nature portfolio

https://doi.org/10.1038/s41586-024-08319-7

Supplementary information

Centrophilic retrotransposon integration via CENH3 chromatin in *Arabidopsis*

In the format provided by the authors and unedited

Centrophilic retrotransposon integration via CENH3 chromatin in Arabidopsis

Sayuri Tsukahara, Alexandros Bousios, Estela Perez-Roman, Sota Yamaguchi, Basile Leduque, Aimi Nakano, Matthew Naish, Akihisa Osakabe, Atsushi Toyoda, Hidetaka Ito, Alejandro Edera, Sayaka Tominaga, Juliarni, Kae Kato, Shoko Oda, Soichi Inagaki, Zdravko Lorković, Kiyotaka Nagaki, Frédéric Berger, Akira Kawabe, Leandro Quadrana, Ian Henderson & Tetsuji Kakutani

Supplementary Information

Table of Contents

Supplementary Discussion 1, 2, 3, 4 2

References for Supplementary Discussion 4

Supplementary Figure 1 6

Supplementary Figure 2 7

Supplementary Figure 3 8

Supplementary Figure 4 9

Supplementary Figure 5 10

Supplementary Table 1 11

Supplementary Table 2 12

Supplementary Tables 3, 4 13

(Supplementary Tables 3 and 4 are in separate files)

Supplementary Discussion 1 | *ALE4*-related centrophilic elements are prevalent in other plant species.

A. lyrata genome contains both centrophilic and centrophobic *ALE* elements of *Ty1/copia* type, while centrophilic *ALE* is almost absent from *A. thaliana* genome. In order to see if centrophilic *ALE* are prevalent in other plant species, we examined three other species, *Barbarea vulgaris*, *Raphanus sativus*, and *Eutrema japonicum* (Extended Data Fig. 2). Genomes of all these species contain centrophilic copies (copies near TRs), which form clusters. Compared to the other clusters of centrophobic copies, the copies in the centrophilic clusters tend to have shorter terminal branches as is the case for centrophilic *ALE4* copies of *A. lyrata*. Consistently, *Ty1*/*copia* elements are enriched in *Brassica oleracea* centromeres with higher LTR identities [81](Li et al 2024). In addition, phylogenic organization of centrophilic and centrophobic clusters suggest their recurrent conversions. In sum, the genome-wide analyses of other plant species suggest that centrophilic *Ty1*/*copia* copies are prevalent in other plant species, although it is almost absent in the *A. thaliana* genome.

Supplementary Discussion 2 | Centrophilic retrotransposon integration via targeting CENH3 chromatin

We concluded that *Tal1* is targeted to CENH3 chromatin. It is formally possible that *Tal1* is targeted to DNA sequences in TR or PC regions and the presence of CENH3 chromatin somehow makes these regions accessible for integration. However, proportion of TR-related sequences generally reduce sharply outside the block of the TR regions, and the sequences outside the TR are diverse [3](Naish et al 2021). Therefore, the very specific integration of *Tal1* into the regions of CENH3-covered PC in CENH3-OX lines (Fig 3d) is hard to be accounted for by presence of specific sequences in PC.

We believe chromatin features, rather than specific sequences, are the primary determinant of the centrophilic integration.

In regard to the sequence preference for the integration sites, we detected local sequence bias conserved between *Tal1* and *EVD*. (Supplementary Discussion 3 and Extended Data Fig. 9, 10), which evolved independently of centrophilic and centrophobic features. In addition, the TR sequence unit associated with centromeres are very different between *A. thalaiana* and *A. lyrata* (~30% different). Still, *Tal1*, which is localized in the *A. lyrata* TR regions, is targeted to *A. thaliana* TR regions, further suggesting that the centrophilic targeting is not based on specific sequences.

Supplementary Discussion 3 | Evolution of local sequence bias in integration

Our results demonstrate that *Tal1* and *EVD* have contrasting integration specificity at genome-wide level, which is associated with specific chromatin features. In addition, *Tal1* and *EVD* integrations showed a local sequence bias around the five base pair target site duplication (Extended Data Fig. 9). Interestingly, an almost identical local sequence bias can be seen for *Tal1* and *EVD*. Despite the similarity of *Tal1* and *EVD* in the local sequence bias, their integration specificities are contrasting at the genome-wide level. Further characterization of local integration bias in *ALE1/2/3/4* copies revealed that the bias similar to *Tal1* and *EVD* are conserved within *ALE1* and *ALE4*, while other types of bias is seen in *ALE2* and *ALE3* copies (Extended Data Fig. 10). Taken together, these observations suggest that local sequence bias can also evolve, but the evolution is independent of conversions between centrophilic and centrophobic features. These observations are consistent with our interpretation that the centrophilic and

centrophobic features are defined by rapidly evolving targeting to chromatin, rather than to specific nucleotide sequences.

Supplementary Discussion 4 | Role of C-terminal regions of integrases for specific targeting

Retrotransposon integrases show target preferences for specific chromatin features, as well as specific sequences [41,42,44]. Our results demonstrate that *Tal1* and *EVD* have contrasting centrophilic/centrophobic integration specificities at genome-wide level. These contrasting integration specificities are defined by C-terminal regions of their integrases. The C-terminal regions of integrases are generally not conserved and predicted to form disordered structure, which is often engaged in protein-protein interaction. For example, yeast *Ty1* is targeted to tRNA promoter loci, which can be accounted for by direct binding of the C-terminal region of the integrase to PolIII [83]. *Tal1* is targeted to CENH3 chromatin, suggesting that the C-terminal region of *Tal1* integrase interacts with component(s) of the kinetochore complex, which should be identified in future. Interestingly, the region with critical K/R site of *EVD*/*Tal1* localizes within predicted NLS (nuclear localization signal) of the integrase [84], and NLS region is also within the critical region of *Ty1* integrase for the targeting to tRNA loci [83]. Although targeting loci and binding partners are different, similar molecular mechanisms can be involved in the centrophilic targeting of *Tal1*.

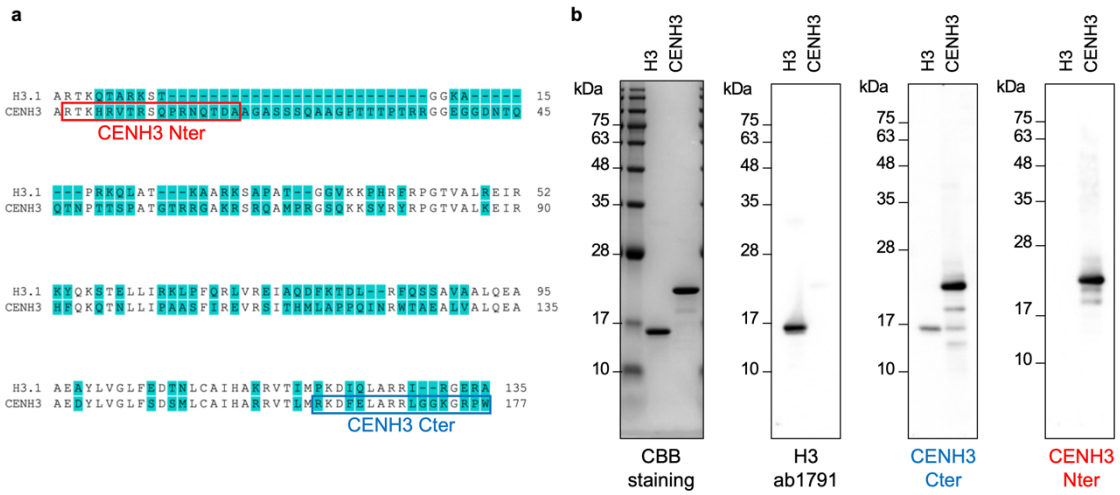
References for Supplementary Discussion

82. Li, X. et al. Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in Brassica oleracea. *Nat Genet* **56,** 517–529 (2024)

83. Asif-Laidin, A., Conesa, C., Bonnet. A., Grison, C., Adhya, I., Menouni, R., Fayol, H., Palmic, N., Acker, J., and Lesage, P. A small targeting domain in Ty1 integrase is sufficient to direct retrotransposon integration upstream of tRNA genes. *EMBO J* **39**, e104337 (2020).

84. Borredá, C., Leduque, B., Colot, V., Quadrana, L. Transposable element products, functions, and regulatory networks in Arabidopsis. *bioRxiv* 2024.04.02.587720

Supplementary Figure 1

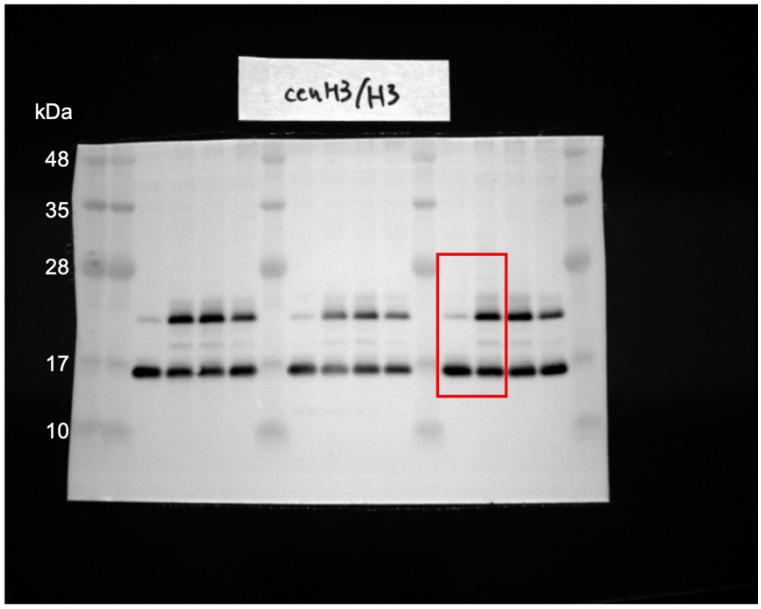


Validation of CENH3 antibody used. a, Alignment of amino acid sequence between H3 and CENH3 from *A. thaliana*. Amino acid residues different between H3 and CENH3 are shown on light blue background. Red and blue boxes indicate the peptide sequences recognized by CENH3 N-terminal and C-terminal antibodies, respectively. The former was used for ChIP-seq, while the latter was used for Western analyses. **b,** The validation of antibodies against CENH3 N-terminal and C-terminal regions. AtH3 and AtCENH3 (1 µg) were separated using 15% SDS-PAGE and visualized using Coomassie Brilliant Blue staining. Western blotting was performed using recombinant AtH3 and AtCENH3 (50 ng) proteins. Uncropped images of Western blotting results here are in Supplementary Fig. 5.

Supplementary Figure 2

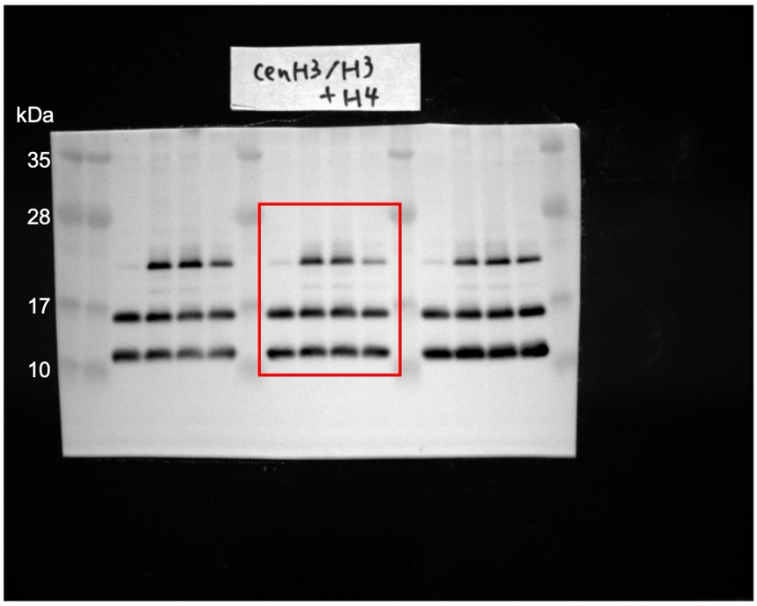
Structures of polyproteins encoded by *Tal1* and *EVD*. a, Schematic diagram of *Tal1* and *EVD* structures. Amino acid sequence similarity from the alignment between *Tal1* and *EVD* are shown. P: protease core domain, IN: integrase core domain, RT: reverse transcriptase core domain. IN1 and IN2 contain N-terminus and C-terminus region of integrase, respectively. **b,** Amino acid sequences of polyproteins encoded by *Tal1* and *EVD*. The position of R/K polymorphism within IN2 examined in Fig. 4b,c is indicated.

Supplementary Figure 3



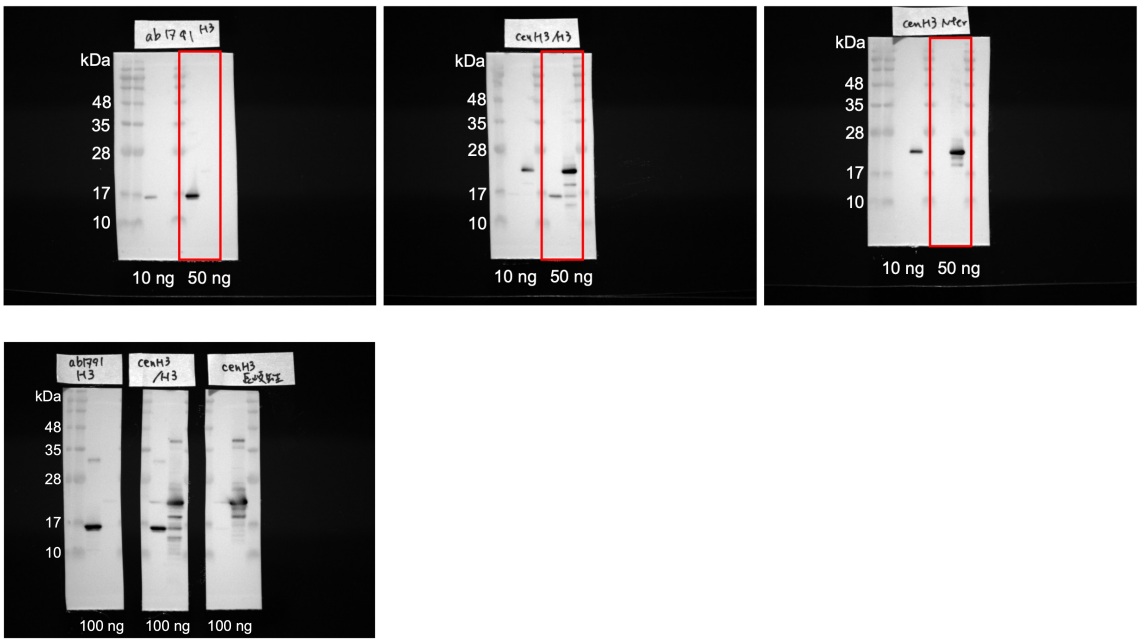
Uncropped image of results shown in Fig. 3a. The parts with red rectangles are shown in Fig. 3a. Other parts include biological replicates and molecular weight markers.

Supplementary Figure 4



Uncropped image of results shown in Extended Data Fig. 4a. The part with red rectangle is shown in Extended Data Fig. 4a. Other parts include biological replicates and molecular weight markers.

Supplementary Figure 5



Uncropped images of results shown in Supplementary Fig. 1. The parts with red rectangles are shown in the Western blotting results in Supplementary Fig. 1. Other parts include biological replicates (n=3 in total) and molecular weight markers. The amount of recombinant AtH3 and AtCENH3 (10 ng, 50 ng, 100 ng) are shown at the bottom of each panel.

Supplementary Table 1

Positions of *ALE* copies shown in Fig. 1d in regard to relationship to centromeric TRs

species	family	relationship to TR	copy number
A. lyrata	*ALE1*	in	0
A. lyrata	*ALE1*	out	104
A. lyrata	*ALE2*	in	1
A. lyrata	*ALE2*	out	113
A. lyrata	*ALE3*	in	0
A. lyrata	*ALE3*	out	142
A. lyrata	*ALE4*	in	229
A. lyrata	*ALE4*	out	50
A. thaliana	*ALE1*	in	0
A. thaliana	*ALE1*	out	21
A. thaliana	*ALE2*	in	0
A. thaliana	*ALE2*	out	10
A. thaliana	*ALE3*	in	0
A. thaliana	*ALE3*	out	14
A. thaliana	*ALE4*	in	0
A. thaliana	*ALE4*	out	2

Supplementary Table 2

Positions of *ALE* copies shown in Fig. 4d in regard to relationship to centromeric TRs

species	*ALE4* group	relationship to TR	copy number
A. lyrata	G1	in	117
A. lyrata	G1	out	2
A. lyrata	G2	in	36
A. lyrata	G2	out	2
A. lyrata	G3	in	47
A. lyrata	G3	out	6
A. lyrata	G4	in	0
A. lyrata	G4	out	12
A. lyrata	G5	in	25
A. lyrata	G5	out	0
A. lyrata	G6	in	0
A. lyrata	G6	out	21
A. lyrata	G7	in	0
A. lyrata	G7	out	4
A. lyrata	G8	in	4
A. lyrata	G8	out	2
A. thaliana	G8	in	0
A. thaliana	G8	out	2

Supplementary Table 3

Experimental conditions for TEd-seq (in a separate file)

Supplementary Table 4

Plasmids and primers used for PCR (in a separate file)