

The role of DNA sequence in centromere formation

Jonathan C Lamb and James A Birchler

Address: University of Missouri, Division of Biological Sciences, Columbia, MO 65211, USA.

Correspondence: James A Birchler. E-mail: BirchlerJ@Missouri.edu

Published: 29 April 2003

Genome Biology 2003, **4**:214

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2003/4/5/214>

© 2003 BioMed Central Ltd

Abstract

Centromeres are key to the correct segregation and inheritance of genetic information. Eukaryotic centromeres, which are located in large blocks of highly repetitive DNA, have been notoriously difficult to sequence. Several groups have recently succeeded in analyzing centromeric sequences in human, *Drosophila* and *Arabidopsis*, providing new insights into the importance of DNA sequence for centromere function.

Centromeres are essential for the proper segregation of chromosomes during cell division in eukaryotes. They are characterized by highly repetitive DNA regions and bound kinetochore proteins, which are required for the attachment of microtubules to the chromosomes during mitosis. Centromeres are a paradox in that their basic function is highly conserved across eukaryotes but their sequences are divergent, even between closely related species [1]. Several investigators have therefore suggested that the DNA sequence may not be essential in centromere formation [2]. It has been difficult to address this issue because of a lack of complete sequence for any higher eukaryotic centromere. Sequencing efforts have been confounded because centromeres are located in regions of highly repetitive DNA. Several groups [3-7] have recently developed novel methods to overcome these difficulties and report extensive centromeric sequence data from human, *Drosophila* and *Arabidopsis*.

Centromere sequences in different species

Deletion of large regions of the human Y chromosome has shown that centromere activity is associated with a block of tandemly repeated 171 base-pair (bp) units, termed α -satellite DNA [8]. Further work has demonstrated that every human centromere is associated with arrays of this α -satellite DNA that can be several megabases (Mb) in size. These massive arrays are imbedded between blocks of pericentric heterochromatin containing highly repetitive DNA [9]. *In situ*

hybridization with α -satellite and immunolabeling using antibodies against kinetochore proteins also confirms that centromeres are located in these regions [10].

Schueler *et al.* [3] used variation among the 171 bp repeats of α -satellite DNA in the human centromere to design PCR markers. The markers were used for constructing a 500 kilobase (kb) contig of bacterial artificial chromosomes (BACs) that covers a region that is immediately adjacent to, and including part of, a 3 Mb array of α -satellite located at the centromere of the human X chromosome. Shotgun and BAC-end sequencing gave a sampling of this region that consisted of approximately 62% diverged α -satellite DNA, about 24% other satellite repeats, and about 16% LINE-type retroelements, as well as other sequences. The 3 Mb array of α -satellite DNA consists of nearly identical copies of the 171 bp unit that have more than 99% sequence identity and are all oriented in the same direction. At the edge of the array is approximately 40 kb of α -satellite DNA that becomes more divergent with distance from the center of the 3 Mb array, moving from 98% to 70% identity at the edge.

Arabidopsis centromeres include a 178 bp satellite repeat, which is organized in tandem arrays that range in size from 0.4 Mb to 1.4 Mb on different chromosomes and are located between regions enriched for various satellites and other repetitive elements [6,11]. The clusters of α -satellite DNA in human and the 178 bp centromeric element in *Arabidopsis*

are organized in similar ways, although their primary sequences are completely unrelated. Interestingly, centromeres of other plants have also been shown to contain DNA elements of similar length, and this may reflect a common requirement for centromere function (see, for example, [12]).

To overcome difficulties in sequencing repetitive DNA from *Drosophila* centromeres, a novel approach [5] was used involving the *Drosophila* minichromosome Dp1187, which is derived from the X chromosome and retains a fully functional centromere. Several deletion derivatives of this minichromosome were recovered after irradiation and were used to map the centromere to a 420 kb region. One derivative chromosome of 620 kb was isolated by electrophoresis and gel extraction. Its DNA was fractionated and cloned and bacterial transposons were inserted into the cloned DNA [5]. Previous work [13] had demonstrated that the centromere of the *Drosophila* X chromosome is composed of arrays of two types of simple 5 bp satellites, AATAT and AAGAG, that are interrupted by five retrotransposons and an 'island' of complex DNA. Using primers specific to the inserted bacterial transposons or tagged primers that consisted of satellite sequence attached to non-homologous sequence, Sun *et al.* [5] were able to sample 31 kb of the AATAT and AAGAG satellites. This study [5] and previous work [13] showed that the arrays in the *Drosophila* centromere are highly similar - the AATAT sequence had 2.2% variation and AAGAG had only 0.3% variation in sequence - and that the repeats in each satellite are in the same orientation. Whereas transposon-like sequences previously found in *Drosophila* heterochromatin often consisted of scrambled clusters of different elements [5], the retrotransposons in the centromere of the X chromosome were intact. This suggested that they had recently been inserted into the genome or that their sequence is functionally conserved. The island of complex DNA was shown to be 39 kb long, including 16.2 kb of AT-rich sequence and retrotransposon-like elements that are arranged in blocks in different orientations. The beginning and end of this island contain a similar sequence, but are oriented in opposite directions - an arrangement analogous to fission yeast centromeres [5].

All of the elements identified by Sun *et al.* [5] are also found at non-centromeric locations in the *Drosophila* genome; the AATAT and AAGAG satellites are present in other but not all centromeres. Indeed, in *Drosophila* there are no DNA sequences that are located at every centromere, suggesting that primary centromeric sequence alone is neither sufficient nor necessary for centromere formation. The arrays identified in the X chromosome may therefore be merely permissive for centromere organization.

Insights from aberrant centromeres

Drosophila centromeres are unusual in being composed of sequences that are abundant elsewhere in the genome whereas in plants or mammals this is not the case under

normal circumstances. But there are some cases, in which the usual human centromeric sequences can be found at other chromosomal locations, where they display no detectable centromeric activity. For example, Robertsonian translocations, which are whole-arm rearrangements between acrocentric chromosomes can link two centromeres and yet the resulting chromosome is stably transmitted through mitosis and meiosis. Furthermore, *in situ* analysis using antibodies against essential kinetochore proteins, such as CENP-C, an essential component of the inner kinetochore plate, and CENP-A, the centromere-specific variant of histone H3 in human, has shown that only one of the two centromeric locations retains function [10].

Also in humans, rearranged chromosomes have been found that lack the region in which the centromere is usually present, and in these cases a new location has acquired centromeric activity. The new site ('neocentromere') has the usual hallmarks of a centromere - it forms a cytologically discernible constriction on the centromere and has kinetochore proteins bound [10,14]. The DNA sequences that gave rise to two of these neocentromeres were determined by immunoprecipitation of chromatin with antibodies against the centromeric histone H3 protein CENP-A. Analysis of the isolated DNA region showed that there are no elements in common between the two neocentromeres and normal centromeres [15,16].

Human artificial chromosomes can be generated by introducing α -satellite DNA arrays into cells [17], but not by introducing the DNA sequences of neocentromeres in a similar fashion [18]. Nevertheless, when the chromosome arms surrounding the neocentromere are truncated by insertion of telomere sequences, the resulting minichromosomes composed of the neocentromere DNA can be perpetuated through cell divisions [18]. This indicates that the satellite array of normal centromeres can direct *de novo* centromere formation, whereas the neocentromere DNA cannot. Nevertheless, the chromatin structure of the neocentromere appears to be stably maintained throughout the cell cycle. Because the primary sequences are not similar between neocentromeres and usual centromeres, the presence of neocentromeres suggests that centromere function may be regulated on an epigenetic level independent of DNA sequence.

Models of centromere determination

The importance of chromatin structure for centromere function is supported by the presence of species-specific variants of histone H3 found in the centromeric chromatin of all eukaryotes. The variants interact with the other core histone proteins, H2a, H2b and H4, to form a type of nucleosome that is present only at functional centromeres. It has been suggested that nucleosomes containing centromeric histone H3 are indispensable for centromere function and likely to serve as anchors for kinetochore formation. A model proposing

that correct spacing of centromeric and normal nucleosomes is required for centromere function is supported by recent data from *Drosophila* and human cells showing that stretched chromatin from centromeres is organized into blocks of centromeric nucleosomes interspersed between blocks of nucleosomes containing the normal core histone H3 [19]. This spacing may be facilitated by the satellites present at centromeres. Centromeric satellites from mammals and plants are approximately the length required to wrap around a nucleosome, and even in *Drosophila* multiples of the 5 bp satellites could add up to a unit of nucleosomal length.

Analysis of centromeric histone H3 in related species of mammals, flies, and plants has shown that the variants are highly similar to core histone H3 proteins in the regions that interact with the other histone proteins [20-22]. But in the region that is likely to contact the DNA strand centromeric histone H3 proteins appear to be under adaptive selection. Because the DNA sequence elements that are in contact with the centromeric H3 histones are divergent between species, it has been suggested that the centromeric histone H3 protein and the DNA are coevolving. Meiotic drive (a distortion of chromosome segregation) resulting from preferential positioning of 'stronger' centromeres to the egg during female meiosis might be the mechanism for this coevolution [20,21].

Many models for centromere determination predict that centromere function is independent of the underlying sequence. Such models are formulated to explain how nucleosomes containing centromeric histone H3 are maintained at all functional centromeres regardless of the DNA sequence with which they are associated. Spatial or temporal sequestration of the centromeres within nuclear compartments coupled to the availability of centromeric nucleosomes within these compartments or time phases has been suggested as a mechanism. Another model predicts that extant nucleosomes containing centromeric histone H3 are distributed to each strand during replication and subsequently used in post-replication recruitment of additional centromeric nucleosomes (for further discussion see [2]).

Models for centromere formation that do not rely on sequence must account for certain elements, such as the human α -satellite DNA and the *Arabidopsis* 178 bp repeat, that are present at every centromere in a normal karyotype within a given species. It seems that there must be mechanisms that homogenize repetitive elements such as centromeric repeats. For example, unequal crossing-over has been postulated to explain homogenization of α -satellite DNA within a chromosome [3], but there must also be a process that homogenizes the repeats between nonhomologous chromosomes. Unless the homogenization mechanism imposes constraints on the substrate sequence, changes to centromeric elements that become fixed in different populations would become randomly distributed in the absence of selection for sequence content. The analysis of *Arabidopsis*

and human centromeric satellites identified regions that were conserved among the various iterations, as well as regions that were more variable than average, implying that selection pressures act on the sequence of centromeric elements [7]. The observed non-random distribution of centromeric satellite DNA is not consistent with a model proposing complete irrelevance of sequence.

Some investigators [23,24] have raised the possibility that secondary structure or even higher order DNA structure could be a factor in determining centromere position and function. This idea may reconcile data showing irrelevance of primary sequence on the one hand with data that show conservation of DNA elements on the other. Conservation of DNA secondary structure allows for large variation in sequence, but does not exclude fine-tuning of the primary sequence, perhaps through coevolution with the domain of the centromeric histone H3 that associates with DNA. Similarly, epigenetic models of centromere formation, proposing regulation at the chromatin level, would not exclude fine-tuning of primary sequence. In either model, formation of a centromere with a new sequence would be allowed as long as the region permitted the proper higher-order DNA organization.

Data from neocentromere analysis do provide support for the idea that centromeres self-perpetuate without the need for a specific underlying sequence. In contrast, conservation of human and *Arabidopsis* centromeric repeat sequences suggests specific requirements at this level. Extreme models advocating a specific DNA element at centromeres versus no requirement at all will probably require a new synthesis. The means by which the position of the centromere on the chromosome is determined has yet to be resolved, but the recent elucidation of DNA sequence from the centromeres of various species is valuable information for making new predictive models. To determine the importance of various DNA elements found in or near the centromere, the mechanisms that drive evolution of centromeric DNA need to be clarified. For example, the lack of any centromeric elements common to all centromeres in *Drosophila* may be the result of a homogenization mechanism that is fundamentally different from the one that seems to function in mammals and plants. As additional centromeric sequences continue to become available from many different species, insights into the homogenization of sequences and their involvement in centromere formation will grow.

References

1. Henikoff S, Ahmad K, Malik HS: **The centromere paradox: stable inheritance with rapidly evolving DNA.** *Science* 2001, **293**:1098-1102.
2. Sullivan BA, Blower MD, Karpen GH: **Determining centromere identity: cyclical stories and forking paths.** *Nat Rev Genet* 2001, **2**:584-596.
3. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF: **Genomic and genetic definition of a functional human centromere.** *Science* 2001, **294**:109-115.

4. Guy J, Hearn T, Crosier M, Mudge J, Viggiano L, Koczan D, Thiesen H, Bailey JA, Horvath JE, Eichler EE, et al.: **Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p.** *Genome Res* 2003, **13**:159-172.
5. Sun X, Le HD, Wahlstrom JM, Karpen GH: **Sequence analysis of a functional *Drosophila* centromere.** *Genome Res* 2003, **13**:182-194.
6. Copenhaver GP, Nickel K, Kuromori T, Benito M, Kaul S, Lin X, Bevan M, Murphy G, Harris B, Parnell LD, et al.: **Genetic definition and sequence analysis of *Arabidopsis* centromeres.** *Science* 1999, **286**:2468-2474.
7. Hall SE, Kettler G, Preuss D: **Centromere satellites from *Arabidopsis* populations: maintenance of conserved and variable domains.** *Genome Res* 2003, **13**:195-205.
8. Heller R, Brown KE, Burgtorf C, Brown WR: **Mini-chromosomes derived from the human Y chromosome by telomere directed chromosome breakage.** *Proc Natl Acad Sci USA* 1996, **93**:7125-7130.
9. Murphy TD, Karpen GH: **Centromeres take flight: alpha satellite and the quest for the human centromere.** *Cell* 1998, **93**:317-320.
10. Warburton PE, Cooke CA, Bourassa S, Vafa O, Sullivan BA, Stetten G, Gimelli G, Warburton D, Tyler-Smith C, Sullivan KF, et al.: **Immunolocalization of CENP-A suggests a distinct nucleosome structure at the inner kinetochore plate of active centromeres.** *Curr Biol* 1997, **7**:901-904.
11. Haupt W, Fischer TC, Winderl S, Franz P, Torres-Ruiz RA: **The CENTROMERE1 (CEN1) region of *Arabidopsis thaliana*: architecture and functional impact of chromatin.** *Plant J* 2001, **27**:285-296.
12. Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, Birchler JA, Jiang J, Dawe RK: **Centromeric retroelements and satellites interact with maize kinetochore protein CENH3.** *Plant Cell* 2002, **14**:2825-2836.
13. Sun X, Wahlstrom J, Karpen GH: **Molecular structure of a functional *Drosophila* centromere.** *Cell* 1997, **91**:1007-1019.
14. Saffery R, Irvine DV, Griffiths B, Kalitsis P, Wordeman L, Choo KH: **Human centromeres and neocentromeres show identical distribution patterns of >20 functionally important kinetochore-associated proteins.** *Hum Mol Gen* 2000, **9**:175-185.
15. Lo AW, Craig JM, Saffery R, Kalitsis P, Irvine DV, Earle E, Magliano DJ, Choo KH: **A 330 kb CENP-A binding domain and altered replication timing at a human neocentromere.** *EMBO J* 2001, **20**:2087-2096.
16. Lo AW, Magliano DJ, Sibson MC, Kalitsis P, Craig JM, Choo KH: **A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA.** *Genome Res* 2001, **11**:448-457.
17. Grimes BR, Rhoades AA, Willard HF: **Alpha-satellite DNA and vector composition influences rates of human artificial chromosome formation.** *Mol Ther* 2002, **5**:798-805.
18. Saffery R, Wong LH, Irvine DV, Bateman MA, Griffiths B, Cutts SM, Cancilla MR, Cendron AC, Stafford AJ, Choo KH: **Construction of neocentromere-based human minichromosomes by telomere-associated chromosomal truncation.** *Proc Natl Acad Sci USA* 2001, **98**:5705-5710.
19. Blower MD, Sullivan BA, Karpen GH: **Conserved organization of centromeric chromatin in flies and humans.** *Dev Cell* 2002, **2**:319-330.
20. Malik HS, Vermaak D, Henikoff S: **Recurrent evolution of DNA-binding motifs in the *Drosophila* centromeric histone.** *Proc Natl Acad Sci USA* 2002, **99**:1449-1454.
21. Henikoff S, Malik HS: **Selfish drivers.** *Nature* 2002, **417**:227.
22. Yoda K, Ando S, Morishita S, Houmura K, Hashimoto K, Takeyasu K, Okazaki T: **Human centromere protein A (CENP-A) can replace histone H3 in nucleosome reconstitution *in vitro*.** *Proc Natl Acad Sci USA* 2000, **97**:7266-7271.
23. Koch J: **Neocentromeres and alpha satellite: a proposed structural code for functional human centromere DNA.** *Hum Mol Gen* 2000, **9**:149-154.
24. Grady DL, Ratliff RL, Robinson DL, McCanlies EC, Meyne J, Moyzis RK: **Highly conserved repetitive DNA sequences are present at human centromeres.** *Proc Natl Acad Sci USA* 1992, **89**:1695-1699.