored by regional patterns of transposable element insertions, as LINE are documented to increase in number and molecular dating when distanced from the homogenized HOR array (Schueler et al. 2005, 2001). Even at the short transition between monomeric and HOR repeats, increased levels of HOR repeat divergence are observed, thus promoting a generally accepted model of alpha satellite sequence evolution where array turnover and satellite variant innovation and expansion promote displacement and divergence of those sequences at the edge of the array (Schueler et al. 2005; Shepelev et al. 2009).

The transition from monomeric to HOR alpha satellite is only reported in the reference genome

assembly for 7 of the 43 available p and q arm transitions in the human genome (Rudd and Willard 2004; She et al. 2004). This incomplete representation is largely due to the presence and enrichment of other expansive tracts of multi-copy DNA, often characterized as segmentally duplicated DNA and/or other adjacent satellite families that prematurely interrupt assembly efforts across centromere transitions (She et al. 2004). Segmental duplications, or stretches of genomic sequences that share high sequence identity among chromosomally distributed copies, are commonly enriched at centromeric transitions (Bailey et al. 2002; She et al. 2004). These multi-copy sequences largely represent interchromosomal duplications, inhibiting chromosome-specific contig assignment extending into centromeric regions (She et al. 2004). Additionally, alpha satellite is commonly found adjacent to other satellite families, including classical human satellite families (HSAT I, II, and III) and beta and gamma satellites, each defined by their respective individual sequence composition and evolution (Lee et al. 1997; Warburton et al. 2008). In some cases, these adjacent, smaller satellite arrays within the centromeric transition regions can be traversed by paired reads and standard assembly and are included in the reference sequence (She et al. 2004; Warburton et al. 2008), yet larger arrays of satellites, such as HSAT II and III, encompass millions of bases that are difficult to distinguish between chromosomal subsets, thereby limiting assembly efforts to reach alpha satellite. These common features of human centromeric regions highlight the need for a new genomic strategy capable of studying alpha satellite sequence organization within the genomic context of adjacent sequence complexities.

Notwithstanding experimental advances in our broad understanding of human centromere sequence organization, gaps remain in our current genomic model. Centromeric regions provide an exciting new frontier for novel sequence variation—possibly even novel gene discovery. Indeed, in addition to satellite-based sequence variation, regions of uncharacterized DNA may reside between adjacent homogenized arrays. Non-satellite sequences could define many of these inter-array sequences, as observed on human chromosome 7, where two distinct alpha satellite arrays, D7Z1 and D7Z2, are separated by approximately 1 Mb of unknown, seemingly single-copy DNA (Wevrick and Willard 1991). Extension of the current model of sequence organization within centromere transitions, as observed in the reference assembly, may provide some clues about the sequences occupying uncharacterized regions (Fig. 1b). For example, these internal regions could contain a progression of satellite divergence, once again transitioning from HOR to divergent monomeric DNA. Interestingly, segmental duplications that are found directly adjacent to alpha satellite offer a mechanism to incorporate non-satellite DNA within inter-array regions. Although centromeric regions are commonly associated with compact, gene-poor heterochromatin, segmental duplications have been credited with "euchromatic colonization," introducing transcriptionally active landscapes capable of genetic innovation in the vicinity of centromeres (She et al. 2004). In the human genome, characterized non-satellite euchromatic-like islands exist within the satellite-rich regions on chromosome 21p and in the pericentromeric region of chromosome Yq11 and, similarly, may represent a new dataset of previously uncharacterized segmental duplications within the inter-array "transitions" missing from our genomic model (Kirsch et al. 2005; Lyle et al. 2007). One may also speculate that even novel, single-copy sequences could occupy these regions through centromeric repositioning, establishing centromere identity within a region of the genome that, over time, is enclosed by satellite expansion (Montefalcone et al. 1999). Future genomic surveys across centromeric satellite domains are expected to address the nature of these unknown sequences within intersatellite domains, thereby completing our generalized genomic model of centromere sequence organization.

Chromosomal distributions of centromeric sequence

Although human centromeric regions are defined by this general model of sequence organization (as shown in Fig. 1a), the distributions and sequence composition of both alpha and non-alpha sequences provide an opportunity to extend centromere genomics studies in a chromosome-specific manner (Willard 1985). Our understanding of sequence organization and content within each human centromeric region results from the pairing of experimentally derived satellite array maps with the annotation of each centromeric transition region currently available in the reference assembly (Rudd and Willard 2004; She et al. 2004;

Willard 1985). Here, I briefly consolidate these two datasets (Fig. 2), with a focus on alpha satellite HOR array organization and corresponding intra- and inter-array similarity between satellite subsets capable of challenging future chromosomal assignment of centromere sequences (Alexandrov et al. 1988).

Extensive experimental and reference assembly-based annotation studies of sequence organization on the X chromosome have provided our best genomic definition of a human centromere (Schueler et al. 2001). Sequence assembly within the X centromeric region currently represents the monomeric to higher-order transition from both p arm and q arm, linking both sides to a single HOR array (DXZ1) defined by a 12-monomer repeat unit (Ross et al. 2005; Schueler et al. 2005, 2001). Physical maps and direct sequencing across DXZ1 reveal that the centromeric gap can be defined as a multi-megabase array with limited pairwise sequence divergence (~1–2 %) to differentiate HOR repeats (Durfy and Willard 1989). Kinetochore proteins interact with DXZ1, demonstrated by human artificial chromosome assays and immunoprecipitation, defining the HOR repeat unit as the basic genetic unit for centromere identity (Schueler et al. 2001). Similarly, 11 human centromeric regions are currently defined by a single HOR array, where it is expected that each single, representative HOR sequence would serve as the site of kinetochore assembly (Alexandrov

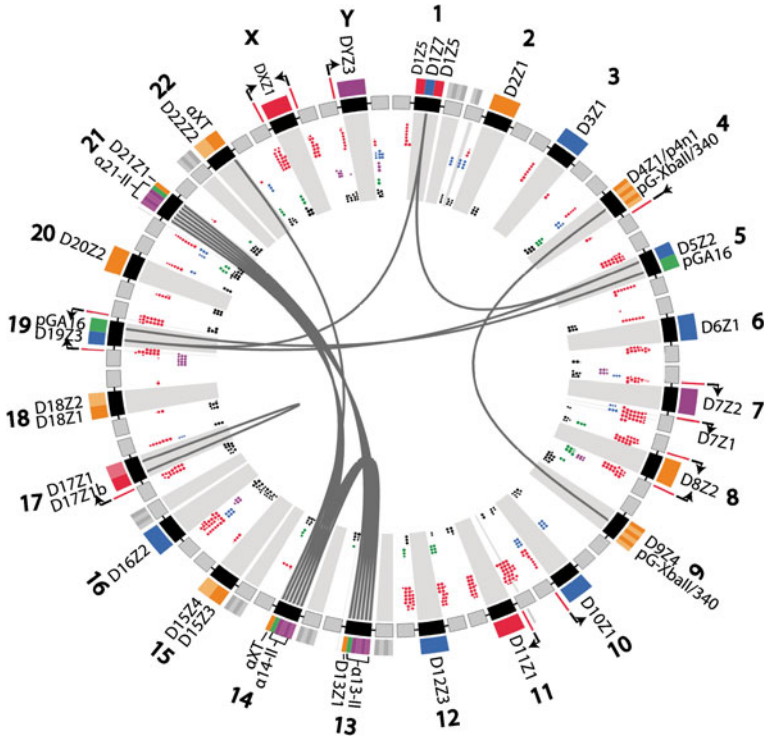

Fig. 2 Chromosomal distributions of centromeric sequences in the human genome. Human centromeric regions are defined as the 3-Mb assigned centromeric gap in the GRCh37/hg19 human assembly (*black*, *center*) and the 2 Mb of sequences found directly adjacent to the centromere gap on either the p arm (*gray*, *left*) or q arm (*gray*, *right*). Experimental characterization and nomenclature of alpha satellite higher-order repeat arrays that occupy each centromere-assigned gap are shown above, as previously summarized: (Alexandrov et al. 2001a; Finelli et al. 1996; Vissel and Choo 1991, 1992). *Dark gray*/*light gray*-banded sites indicate "gapped" heterochromatic regions within the portrayed centromeric regions. Color assignment for each HOR array block provides information for previously established phylogenetic similarity: J1/J2, dimeric family 1 (*blue*); D1/D2, dimeric family 2 (*orange*); W1-5, pentameric family (*red*); M, monomeric family 4 (*purple*); and R1/R2, irregular family 5 (*green*). *Dark gray lines* between these families illustrate those that are observed to be greater than 90 % identical and are difficult to study in a chromosome-specific manner. Below each *gray*, adjacent p arm and q arm block are the available satellite annotations in the hg19 reference assembly—alpha satellite (*red*), HSATII&III/(CATTC)n (*blue*), beta satellite (*purple*), gamma satellite (*green*), other satellite (*black*)—as presented in genome annotation tracks (Kent et al. 2002). Regions of monomeric to higher-order transition are highlighted in *red bars* that extend from the p arm and/or q arm; *black arrows* show membership to the array, with GRCh37/hg19 liftOver coordinates provided (Rudd and Willard 2004). The image was created using the circos software (Krzywinski et al. 2009)

et al. 2001b). However, few studies have examined how this model of kinetochore interaction changes in human centromeric regions containing two or more arrays that differ considerably in array length and HOR monomer arrangement. Abnormal chromosomes containing rearranged organization of alpha satellite DNAs, often studied in the context of HOR array deletion maps and regional duplication of satellite arrays, suggest that array sequence arrangement as well as spatial distance between arrays may influence the location of kinetochore assembly (Sullivan and Willard 1998; Tyler-Smith et al. 1993). Additionally, limited studies that focus on normal centromeric sites with more than one HOR array provide evidence for a single, "active" HOR array, demonstrated by the presence and abundance of proteins involved in proper kinetochore assembly (Haaf and Ward 1994; Vafa and Sullivan 1997). It is unclear whether certain genomic features, such as array size, homogenization patterns, or level of divergence, are predictive of centromere activity. Such studies would rely on a comprehensive inventory of all "active" and "non-active" HOR arrays within individual genomes to survey potential functional associations with genomic organization.

Patterns of intra- and inter-array sequence variation in alpha satellite provide an additional opportunity to study centromere identity of common HOR sequences within different, chromosomally defined, subsets. For example, a dimeric HOR family on chromosomes 1 (D1Z7), 5 (D5Z2), and 19 (D19Z1) is nearly indistinguishable between each centromeric distribution (Baldini et al. 1989). Such extreme interchromosomal homology confounds assembly across centromeric regions and prompts consideration of diagnostic single-base-pair changes to study centromere activity in a chromosome-specific manner. Studies aimed to correlate centromere identity within an all-inclusive alpha satellite dataset will be guided by this general understanding of shared sequence variation, genome distribution, and correlation with kinetochore assembly. It is very likely that our current maps underrepresent the total number of HOR arrays, as predictions in the past relied on a certain level of array abundance for detection. Novel detection of a few examples of these HOR arrays in the reference and unassembled reads provides evidence that less abundant chromosomally assigned HOR arrays remain uncharacterized (Alkan et al. 2007; Rudd and Willard 2004; Warburton et al. 2008). Acknowledging the incomplete nature of our current maps challenges our confidence in assigning any stretch of alpha satellite sequence represented on the reference assembly, as short windows of identical sequence homology could be mapped to multiple, distinct genomic locations, complicating the interpretation of mapping in short-read functional datasets. This general lack of "mappability" increases the urgency to extend our current satellite maps for each chromosome region and present a genome-wide measure of sequence relatedness and chromosome specificity within alpha satellite DNAs.

Alpha satellite in population studies

Centromeric sequences are expected to represent a rich source of molecular variation in the human population. Nonhomologous sequence exchange between near-identical copies of each repeat can result in regional duplication or deletion within or between alpha satellite arrays, providing opportunities for rapid sequence evolution of select variants (Dover 1982). As a result, chromosome-assigned HOR array sequence libraries more often represent a mixture of homologous arrays that may vary considerably in sequence composition and organization between maternally and paternally derived chromosomes (Fig. 3a).

In contrast to the predicted variation in satellite arrays among unrelated individuals, alpha satellite arrays are stable in the context of a pedigree. Three-generational pedigree studies, in which the transmission of individual arrays was monitored by chromosome-assigned satellite markers, provide little evidence of restriction-site or array length changes throughout meiosis (Wevrick and Willard 1989). The absence of observable recombination within centromeric regions, at least within the short time of familial inheritance, ensures stable, Mendelian transmission of centromeric sequences (Fig. 3b) (Wevrick and Willard 1989).

Although meiotically stable in the context of a pedigree, alpha satellite arrays can vary considerably across unrelated individuals, providing substantial DNA variation to investigate polymorphic centromeric patterns within the population. For example, collective sequence-based markers on the Y chromosome higher-order array (DYZ3) indicate that centromeric features are polymorphic in human populations, demonstrated using both patterns of retrotransposition and

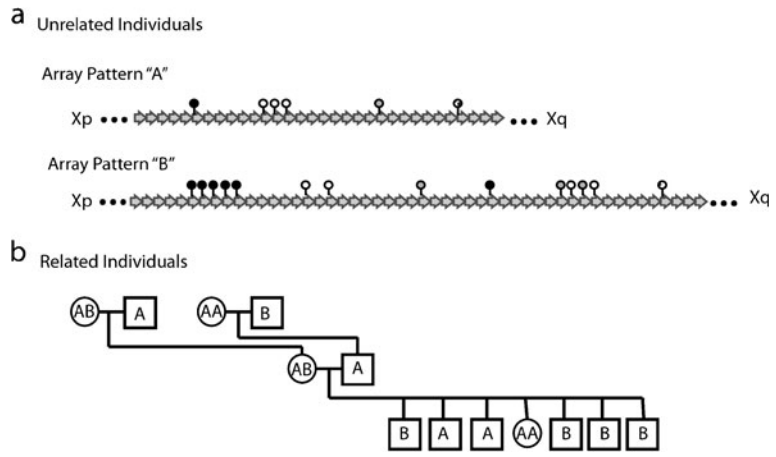

Fig. 3 Alpha satellite population-based studies. **a** An example of an alpha satellite array from the X chromosome from two unrelated individuals is shown to differ in array length, illustrated as the sum number of repeat units, and by individual higher-order repeat unit base differences, indicated as *circles* (shading to demonstrate shared patterns of sequence-based changes). **b** These arrays, although capable of differentiating maternally and paternally derived chromosomes, are observed to have stable, Mendelian transmission within the context of three-generation pedigree studies

array length estimates to predict independent grouping of individuals of European and Asian descent (Oakey and Tyler-Smith 1990; Santos et al. 2000). Additionally, variations in HOR monomer organization from the centromeric array on chromosome 17 (D17Z1) were reported in different population-based proportions within human populations (Warburton et al. 1991). Gradual fixation of any one repeat unit is expected to occur over time at a rate sensitive to satellite sequence variants and effective population size (Ohta and Dover 1984). In line with this hypothesis, HOR repeat unit fixation within D17Z1 has been detected within a cohort of Pygmies and is estimated to have occurred due to limited alpha satellite sequence variance within the small, isolated breeding population (Cavalli-Sforza et al. 1986). These studies illustrate an early understanding of sequence variation between individuals in the human population and provide a basis to extend this analysis to genomes of humans existing today.

Centromeric alpha satellite arrays are thought to evolve by molecular drive, in which variants are able to spread quickly—driven by homology-based sequence exchange through a sequence family—and fix in a population independent of both selection and drift (Dover 1982). However, given the importance of proper chromosome segregation in organismal viability, individual arrays that directly associate with centromere function may be subject to selection, presenting a model where intragenomic conflict is resolved by meiotic drive (Henikoff et al. 2001; Malik and Henikoff 2002). Further, recent studies have observed consistent evidence for selective sweeps in human centromeric regions, suggesting that centromeres, or some genetic elements within centromeric gaps, may be under selection in the human population (Williamson et al. 2007). Centromere regions may also be placed under evolutionary pressure by an association with human disease. Indeed, several disease-mapping studies have reported associations with regions overlapping centromeric gaps in the human genome, including multiple sclerosis (centromere gap on chromosome 1), schizophrenia (centromere gap on chromosomes 3, 5, 8, 11, 16, and 19), and cancer (centromere gap on chromosome 5) (Lencz et al. 2007; Reich et al. 2005; Stacey et al. 2008). Efforts to characterize sequences within centromere-assigned gaps will offer new genetic information that is expected to contribute to our understanding of human evolution and disease.

Centromere identity: a sequence perspective

Centromere sequences interact with an increasingly identified number of DNA-binding kinetochore proteins and tolerate the unique level of DNA catenation, supercoiling, and tension necessary for proper chromosome segregation. The genomic definition of a centromere–kinetochore interface is revealed through the specific

DNA contact of a subset of constitutive, inner kinetochore proteins, including CENP-A, CENP-B, CENP-C, CENP-T/W, and CENP-S/X (Fig. 4) (Hori et al. 2008). Apart from CENP-B, which binds specifically to a 17-bp motif (CENP-B box) (Masumoto et al. 1989), the remaining inner kinetochore proteins appear to bind centromeric DNA in a complex manner independent of a clear DNA binding site. From the perspective of the underlying DNA, these interactions introduce a combination of multi-protein interactions assuming sequence flexibility to tolerate the induced supercoiling, spacing, and bending necessary for nucleosome wrapping to accommodate centromere-specific chromatin (Furuyama and Henikoff 2009; Hori et al. 2008; Nishino et al. 2012). In addition to direct sequence-based interactions with inner kinetochore proteins, centromeric sequences also accommodate a hierarchy of constitutive and dynamic protein-based interactions that bridge the inner kinetochore and the outer kinetochore in a protein-mediated transfer of tension from spindle forces to those proteins that directly bind DNA (Hori et al. 2008; Screpanti et al. 2011). Moreover, proper assembly of pericentromeric heterochromatin, typically correlated with functional kinetochores, may be influenced by sequence-based factors, such as transcriptional regulation and specialized chromatin compaction (Folco et al. 2008). Our understanding of the genomic definition of human centromeres will require a much broader understanding of the sequence-based interactions that define the centromere–kinetochore interface.

Current studies—in the absence of a human centromere reference sequence—effectively remove the larger genomic context from our understanding of the sequence role in establishing and maintaining the centromere–kinetochore interface. Currently, efforts to understand the genomic contribution to centromere function are largely confined to human artificial chromosome (HAC) studies. In these assays, HOR alpha satellite sequences consistently demonstrate a "sequence code" to establish and stably maintain centromere identity (Harrington et al. 1997; Schueler et al. 2001). At least in rare instances, sequences outside of normal centromere DNA are able to recruit and stably maintain centromere-specific proteins (CENPs) (Choo 1997). Yet, in contrast to alpha satellite, these studies

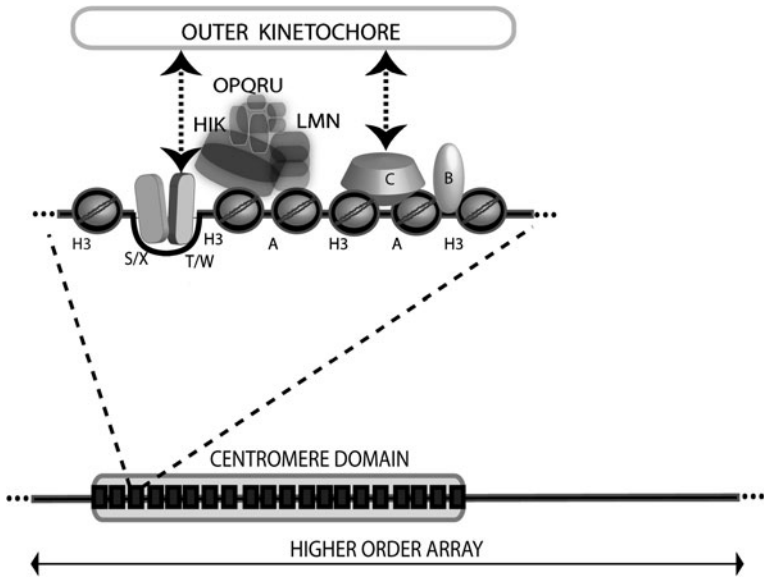

Fig. 4 Kinetochore protein interactions with centromeric DNA. Centromeric domains are organized through the connection of inner and outer kinetochore protein interactions that facilitate spindle attachment. A model for kinetochore proteins is shown here with a focus on their interaction with the underlying DNA, and centromeric chromatin (CENP-A noted as *A*, nucleosomes containing histone H3 as *H3*, and centromere proteins that are shown to directly interact with the underlying DNA: CENP-B (*B*), CENP-C (*C*), CENP-T-W complex (*T/W*), and CENP-S-X complex (*S/X*)). The protein-determined connection to the outer kinetochore is indicated with *black arrows*. Additional protein complexes that characterize the constitutive centromere-associated network (CCAN) are indicated in *gray*. This organization, as currently understood, contributes to our understanding of the centromere–kinetochore interface. These sequence-kinetochore-mediated interactions are observed to occur at multiple, localized sites within a centromere domain, shown here to only represent a proportion of an active HOR array

have not demonstrated an ability for de novo centromere formation in human artificial chromosomes in the absence of direct experimental adherence of key kinetochore proteins, thereby indicating a lower sequence-based efficiency for centromere identity (Guse et al. 2011; Saffery et al. 2001). The current scope of sequence evaluation for centromere competency, as observed through stable inheritance of artificial chromosome assay, is currently limited by the paucity of centromeric sequences available for further study and the time consuming and experimentally extensive standards to test sequences by HACs. As our genomic maps of human centromeres improve, genome-wide quantitative studies are expected to provide a comprehensive survey of the range of centromeric sequence variation competent for function, thereby enabling further hypothesis-driven HAC-based studies to test predictive sequence-based trends in centromere identity.

Genomic studies of centromere function are expected to depend not only on mapping of sites of kinetochore assembly, but also on the regional arrangement of these binding sites with respect to the centromeric chromatin domain. Inner kinetochore proteins are observed by extended chromatin fibers to bind only within a portion of a given individual HOR array (Blower et al. 2002). Further, direct comparisons between multiple cell lines demonstrated that the length of the sequence interface with the kinetochore, or the domain of centromeric chromatin, is proportional to the size of alpha satellite array (Sullivan et al. 2011). When the abundance of kinetochore proteins increases, commonly observed in some human cancers, the domain of kinetochore–DNA interaction expands within the given array, providing a larger genomic interface to kinetochore assembly (Sullivan et al. 2011). In consideration of this variance in sequence organization and protein abundance, future studies aimed to define the comprehensive range of centromere-competent sequences will benefit from an accumulation of personalized centromere genomics within the epigenetic context of matched cell types.

Centromere model of sequence optimality

Centromeres are represented by the direct sequence contacts that are made between inner kinetochore proteins and the underlying DNA, and the spatial distribution and local enrichment of those contacts, or specialized chromatin domain. Provided with this definition, one could initially assign the fundamental genomic unit for centromere identity to a minimal locus that is competent to recruit and maintain inner kinetochore proteins (Fig. 4), thus supplying an opportunity for local kinetochore assembly. Previous human artificial chromosome assays provided evidence that certain alpha satellite HOR repeat units are centromere-competent genomic units, and other sequences, such as monomeric alpha satellite and non-alpha satellite DNA, appear to be considerably less efficient at de novo establishment and stable maintenance (Harrington et al. 1997; Saffery et al. 2001). In line with these observations, the human genome is expected to contain sites with varying probabilities for proper protein binding affinities of inner kinetochore proteins. This argues against a strict epigenetic definition, where it is assumed that all sequences are *equally* capable or optimal for centromere maintenance and inheritance, and rather emphasizes that the genome could, in theory, be annotated by likelihood of involvement with centromere function due to the presence of inherent sequence features (Fig. 5a).

Human centromeres operate with respect to a domain, or linear arrangement of kinetochore binding sites. The length and spatial organization of inner kinetochore proteins within these domains are considered critical for proper secondary chromosomal structure, kinetochore assembly, and interactions with the spindle apparatus (Blower et al. 2002; Hori et al. 2008). These observations emphasize an additional genomic pressure to organize centromere-competent sequences within a broader, yet regional, context, resulting in a measure of centromere-domain efficiency. Higher-order arrays of alpha satellite provide an ideal sequence organization for this genomic model, as they ensure an expansive region defined by tandemly organized genomic units, each with high kinetochore binding efficiency (Fig. 5b). Mechanisms underlying array homogenization, under this hypothesis, could serve as a "molecular dial" to expand or contract these sequence arrays for participation in centromere function, potentially placing them under selection (Henikoff et al. 2001; Malik and Henikoff 2002). Regional enrichment of centromere-competent sequences is expected to exist outside of alpha satellite regions, yet these regions are suspected to have an inconsistent arrangement of

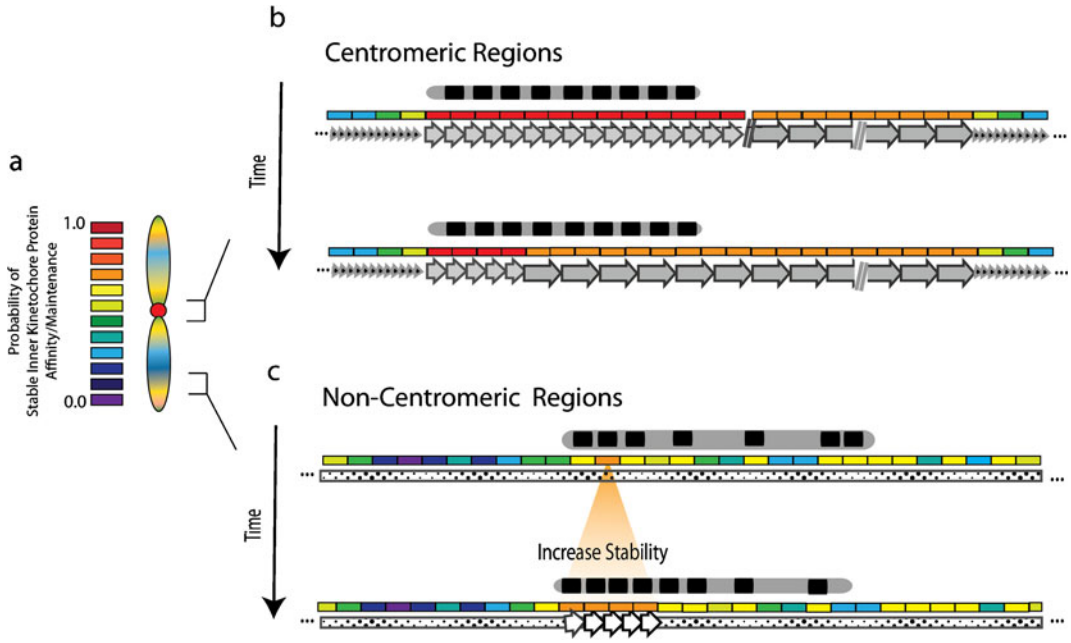

Fig. 5 Sequence optimality and centromere identity. **a** In this model, the genome is redefined by the regional probabilities of maintaining the fundamental genomic unit necessary for inner kinetochore binding. These complex, multi-protein binding sites are proposed to vary in centromere competency, as illustrated in the color spectrum from *red*, demonstrating high efficiency, to *purple*, demarcating regions of poor efficiency. In the human genome, alpha satellite—the predominant sequence in centromeric regions—has been shown to contain sequence features important for centromere identity and is noted in the chromosome organization as *red*. An expanded view of this region is shown (**b**), providing the expected variability in efficiency within these regions, distributed between different higher-order arrays (*red* versus *orange*), and over adjacent, divergent monomeric alpha satellite sequences (*yellow* to *blue*). The inner kinetochore binding sites are noted in *black*, and the domain—or linear arrangement of binding sites—is shown in *light gray* as a collection of these sites. In this example, the kinetochore is shown to interact with the most optimal sequences within the centromeric region. Over time, satellites in these regions are expected to expand and contract, thereby altering the sequence affinity distributions in the region. Less favorable interactions are still tolerated in these regions, creating an opportunity for an evolving sequence definition. Highly divergent, monomeric alpha satellite is suspected to have a lower, inconsistent range of sequence efficiency, providing a less likely substrate for kinetochore formation. Similar to monomeric, (**c**) the rest of the genome is expected to have fewer sequence features competent for centromere identity. Although such regions are expected to exist with low, and inconsistent, distribution in the genome, it remains possible to form functional centromeres. It is hypothesized that, over time, these regions are capable of stabilizing by increasing the available, linear arrangement of sites that are optimal for kinetochore formation, as shown here as the introduction of an expanding tandem repeat

binding efficiencies (Fig. 5c). Such sites that demonstrate centromere identity represent a short-term genomic solution, as they lack a sequence-based mechanism to propagate, or increase sites of, kinetochore binding efficiency over time.

Dosage and dynamic exchange of inner kinetochore proteins throughout the cell cycle place an additional form of genomic regulation on human centromeres. Once established, centromeric domains are expected to tolerate variability in adjacent sequence efficiencies to accommodate increased abundance of inner kinetochore proteins or general stochastic shifting of inner kinetochore proteins. Therefore, it is possible for sequences that are less optimal to acquire membership into this domain due to spatial context, conveying a genomic advantage. Changes in inner kinetochore protein abundance in human cancer cell lines have demonstrated that centromeric domains expand relative to the underlying satellite array (Sullivan et al. 2011). Human artificial chromosome assays provide further support for establishment and dosage-dependent spreading of kinetochore-binding proteins to adjacent DNA (Lam et al. 2006). With increased dosage, it is likely that other optimal regions of the genome might acquire ectopic localization of inner kinetochore proteins, perhaps coupling epigenetic

regulation and sequence-based affinity in establishing centromere identity.

This sequence-based model predicts how centromere identity may behave over time and within a population. Each human centromeric region offers an abundance of alpha satellite sequences, often organized into multiple HOR arrays that are thought to vary considerably between individuals in the population. Abundance of highly efficient sequences may confer stability to the centromere region, without a strict sequence definition within a population of individuals. This excess of regional centromere-competent sequences may provide some security of centromere function, as rapid gain or loss of sequence in this region may be generally tolerated. Areas outside of the normal centromeric regions are expected to lack this general excess of optimal sequences; however, over time, these regions may acquire tandemly arrayed sequences, increasing the genomic prospect for stable centromere identity. Over evolutionary time, it is likely that kinetochore binding affinities will change, conferring an advantage to sequences that were previously considered less efficient and leading to sequence family expansion throughout a given population (Henikoff et al. 2001; Malik and Henikoff 2002). Attention to the range of sequences that demonstrate centromere competency in vivo—as observed through surveys of the kinetochore interface in genomes that vary in centromeric sequence organization and/or kinetochore protein abundance—may guide our understanding of how these regions of the genome function and change over time.

Conclusions and closing remarks

Centromeric sequence discovery, annotation, and eventual assembly will be guided by previous, satellite-based experimental studies. These collective studies not only support our current genomic model for centromere sequence organization and function, but also provide evidence of substantial chromosome-assigned sequence variation within these multi-megabase-sized gaps in our reference assemblies. Understanding the functional relevance of this sequence variability in the human population, and in the context of human disease, will rely on carefully matched epigenetic and genomic datasets to study the sequence interface with sites of kinetochore assembly. The field of centromere genomics is expected to progress through the availability of personalized genomes, capable of presenting sequence-based comparisons of centromere-competent sites within the human population. Each "genomic instance" of the centromere interface, with consideration to each individual sequence and kinetochore protein abundance, is expected to identify a range of sequences that are competent for function. Accumulation of these studies should dramatically expand our understanding of the ordinal range of sequence efficiencies and promote a broad genomic definition of centromere identity.

Acknowledgments I would like to gratefully acknowledge Dr. B. Sullivan, Dr. K. Scott, Dr. E. Strome, and Dr. HF Willard for helpful discussions that guided this review. I would also like to thank Dr. J. Kent and my colleagues within Dr. D. Huassler's group, at the University of California, Santa Cruz, for their support and evaluation of this manuscript.

References

Alexandrov IA, Mitkevich SP, Yurov YB (1988) The phylogeny of human chromosome specific alpha satellites. Chromosoma 96:443–453

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 21:2209–2215

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

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 3:1807–1818

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV et al (2002) Recent segmental duplications in the human genome. Science 297:1003–1007

Baldini A, Smith DI, Rocchi M, Miller OJ, Miller DA (1989) A human alphoid DNA clone from the EcoRI dimeric family: genomic and internal organization and chromosomal assignment. Genomics 5:822–828

Blower MD, Stockwell TB, Karpen GH (2002) Conserved organization of centromeric chromatin in flies and humans. Developmental cell 2:319–330

Cavalli-Sforza LL, Kidd JR, Kidd KK, Bucci C, Bowcock AM et al (1986) DNA markers and genetic variation in the human species. Cold Spring Harb Symp Quant Biol 51 (Pt 1):411–417

Cheeseman IM, Desai A (2008) Molecular architecture of the kinetochore-microtubule interface. Nat Rev Mol Cell Biol 9:33–46

Choo KHA (1997) Centromere DNA dynamics: latent centromeres and neocentromere formation. Am J Hum Genet 61:1225–1233

Dover GA (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

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 5:810–821

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 38:325–330

Folco HD, Pidoux AL, Urano T, Allshire RC (2008) Heterochromatin and RNAi are required to establish CENP-A chromatin at centromeres. Science 319:94–97

Furuyama T, Henikoff S (2009) Centromeric nucleosomes induce positive DNA supercoils. Cell 138:104–113

Gray KM, White JW, Costanzi C, Gillespie D, Schroeder WT et al (1985) Recent amplification of an alpha satellite DNA in humans. Nucleic Acids Res 13:521–535

Guse A, Carroll CW, Moree B, Fuller CJ, Straight AF (2011) In vitro centromere and kinetochore assembly on defined chromatin templates. Nature 447:354–358

Haaf T, Ward DC (1994) Structural analysis of alpha-satellite DNA and centromere proteins using extended chromatin and chromosomes. Hum Mol Genet 3:697–709

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

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

Hori T, Amano M, Suzuki A, Backer CB, Welburn JP et al (2008) CCAN makes multiple contacts with centromeric DNA to provide distinct pathways to the outer kinetochore. Cell 135:1039–1052

Kent W, Sugnet C, Furey T, Roskin K et al (2002) The human genome browser at UCSC. Genome Res 12:996–1006

Kirsch S, Weiss B, Miner TL, Waterston RH, Clark RA et al (2005) Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. Genome Res 15:195–204

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R et al (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645

Lam AL, Boivin CD, Bonney CF, Rudd MK, Sullivan BA (2006) Human centromeric chromatin is a dynamic chromosomal domain that can spread over noncentromeric DNA. Proc Natl Acad Sci U S A 103:4186–4191

Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC (1997) Human centromeric DNAs. Hum Genet 100:291–304

Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV et al (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci U S A 104:19942–19947

Lyle R, Prandini P, Osoegawa K, ten Hallers B, Humphray S et al (2007) Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. Genome Res 17:1690–1696

Malik HS, Henikoff S (2002) Conflict begets complexity: the evolution of centromeres. Curr Opin Genet Dev 12:711–718

Manuelidis L (1976) Repeating restriction fragments of human DNA. Nucleic Acids Res 3:3063–3076

Manuelidis L (1978) Chromosomal localization of complex and simple repeated human DNAs. Chromosoma 66:23–32

Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T (1989) A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol 109:1963–1973

Montefalcone G, Tempesta S, Rocchi M, Archidiacono N (1999) Centromere repositioning. Genome Res 9:1184–1188

Nishino T, Takeuchi K, Gascoigne KE, Suzuki A, Hori T et al (2012) CENP-T-W-S-X forms a unique centromeric chromatin structure with a histone-like fold. Cell 148:487–501

Oakey R, Tyler-Smith C (1990) Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. Genomics 7:325–330

Ohta T, Dover GA (1984) The cohesive population genetics of molecular drive. Genetics 108:501–521

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

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 37:1113–1118

Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K et al (2005) The DNA sequence of the human X chromosome. Nature 434:325–337

Rudd MK, Willard HF (2004) Analysis of the centromeric regions of the human genome assembly. Trends Genet 20:529–533

Saffery R, Wong LH, Irvine DV, Bateman MA, Griffiths B et al (2001) Construction of neocentromere-based human minichromosomes by telomere-associated chromosomal truncation. Proc Natl Acad Sci USA 98:5705–5710

Santos FR, Pandya A, Kayser M, Mitchell RJ, Liu A et al (2000) A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. Hum Mol Genet 9:421–430

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

Schueler MG, Dunn JM, Bird CP, Ross MT, Viggiano L et al (2005) Progressive proximal expansion of the primate X chromosome centromere. Proc Natl Acad Sci USA 102:10563–10568

Screpanti E, De Antoni A, Alushin GM, Petrovic A, Melis T et al (2011) Direct binding of Cenp-C to the Mis12 complex joins the inner and outer kinetochore. Curr Biol 21:391–398

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 430:857–864

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: e1000641

Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA et al (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor–positive breast cancer. Nat Genet 40:703–706

Sullivan BA, Willard HF (1998) Stable dicentric X chromosomes with two functional centromeres. Nat Genet 20:227–228

Sullivan LL, Boivin CD, Mravinac B, Song IY, Sullivan BA (2011) 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(4):457–470

Tyler-Smith C, Oakey RJ, Larin Z, Fisher RB, Crocker M et al (1993) Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. Nat Genet 5:368–375

Vafa O, Sullivan KF (1997) Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. Curr Biol 7:897–900

Vissel B, Choo KHA (1991) Four distinct alpha satellite subfamilies shared by human chromosomes 13, 14 and 21. Nucleic Acids Res 19:271–277

Vissel B, Choo KHA (1992) Evolutionary relationships of multiple alpha satellite subfamilies in the centromeres of human chromosomes 13, 14, and 21. J Mol Evol 35:137–146

Warburton PE, Greig GM, Tea H, Willard HF (1991) PCR amplification of chromosome-specific alpha satellite DNA: definition of centromeric STS markers and polymorphic analysis. Genomics 11:324–333

Warburton PE, Wevrick R, Mahtani MM, Willard HF (1992) Pulsed-field and two-dimensional gel electrophoresis of long arrays of tandemly repeated DNA: analysis of human centromeric alpha satellite. Methods Mol Biol 12:299–317

Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G (2008) Analysis of the largest tandemly repeated DNA families in the human genome. BMC Genomics 9:533

Waye JS, Willard HF (1987) Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. Nucleic Acids Res 15:7549–7569

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. Proc Natl Acad Sci U S A 86:9394–9398

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 19:2295–2301

Willard HF (1985) Chromosome-specific organization of human alpha satellite DNA. Am J Hum Genet 37:524–532

Willard HF, Waye JS (1987) Hierarchical order in chromosome-specific human alpha satellite DNA. Trends Genet 3:192–198

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R (2007) Localizing recent adaptive evolution in the human genome. PLoS genetics 3:e90