

# Retrotransposon insertion targeting: a mechanism for homogenization of centromere sequences on nonhomologous chromosomes

James A. Birchler<sup>1,3</sup> and Gernot G. Presting<sup>2</sup>

<sup>1</sup>Division of Biological Sciences, University of Missouri at Columbia, Columbia, Missouri 65211, USA; <sup>2</sup>Department of Molecular Biosciences and Bioengineering, University of Hawaii, Honolulu, Hawaii 96822, USA

**The centromeres of most eukaryotic organisms consist of highly repetitive arrays that are similar across non-homologous chromosomes. These sequences evolve rapidly, thus posing a mystery as to how such arrays can be homogenized. Recent work in species in which centromere-enriched retrotransposons occur indicates that these elements preferentially insert into the centromeric regions. In two different *Arabidopsis* species, a related element was recognized in which the specificity for such targeting was altered. These observations provide a partial explanation for how homogenization of centromere DNA sequences occurs.**

The centromere paradox notes that despite a strong selection for the functional components of the kinetochore across eukaryotes, the sequences present at the centromeres of chromosomes evolve quite rapidly (Henikoff et al. 2001). Indeed, centromere sequences can become inactive, or kinetochores can form over entirely unique sequences. These observations have led to the idea that there is an epigenetic component to centromere function in most eukaryotes so that the exact sequence is not critical for the establishment of the kinetochore at the same site in each cell division (Karpen and Allshire 1997; Han et al. 2006; Birchler et al. 2011). The basal molecule in association with DNA is a variant of histone H3, CENH3, which is present in nucleosomes of active centromeres. The N terminus of CENH3 evolves rapidly as well, which has led to the idea that an antagonism between this protein sequence and the DNA sequence drives the rapid evolution (Henikoff et al. 2001).

In most eukaryotes, the DNA underlying the kinetochore sites is a highly repetitive array that includes a centromeric satellite ranging from 150 to 180 base pairs

(bp) in length, depending on the species. This satellite is similar throughout the arrays and usually among non-homologous chromosomes. Nevertheless, there is sequence diversity within arrays that is quite common. In many plant species, there is also a centromere-enriched retrotransposon that is an integral part of the array (Presting et al. 1998).

## The homogenization conundrum

With such a rapid evolution, an interesting question arises as to how a similar sequence repeat can become established within and between centromeres. Within an array, one might imagine that unequal crossing over might homogenize an array, although such events are extremely rare or nonexistent in centromeric regions. Indeed, if crossing over were allowed to occur between centromere arrays, the size might change dramatically and produce detrimental effects, or if the similar sequences on non-homologous chromosomes were to enter into recombination, a translocation would be generated between the two chromosomes, which would lead to semisterility in many species. Thus, crossing over is suppressed in these regions.

It is perhaps possible that gene conversion events between homologs can act to exchange sequences between homologs without crossing over (Shi et al. 2010). Such a mechanism could act to homogenize the sequences between homologs but leaves open the question of how homogenization across nonhomologous chromosomes occurs. An understanding of how rapid evolution of related sequences on different chromosomes in specific locations can occur is needed.

## Centromeric retrotransposon targeting in *Arabidopsis*

Recent insight into a partial explanation involves the fact that members of certain retrotransposon families primarily target the centromere region for insertion and become a major component of the centromere array. Tsukahara et al. (2012) have examined this issue with a study of targeting of a retrotransposon in *Arabidopsis*. The preferential insertion of these elements into the centromeric

[Keywords: centromere; tandem repeat; retrotransposon; evolution]

<sup>3</sup>Corresponding author.

E-mail birchlerj@missouri.edu.

Article is online at <http://www.genesdev.org/cgi/doi/10.1101/gad.191049.112>.

regions appears to be a property of the transposons themselves and can contribute to the rapid evolution of centromere sequences.

In *Arabidopsis thaliana*, the centromeric regions consist of a satellite repeat that is canonically 178 bp in length. A particular retrotransposon, Athila, is also commonly found in these regions, but mobile copies have not been identified. In a related species, *Arabidopsis lyrata*, a retrotransposon with similarity to a low-copy dispersed version in *thaliana*, ATCOPIA93, is present at a much higher copy number and concentrated at the centromeric regions. The centromeric satellite is considerably diverged (30%) between the two species. The difference in satellite sequence and centromere-clustered retrotransposons illustrates how quickly the composition of centromeres can change over evolutionary time.

### Targeting: property of the element or cellular environment?

The fact that this retrotransposon, although similar between the two species, has a distinct genomic distribution might be attributed to a different cellular and genetic environment that affects the behavior of this element in the two species. Alternatively, it might be the case that the two versions of this element have diverged integration mechanisms that target the centromeric regions.

To examine aspects of this presumed targeting process, Tsukahara et al. (2012) identified a copy of this transposon of *A. lyrata* (Tal1) that appears to be recently transposed based on the high sequence similarity of the 5' and 3' long terminal repeats (LTRs), which are identical upon insertion. This copy was then transformed into *A. thaliana*. Its expression there could be confirmed by detection of RNA specific to this element. Southern analysis revealed new insertions of this element in the *A. thaliana* genome, which restriction analysis suggested might be in the centromeric satellite arrays. To confirm this conjecture, whole-genome sequencing was performed. Sequences flanking the newly inserted elements were highly biased toward centromere satellite arrays, suggesting that targeting might be mediated by either recognition of the satellite sequences or CENH3.

Because only the central portion of the centromeric satellite repeat in *A. thaliana* is actually covered by the centromeric histone CENH3 (Shibata and Murata 2004), and because this region also is hypomethylated relative to the flanking satellite DNA (Zhang et al. 2008), it would be interesting to map the new insertions with respect to the CENH3 footprint either by CENH3 chromatin immunoprecipitation (ChIP) or using cytogenetic methods.

Interestingly, within *A. thaliana*, ATCOPIA93 can be mobilized in a background of the mutation *ddm1*, which reduces the amount of DNA methylation in the genome. When these mobilizations occur, the sites of integration appear to be random within the genome and certainly not clustered in centromeric regions.

### Target switching

To address the ancestral state, differently diverged lineages or clusters of the element were examined for their

flanking insertion sequences in *A. lyrata*. Most of these insertions were flanked by centromeric sequences, suggesting that the ancestral state of this transposon was for insertion into centromeric arrays. Presumably, this property was lost by ATCOPIA93 in *A. thaliana*. One should note, however, that the majority of known retrotransposon insertions in *Arabidopsis* genome are near, but not in, the centromere (Peterson-Burch et al. 2004).

Interestingly, the centromere retrotransposon (CR) elements of the grasses also contain some lineages that insert preferentially in the centromere and others that appear to have lost this ability (Sharma and Presting 2008). For example, CRM1 and CRM2 of maize are highly enriched in chromatin fractions immunoprecipitated with anti-CENH3 antibody (Wolfgruber et al. 2009). Moreover, these elements have been targeting the centromeres over millions of years, thus providing a tool to determine the historical location of maize centromeres as deduced from the genomic sequence (Wolfgruber et al. 2009). These footprints illustrate that the centromere 5 of maize has shifted position over time. Because the DNA sequence of the functional maize centromeres contains many sequences other than centromeric satellites (Wolfgruber et al. 2009), targeting of these elements is likely to be sequence-independent.

Nevertheless, CR elements belong to the gypsy class of elements and are closely related to chromoviruses, which are retrotransposons whose integrase contains a chromodomain that has been shown to target insertions. Extensive and elegant work has thus far failed to identify the interacting partner of the CR motif, although the CR motif (and group I and group II chromodomains of the chromoviruses) target yellow fluorescent protein (YFP) fusion constructs to the heterochromatic regions of *Arabidopsis* (Gao et al. 2008). Moreover, Gao et al. (2008) showed convincingly that the chromodomain of the MAGGY element interacts with the chromatin modification of histone H3 methyl-K9. Thus, although it seems likely that the interacting partner of the centromere-specific CRM elements and perhaps that of TAL1 is the centromere-specific CENH3 histone variant, this remains to be conclusively proven.

The sequence of the integrase gene of Tal1 and ATCOPIA93 are quite similar. Presumably, this gene is responsible for the differential recognition of sites of insertion, which, as noted above, might involve chromatin states. Tsukahara et al. (2012) suggest that a study involving chimeras between the two integrase genes from the two species transformed back into *Arabidopsis* would help determine whether this is the case.

### The role of retrotransposons at plant centromeres

It is unclear at this time whether retrotransposons that target centromeres do so only because these regions provide them with a safe genomic environment or whether these elements actually play an active role in centromere function. Many retrotransposons have acquired the ability to target nongenic regions, presumably to reduce the likelihood of causing lethal knockout mutations in their host. Centromeres contain few, if any, genes and thus provide a relatively "safe" target. Also, it is conceivable that the centromeric

environment may prevent recombination-mediated loss of retrotransposons, although there is no evidence that the average date of centromeric insertions is older than that of chromosome arm insertions. On the other hand, very few retrotransposon lineages have acquired the ability to integrate specifically into centromere regions, so this must either be difficult to do or carry some disadvantages.

### Targeting as a tool?

Tsukahara et al. (2012) also suggest that the targeting of this element to centromere regions might provide the basis for certain types of chromosomal manipulations. For example, if site-specific recombination cassettes such as *lox* sequences from the Cre-*lox* system were targeted to centromeric regions, then the introduction of Cre recombinase into lines carrying *lox* sites in different centromeres would induce chromosomal exchanges. Moreover, chromosomal truncations might be able to be generated if telomere arrays were targeted to centromeric regions. The introduction of telomere arrays has been shown to catalyze chromosomal truncation at the sites of insertion (Yu et al. 2007; Nelson et al. 2011; Teo et al. 2011). By using a targeting system, the position of the truncations could be manipulated. Telomere-mediated chromosomal truncations have been used to produce engineered minichromosomes when additional genes are included in the truncating transgene (Yu et al. 2007). If such a targeting technology could become a reality, then cleavage of all of the flanking sequences to a centromere could be achieved.

### Concluding remarks

The finding that centromere-enriched retrotransposons have a targeting function that preferentially fosters integration into centromeric regions provides a partial solution to the question of how the rapid evolution of centromere arrays can be achieved across nonhomologous chromosomes. Clearly, elements that can target their integration to centromeres can become inserted at the centromere sites on each of the different chromosomes in the karyotype, and this fact can account for the similarity across chromosomes. Indeed, the fact that, in related species, the targeting function can change provides an explanation for why the retroelements are accumulated at centromeres in one species but not in another. What is not yet explained is how the homogenization of the centromere satellites occurs. These arrays also change rapidly over evolutionary time. However, a mechanism for how they are accumulated on nonhomologous chromosomes awaits.

### Acknowledgments

This research was supported by National Science Foundation grants DBI 0922703 and DBI 0701297.

### References

Birchler JA, Gao Z, Sharma A, Presting GG, Han F. 2011. Epigenetic aspects of centromere function in plants. *Curr Opin Plant Biol* **14**: 217–222.

- Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res* **18**: 359–369.
- Han F, Lamb JC, Birchler JA. 2006. High frequency of centromere inactivation resulting in stable dicentric chromosomes of maize. *Proc Natl Acad Sci* **103**: 3238–3243.
- Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science* **293**: 1098–1102.
- Karpen GH, Allshire RC. 1997. The case for epigenetic effects on centromere identity and function. *Trends Genet* **13**: 489–496.
- Nelson AD, Lamb JC, Kobrossly PS, Shippen DE. 2011. Parameters affecting telomere-mediated chromosomal truncation in *Arabidopsis*. *Plant Cell* **23**: 2263–2272.
- Peterson-Burch BD, Nettleton D, Voytas DF. 2004. Genomic neighborhoods for *Arabidopsis* retrotransposons: A role for targeted integration in the distribution of the Metaviridae. *Genome Biol* **5**: R78. doi: 10.1186/gb-2004-5-10-r78.
- Presting GG, Malysheva L, Fuchs J, Schubert I. 1998. A Ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant J* **16**: 721–728.
- Sharma A, Presting GG. 2008. Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity. *Mol Genet Genomics* **279**: 133–147.
- Shi J, Wolf SE, Burke JM, Presting GG, Ross-Ibara J, Dawe RK. 2010. Widespread gene conversion in centromere cores. *PLoS Biol* **8**: e1000327. doi: 10.1371/journal.pbio.1000327.
- Shibata F, Murata M. 2004. Differential localization of the centromere-specific proteins in the major centromeric satellite of *Arabidopsis thaliana*. *J Cell Sci* **117**: 2963–2970.
- Teo CH, Ma L, Kapusi E, Hensel G, Kumlehn, Schubert I, Houben A, Mette MF. 2011. Induction of telomere-mediated chromosomal truncation and stability of truncated chromosomes in *Arabidopsis thaliana*. *Plant J* **68**: 28–39.
- Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, Toyoda A, Fujiyama A, Tarutani Y, Kakutani T. 2012. Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*. *Genes Dev* (this issue). doi: 10.1101/gad.183871.111.
- Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, Shi J, Gao Z, Han F, Lee H, Xu R, et al. 2009. Maize centromere structure and evolution: Sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genet* **5**: e1000743. doi: 10.1371/journal.pgen.1000743.
- Yu W, Han F, Gao Z, Vega JM, Birchler JA. 2007. Construction and behavior of engineered minichromosomes in maize. *Proc Natl Acad Sci* **104**: 8924–8929.
- Zhang W, Lee HR, Koo DH, Jiang J. 2008. Epigenetic modification of centromeric chromatin: Hypomethylation of DNA sequences in the CENH3-associated chromatin in *Arabidopsis thaliana* and maize. *Plant Cell* **20**: 25–34.