

Minireview

## What's in a centromere?

Jonathan C Lamb, James Theuri and James A Birchler

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

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

Published: 17 August 2004

Genome **Biology** 2004, **5**:239

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2004/5/9/239>

© 2004 BioMed Central Ltd

### Abstract

The complete sequence of rice centromere 8 reveals a small amount of centromere-specific satellite sequence in blocks interrupted by retrotransposons and other repetitive DNA, in an arrangement that is strikingly similar in overall size and content to other centromeres of multicellular eukaryotes.

Shakespeare's Juliet posed the question "What's in a name?" to explore the connotations that a single word can hold. The name 'centromere' conjures many ideas from classical biology, but genome projects have had a difficult time defining exactly what is present at the portion of the chromosome responsible for microtubule association and segregation at mitosis and meiosis. In humans [1], *Arabidopsis thaliana* [2], and other model organisms, centromeres appear to contain a core of megabase-sized arrays of a single element (or, in flies, several arrays of a small number of different microsatellite elements [3]). Near the center of this core the repeated elements are arranged in a nearly perfect array, while near the edges the uniformity decreases and the arrays are interspersed by various repetitive elements. Because of the size and uniformity of the cores, they have been impossible to sequence with standard techniques and so have remained as gaping holes of unsequenced DNA in the otherwise well-defined model-organism genomes obtained by various international efforts.

As in other model organisms, each centromere of members of the grass family (including rice and maize) contains large tandem arrays of a species-specific centromeric repeat (CentO in rice [4]; CentC in maize [5]). Fluorescent *in situ* hybridization (FISH) using centromere-specific satellite sequence as a probe reveals that their copy number among different rice and maize centromeres varies considerably - almost 30-fold in rice. Because the copy number of the centromeric satellite in rice chromosome 8 is very low, two

groups - Nagaki *et al.* [6] and Wu *et al.* [7] - were able to sequence the entire centromeric region using standard techniques involving bacterial artificial chromosomes (BACs). The two groups screened BAC libraries, created as part of the ongoing effort to sequence the rice genome, with centromere-specific elements as probes, and then 'walked' from BAC to adjacent BAC, by virtue of overlapping sequence at their ends, so as to form a minimal tiling path, or contig, spanning the genetically defined centromeric region. Their work has resulted in the first complete sequence of a normal centromere from a multicellular organism.

Because CentO is found as a tandem array of repeats and such repetitive DNA tends to be unstable when maintained in *Escherichia coli* (which is used to replicate BACs), Nagaki *et al.* [6] used cytological approaches to confirm the location and completeness of their centromere-containing contig. First, they used BACs that flanked the CentO region from the minimal contig of centromere 8 as FISH probes on spreads of rice pachytene chromosomes, to confirm that the contig included the entire CentO-containing region. Next, they performed 'fiber FISH', probing the same chromosomes in the form of stretched DNA fibers, again using the BACs from the minimal contig as probes and with CentO as a probe, to show that the predicted tiling path reflected the correct physical arrangement of the BACs around the centromere. This procedure also showed that the complete cytologically detectable CentO-containing region was contained in one BAC. Measuring the length of the CentO array in parallel on stretched genomic DNA and on stretched

BAC fibers then confirmed that the CentO array contained in the BAC was intact. Nagaki *et al.* [6] then sequenced 12 BACs containing 1.65 Mb in total, spanning the CentO tract and extending into both the long and short arms of the chromosome. Wu *et al.* [7] independently obtained 1.97 Mb of sequence from the same centromeric region that includes the 1.65 Mb from the Nagaki *et al.* [6] study. They sequenced multiple BACs covering the CentO tracts to confirm the size and integrity of the CentO arrays.

In contrast to human [1] and *Arabidopsis* [2] centromeres, each of which has a large core of nearly homogeneous satellite sequence, the tandem arrays of centromeric satellite in rice chromosome 8 are frequently interrupted by insertions of a particular family of retroelements of the long terminal repeat (LTR) type, called CRR in rice. Using FISH, retroelements of this type can be seen only at the centromere in cytological preparations from numerous grass species [8]. Nagaki *et al.* [6] report that rice centromere 8 contains only 41 kilobases of CentO sequence, arranged as a cluster of three arrays of CentO separated by full and partial CRR elements. One of the arrays is oriented in the opposite direction to the other two. There is also approximately 2.8 kb of CentO that is separated from the main site by over 700 kb of sequence that includes repetitive elements and active genes. Analyzing yet another rice centromere, Zhang *et al.* [9] defined a BAC contig that spans rice centromere 4 and reported sequencing efforts from the single BAC that hybridizes to the CentO element of this centromere. This BAC contained a 124 kb 'core' region made up of 379 copies of CentO arranged in 18 tracts in different orientations interrupted by various repetitive sequences, including CRR elements and other LTR retroelements and repeats not specific to centromeres.

Because many repetitive elements, including the centromeric unit, are highly divergent between maize and oat, it is possible to use FISH to distinguish the centromeres of maize chromosomes that have been artificially transferred to an oat background. Using this type of material, Jin *et al.* [10] examined the DNA arrangement along stretched chromatin fibers from individual maize centromeres and found that tracts of the maize centromere repeat element CentC were interspersed with CRM, the maize homolog of CRR, and unknown sequences. This pattern is consistent with the results of the sequencing efforts for rice centromeres 4 and 8 as well as other rice [4] and maize [11] BACs that contain centromeric satellite sequence. Taken together, these results suggest a consistent pattern of DNA organization at grass centromeres consisting of tracts of centromeric satellite interspersed with various repetitive elements, especially centromere-specific retrotransposons.

### Centromeric chromatin structure

Centromeric chromatin includes a centromere-specific histone H3 variant (CenH3) that is incorporated into nucleosomes

underlying the kinetochore. These nucleosomes remain a part of the chromatin throughout the cell cycle and are essential to both meiotic and mitotic cell divisions [12]. Although it has not been established that CenH3 alone determines centromere identity, the sequence of a complete centromere should at the least include the entire region that is wound around nucleosomes containing CenH3. Nagaki *et al.* [6] used anti-CenH3 antibodies to immunoprecipitate chromatin (ChIP) comprising DNA bound to CenH3-containing nucleosomes, confirming that CenH3 is associated with both the CentO repeats and the CRR family of retrotransposons. Primer pairs were designed that would amplify sequences scattered along the length of the centromere 8 contig, and these were used to sample the immunoprecipitated DNA using a process called ChIP-PCR, showing that the CenH3-containing region is approximately 750 kb and does not include the small 2.8 kb cluster of CentO that is separated from the three main arrays. Although the region immediately around the CentO tracts for both centromeres 4 and 8 consists entirely of repetitive elements, the 750 kb CenH3-binding domain of rice centromere 8 included 14 putative non-retroelement open reading frames (ORFs), including 4 that were shown to be expressed by reverse-transcriptase-coupled PCR [6]. This observation is reminiscent of human neocentromeres - chromosomal regions that have newly acquired centromere activity. Neocentromeres have also been shown to harbor expressed genes [13], and the rice finding shows that the chromatin structure of both plant and mammalian CenH3-binding domains is open and accessible to the transcriptional machinery.

In addition to binding microtubules, centromeres have other functions, including sister chromatid cohesion and preventing microtubules from both poles attaching to the same chromatid. These other functions may be located in domains with distinct chromatin structures [14,15]. To examine the chromatin structure of rice centromere 8, Nagaki *et al.* [6] used ChIP-PCR with antibodies against two different covalent modifications of the canonical H3 histone protein (rather than the centromere-specific CenH3): dimethylation on lysine 9 (dimethyl-K9), which has been shown to be enriched in heterochromatic regions, and dimethyl-K4, which is present in euchromatic portions of the chromosome. The region associated with dimethyl-K9 H3 spans approximately 1.2 Mb and includes all of the CentO arrays. Because this region covers the entire CenH3-binding region (around 750 kb), the authors [6] postulated that CenH3-containing and dimethyl-K9 H3-containing nucleosomes are interspersed and that the position of these nucleosomes is dynamic, so that a population of cells may have the same DNA sequence interacting with both types of nucleosome. Indeed, the interspersed nature of these two types of nucleosome has been observed on stretched chromatin fibers of both *Drosophila* [16] and maize [10]. Immunoprecipitation with antibodies against dimethyl-K4 H3 was limited to the edges of the contig flanking the dimethyl-K9 H3 region [6].

Nakagi *et al.* [6] and Wu *et al.* [7] chose the rice centromere with the fewest copies of CentO for their sequencing efforts. Although this approach allowed an achievement not otherwise possible, the sequence obtained may not be representative of centromeres of other rice chromosomes and of some other model organisms, because of its unusually small size. Despite the reduced copy number of CentO, however, it should not be concluded that the functional domain of rice centromere 8 is smaller than other centromeres. In humans [1,15] and *Arabidopsis* [17], which have centromeres made up of numerous copies of satellite sequences, the CenH3-binding region covers only a portion of the central core of the centromeric satellite array. In rice and maize, ChIP analysis shows that the majority of centromeric satellite is not associated with CenH3 [6,18]. Cytological observation of maize chromosomes shows that while the amount of centromeric satellite varies extensively among centromeres, the amount of CenH3 remains relatively constant [18]. Although it is difficult to determine the precise sizes of centromeres (because they are composed of large arrays of satellite), observations of fragmented centromeres arising from rare events [19,20] have allowed the lengths of some centromeres to be estimated. The rice centromere 8 CenH3-binding domain is consistent with the reported minimal sizes of other centromeres including the maize B chromosome (around 500 kb) [19], the human Y chromosome (not more than 500 kb) [20] and a *Drosophila* minichromosome (around 420 kb) [3], suggesting a common size requirement. Additional requirements for effective passage through meiosis may necessitate additional chromatin configurations and could explain the excess sequences that are present at many centromeres and whose function is not yet apparent. For example, *Drosophila* minichromosomes that lack sequences adjacent to the essential core show reduced meiotic transmission [21].

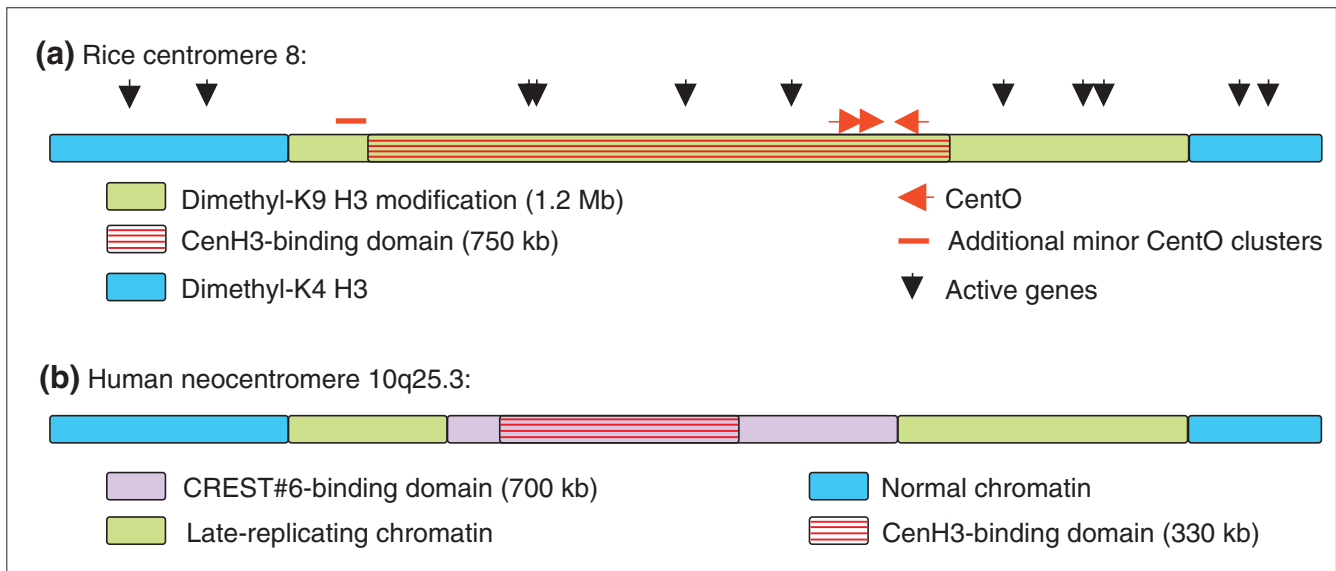
Because human neocentromeres are not composed of repetitive DNA, immunoprecipitation analysis is possible and a direct comparison of chromatin states between neocentromeres and rice centromere 8 is revealing (Figure 1). Human neocentromere 10q25.3 contains a 330 kb CenH3-binding region within a 700 kb domain that can be precipitated by an immune serum containing antibodies against numerous centromeric proteins [22]. These domains are flanked by regions that replicate late in the cell cycle. In total, the region altered by adoption of centromere identity is approximately 1.4–2 Mb, similar in size to the dimethyl-K9 H3-bound region of rice centromere 8. Although dimethyl-K9 H3 antibodies were not used in the study by Lo *et al.* [22], the delayed replication of this region probably reflects the presence of dimethyl-K9 H3 or a similar heterochromatic structure. The similarities in chromatin domain size and arrangement between rice centromere 8 and the human neocentromere (Figure 1) suggest that rice and human have similar chromatin requirements for functional centromeres, including a requirement for flanking heterochromatin that is

shared with *Drosophila* [21]. Additional chromatin domains have been identified within the human neocentromere, including a domain that binds the centromere protein CenP-H and another enriched for chromosomal scaffold/matrix attachment regions [13]. With the availability of the complete sequence for rice centromere 8, similar analysis can now be performed for this centromere and the findings compared to the human neocentromere results.

### Centromere evolution

Taking their cue from the analysis of human neocentromeres, Nagaki *et al.* [6] suggest that the presence of active genes indicates that rice centromere 8 is relatively 'young', evolutionarily, and may have arisen from a neocentromerization event. In humans, neocentromerization is usually initiated by a significant chromosomal rearrangement, such as a translocation that produces an acentric fragment, but neocentromeres can also arise spontaneously in an intact karyotype within a single generation [23]. Consistent with the hypothesis that rice centromere 8 is a relatively new centromere, the amount of CentO it contains is small and sequence analysis of the LTRs of the CRR-class retroelements shows that they have recently inserted into the region. But because the CenH3-binding domain has not been determined for other rice centromeres, the possibility that active genes and frequent retrotransposon insertions are a common feature of grass centromeres cannot yet be ruled out. Also, certain maize centromeres in some lines have virtually undetectable amounts of CentC [5] while homologous centromeres of other lines contain numerous copies of the centromeric satellite and are present at the same genetic location [24]. This suggests that aside from neocentromere formation, mechanisms that reduce satellite copy number could account for the small amount of CentO at rice centromere 8. An example of such a reduction is seen in a study of human cells in which centromere 21 spontaneously lost a specific portion of the centromeric repeat array at a measurable frequency [25].

Although rice centromeres 4 and 8 do not contain massive arrays of CentO, other rice centromeres do (for example, centromeres 1 and 11 [4]), indicating that forces that expand centromeric DNA elements are active in rice. Despite the involvement of epigenetic factors that determine centromere identity, certain DNA sequences seem more suited to life in a centromere than others [26]. In chromosomes that contain very few copies of centromeric satellite, flanking sequences, including genes, will be incorporated into the centromere and forced to conform to local centromeric chromatin requirements. Introduction and subsequent expansion of more suitable sequences would push these sequences away from the active centromere core. Such changes would be strongly selected for, especially if the misexpression of genes incorporated into centromeric regions is detrimental to individual fitness and regular expression could be restored by the expansion of centromere repeats. This type of selection

**Figure 1**

Similarities between a rice centromere and a human neocentromere. **(a)** Rice centromere 8 contains an approximately 750 kb CenH3-binding domain that is positioned off-center inside an approximately 1.2 Mb domain where H3 is dimethylated at the lysine that is residue 9 (dimethyl-K9 H3). Active genes are found in and around the CenH3-binding domain. Rice-specific centromeric repeats (CentO) are indicated. **(b)** Human neocentromere 10q25.3 contains an approximately 330 kb CenH3-binding domain contained in an approximately 700 kb region that can be precipitated with CREST#6 antibodies and is flanked by late-replicating regions. Shading is used to indicate potentially analogous regions, and the sizes shown are approximate.

pressure on new centromeres to expand would complement other forces that could drive centromere satellite expansion, such as competition among centromeres during female meiosis [27].

The two rice centromere 8 sequences derived from Nipponbare varieties by Nagaki *et al.* [6] and Wu *et al.* [7] are essentially identical to each other except for the size of the CentO arrays: 38.2 kb versus 68.5 kb of CentO contained in the major cluster for Nagaki *et al.* [6] and Wu *et al.* [7], respectively. Despite the large differences in satellite copy number, the relative orientation of the tandem arrays is the same for the two groups' sequencing efforts, and the CRR elements that separate the three arrays are identical. Because both groups took steps to confirm that the size of their tracts was accurate, it is unlikely that rearrangements resulting from the cloning process account for the differences between the two groups' findings. Instead, the sequencing efforts probably captured ongoing changes in centromeric satellite copy number and underscore how rapidly such change can occur.

In humans, L1 retroelement insertions are scarce in the heart of the centromeric satellite arrays but are more common in the divergent repeat units found on the periphery. Insertions located at some distance from each other are found to be either present or absent as a group, a phenomenon that can be explained by intra-chromosomal recombination between L1 elements simultaneously removing several

elements and the intervening satellites [28]. The presence of a centromere-specific LTR retroelement has thus far only been observed in the grasses and, in contrast to human L1 retroelements, the grass centromeric retroelements show a preference for, and frequent insertion into, centromeric regions including satellite arrays. Thus, an accelerated process of continual transposition and subsequent rearrangements coupled with satellite expansion may explain the differences between human and grass centromeres, the latter of which contain clusters of centromeric satellite organized in fragmented arrays with different orientations and abundant solo LTR elements.

In conclusion, the completion of the first sequences of a centromere from a multicellular eukaryote thus indicates that the necessary regions span hundreds of kilobases and contain a specific repeat. Some of this region is organized around nucleosomes containing CenH3 or histone H3 dimethylated at lysine 9. As other sequences become available, further generalizations will emerge to answer the question from 'Juliet of the genome', "What's in a centromere?"

## References

- Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF: **Genomic and genetic definition of a functional human centromere.** *Science* 2001, **294**:109-115.
- Copenhaver GP, Nickel K, Kuromori T, Benito MI, 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.

3. Sun X, Le HD, Wahlstrom JM, Karpen GH: **Sequence analysis of a functional *Drosophila* centromere.** *Genome Res* 2003, **13**:182-194.
4. Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, Gu M, Blattner FR, Jiang J: **Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon.** *Plant Cell* 2002, **14**:1691-1704.
5. Ananiev EV, Phillips RL, Rines HW: **Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions.** *Proc Natl Acad Sci USA* 1998, **95**:13073-13078.
6. Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang J: **Sequencing of a rice centromere uncovers active genes.** *Nat Genet* 2004, **36**:138-145.
7. Wu J, Yamagata H, Hayashi-Tsugane M, Hijishita S, Fujisawa M, Shibata M, Ito Y, Nakamura M, Sakaguchi M, Yoshihara R, et al.: **Composition and structure of the centromeric region of rice chromosome 8.** *Plant Cell* 2004, **16**:967-976.
8. Jiang J, Nasuda S, Dong F, Scherrer CW, Woo SS, Wing RA, Gill BS, Ward DC: **A conserved repetitive DNA element located in the centromeres of cereal chromosomes.** *Proc Natl Acad Sci USA* 1996, **93**:14210-14213.
9. Zhang Y, Huang Y, Zhang L, Li Y, Lu T, Lu Y, Feng Q, Zhao Q, Cheng Z, Xue Y, et al.: **Structural features of the rice chromosome 4 centromere.** *Nucleic Acids Res* 2004, **32**:2023-2030.
10. Jin W, Melo JR, Nagaki K, Talbert PB, Henikoff S, Dawe RK, Jiang J: **Maize centromeres: organization and functional adaptation in the genetic background of oat.** *Plant Cell* 2004, **16**:571-581.
11. Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones KM, Dawe RK, et al.: **Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres.** *Genetics* 2003, **163**:759-770.
12. Sullivan BA, Blower MD, Karpen GH: **Determining centromere identity: cyclical stories and forking paths.** *Nat Rev Genet* 2001, **2**:584-596.
13. Saffery R, Sumer H, Hassan S, Wong LH, Craig JM, Todokoro K, Anderson M, Stafford A, Choo KH: **Transcription within a functional human centromere.** *Mol Cell* 2003, **12**:509-516.
14. Bjerling P, Ekwall K: **Centromere domain organization and histone modifications.** *Braz J Med Biol Res* 2002, **35**:499-507.
15. Spence JM, Critcher R, Ebersole TA, Valdivia MM, Earnshaw WC, Fukagawa T, Farr CJ: **Co-localization of centromere activity, proteins and topoisomerase II within a subdomain of the major human X alpha-satellite array.** *EMBO J* 2002, **21**:5269-5280.
16. Blower MD, Sullivan BA, Karpen GH: **Conserved organization of centromeric chromatin in flies and humans.** *Dev Cell* 2002, **2**:319-330.
17. Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang J: **Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres.** *Genetics* 2003, **163**:1221-1225.
18. 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.
19. Kaszas E, Birchler JA: **Meiotic transmission rates correlate with physical features of rearranged centromeres in maize.** *Genetics* 1998, **150**:1683-1692.
20. Tyler-Smith C, Oakey RJ, Larin Z, Fisher RB, Crocker M, Affara NA, Ferguson-Smith MA, Muenke M, Zuffardi O, Jobling MA: **Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes.** *Nat Genet* 1993, **5**:368-375.
21. Murphy TD, Karpen GH: **Localization of centromere function in a *Drosophila* minichromosome.** *Cell* 1995, **82**:599-609.
22. 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.
23. Amor DJ, Bentley K, Ryan J, Perry J, Wong L, Slater H, Choo KH: **Human centromere repositioning "in progress".** *Proc Natl Acad Sci USA* 2004, **101**:6542-6547.
24. Kato A, Lamb JC, Birchler JA: **Chromosome painting in maize using repetitive DNA sequences as probes for somatic chromosome identification.** *Proc Natl Acad Sci USA*, in press.
25. Lo AW, Liao GC, Rocchi M, Choo KH: **Extreme reduction of chromosome-specific alpha-satellite array is unusually common in human chromosome 21.** *Genome Res* 1999, **9**:895-908.
26. Lamb JC, Birchler JA: **The role of DNA sequence in centromere formation.** *Genome Biol* 2003, **4**:214.
27. Henikoff S, Malik HS: **Centromeres: selfish drivers.** *Nature* 2002, **417**:227.
28. Laurent AM, Puechberty J, Roizes G: **Hypothesis: for the worst and for the best, L1s retrotransposons actively participate in the evolution of the human centromeric aliphoid sequences.** *Chromosome Res* 1999, **7**:305-317.