

Celine, a long interspersed nuclear element retrotransposon, colonizes in the centromeres of poplar chromosomes

Haoyang Xin,^{1,2}  Yiduo Wang,³  Wenli Zhang,³  Yu Bao,⁴  Pavel Neumann,⁵  Yihang Ning,¹ Tao Zhang,⁴  Yufeng Wu,³  Ning Jiang,^{6,7,*}  Jiming Jiang,^{2,6,7,*} and Mengli Xi^{1,*}

¹State Key Laboratory of Tree Genetics and Breeding/Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing 210037, China

²Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA

³The State Key Laboratory of Crop Genetics and Germplasm Enhancement/Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing 210095, China

⁴Jiangsu Key Laboratory of Crop Genetics and Physiology/Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding/Co-Innovation Centre for Modern Production Technology of Grain Crops/Key Laboratory of Plant Functional Genomics of the Ministry of Education, Yangzhou University, Yangzhou 225009, China

⁵Biology Centre, Czech Academy of Sciences, Institute of Plant Molecular Biology, České Budějovice 37005, Czech Republic

⁶Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA

⁷Michigan State University AgBioResearch, East Lansing, MI 48824, USA

*Author for correspondence: jiangn@msu.edu (N.J.), jiangjm@msu.edu (J.J.), ximenglinjfu@126.com (M.X.)

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (<https://academic.oup.com/plphys/pages/General-Instructions>) is Mengli Xi.

Abstract

Centromeres in most multicellular eukaryotes are composed of long arrays of repetitive DNA sequences. Interestingly, several transposable elements, including the well-known long terminal repeat centromeric retrotransposon of maize (CRM), were found to be enriched in functional centromeres marked by the centromeric histone H3 (CENH3). Here, we report a centromeric long interspersed nuclear element (LINE), Celine, in *Populus* species. Celine has colonized preferentially in the CENH3-associated chromatin of every poplar chromosome, with 84% of the Celine elements localized in the CENH3-binding domains. In contrast, only 51% of the CRM elements were bound to CENH3 domains in *Populus trichocarpa*. These results suggest different centromere targeting mechanisms employed by Celine and CRM elements. Nevertheless, the high target specificity seems to be detrimental to further amplification of the Celine elements, leading to a shorter life span and patchy distribution among plant species compared with the CRM elements. Using a phylogenetically guided approach, we were able to identify Celine-like LINE elements in tea plant (*Camellia sinensis*) and green ash tree (*Fraxinus pennsylvanica*). The centromeric localization of these Celine-like LINEs was confirmed in both species. We demonstrate that the centromere targeting property of Celine-like LINEs is of primitive origin and has been conserved among distantly related plant species.

Introduction

The centromere was first recognized as the “primary constriction” of metaphase chromosomes and represents the most distinct cytological domain of metaphase chromosomes in higher eukaryotes. Chromatin in the centromeres is defined by the presence of the centromeric histone H3 (CENH3), a centromere-specific H3 histone variant (Henikoff et al. 2001). In most multicellular eukaryotes, centromeres are composed of highly repetitive DNA sequences. Long arrays of satellite repeats and retrotransposons are 2 of the most common types of centromeric repeats (Henikoff et al. 2001; Jiang et al. 2003). The evolutionary dynamics of centromeric satellite repeats have been studied in a number of plant and animal species. Like other satellite repeats, centromeric satellite repeats often evolve rapidly and can be diverged among closely related species (Gong et al. 2012; Zhang et al. 2014; Robledo et al. 2020). However, certain types of satellite repeats appear to specifically fit in the centromeric chromatin environment. For example, the monomeric units of many classical centromeric satellite

repeats are 150–200 bp long, a characteristic length for wrapping a single nucleosome. This 1 repeat–1 CENH3 nucleosome relationship was demonstrated in humans (Hasson et al. 2013) and several plant species (Zhang et al. 2013; Yang et al. 2018; Su et al. 2019). The 155 bp centromeric satellite repeat CentO in rice (*Oryza sativa*) shows both translational and rotational phasing on CENH3 nucleosomes, a feature that may play a role in the stability of centromeric nucleosomes and chromatin (Zhang et al. 2013).

Retrotransposons fall into 2 large groups including long terminal repeat (LTR) and non-LTR elements (Kumar and Bennetzen 1999). A Ty3-gypsy type of centromeric LTR retrotransposon (CR) was first discovered in grass species (Aragon-Alcaine et al. 1996; Jiang et al. 1996; Miller et al. 1998; Presting et al. 1998). CR elements were best characterized in rice (CRR, CR of rice) (Cheng, Buell, et al. 2002; Cheng, Dong, et al. 2002; Nagaki et al. 2004; Nagaki et al. 2005) and maize (CRM, CR of maize) (Zhong et al. 2002; Jin et al. 2004; Wolfgruber et al. 2009). Both CRR and CRM elements are highly enriched with CENH3 nucleosomes (Zhong et al. 2002; Nagaki et al. 2005). Cytologically, both CRR and CRM

Received October 10, 2023. Accepted March 14, 2024. Published online April 23, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of American Society of Plant Biologists.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

appeared to be largely restricted within the primary constriction of metaphase chromosomes (Cheng, Buell, et al. 2002; Cheng, Dong, et al. 2002; Zhong et al. 2002). Based on phylogeny, CR elements belong to a specific lineage of chromoviruses (Chromoviridae), which has been commonly named as CRM, after the CR of maize (Gorinsek et al. 2004; Kordis 2005). Although CRM lineage elements were found in a wide range of distantly related species of spermatophyta, their centromeric localization was confirmed only in angiosperm species (Neumann et al. 2011; Neumann et al. 2019). A distinctive feature of the CRM elements is the presence of an integrase chromodomain, which differs in sequence from that of other chromoviruses and was hypothesized to be responsible for targeting centromeres (Kordis 2005; Neumann et al. 2019). However, the actual mechanism behind the centromeric specificity of the CRM elements in plants remains a mystery.

In addition to the CRM elements discovered in plants, several other transposable elements (TEs) were found to reside in centromeres. The K111 provirus, a human endogenous retrovirus (HERV), has at least 100 copies in the human genome and is spread across the centromeres of 15 human chromosomes. Chromatin immunoprecipitation (ChIP) experiments confirmed the enrichment of K111 sequences in CENP-A-associated chromatin (Contreras-Galindo et al. 2013). Long interspersed nuclear elements (LINEs) are a group of non-LTR retrotransposons. LINE-1 (L1) represents one of the most abundant retrotransposons in mammalian species, including humans (Beck et al. 2011). Interestingly, the L1 elements were found to be enriched in the centromeres of phyllostomid bats (de Sotero-Caio et al. 2017). Similarly, LINEs were reported to be in centromeres in banana (D'Hont et al. 2012) and sunflower (Nagaki et al. 2015). Here, we report the discovery of Celine, a LINE element that has colonized in the CENH3-associated functional centromeres of poplar chromosomes. On the basis of genome-wide CENH3-binding mapping in *Populus trichocarpa* and pachytene chromosome and DNA fiber-based high-resolution cytological mapping in *Populus simonii*, we were able to analyze the structure, organization, and evolution of a centromeric LINE element with an unprecedented scale and details. The underlying mechanism of Celine evolution is discussed.

Results

Pt45, a centromeric repeat related to a LINE element in poplar

To identify DNA sequences associated with the centromeres of poplar chromosomes, we developed an antibody against histone CENH3 of poplar (see **Materials and methods**). The specificity of this antibody to CENH3 was confirmed by immunofluorescence assay on somatic metaphase chromosomes prepared from *P. trichocarpa* (Fig. 1A). We conducted ChIP using chromatin isolated from young leaf tissue of *P. trichocarpa*. Two DNA libraries, prepared from ChIPed DNA and input DNA, respectively, were prepared and sequenced. We obtained 36.1 and 36.5 million (M) of sequence reads from the 2 libraries. We used 5 M of random reads from the input library to computationally identify repeat sequence clusters using a similarity-based sequence clustering approach (Novak et al. 2010). The proportion (%) of each repeat cluster in the poplar genome was estimated based on the number of reads associated with each cluster. We then mapped the ChIP-sequencing (ChIP-seq) reads to the repeat clusters to identify candidate centromeric repeats based on the enrichment of each

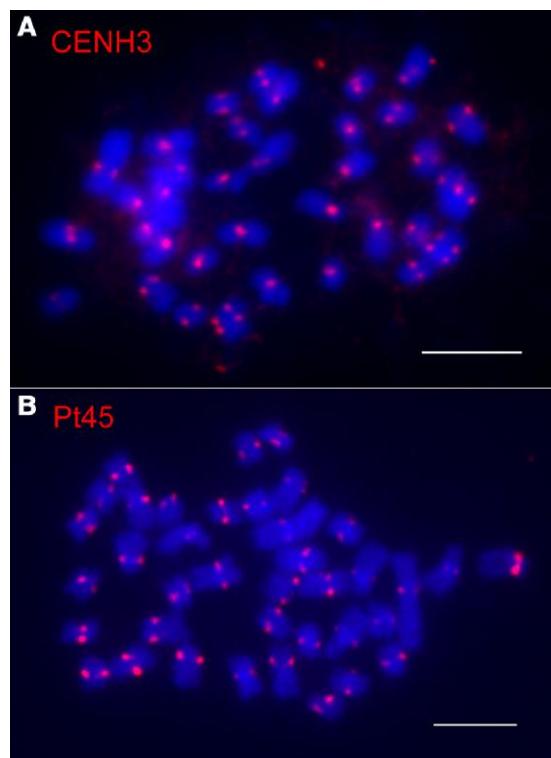


Figure 1. Identification of a centromeric repeat Pt45 in poplar. **A)** Immunofluorescence assay of the anti-CENH3 antibody on somatic metaphase chromosomes prepared from *P. trichocarpa*. **B)** FISH of the Pt45 repeat on the somatic metaphase chromosomes prepared from *P. trichocarpa*. Bars = 5 μ m.

repeat cluster in the centromeres (Gong et al. 2012; Yang et al. 2018). We analyzed the top 9 most centromere-enriched repeat clusters (Supplementary Table S1). Five of the repeats were found to be related to the CRM family. Two other repeats were related to the Athila and Tekay classes of retrotransposons (Supplementary Table S1). Interestingly, a 2,816 bp repeat cluster, Pt45, which exhibited 29 times enrichment in the centromeres and accounted for 0.15% of the poplar genome, showed 99.7% sequence similarity with a LINE-like repeat (L1-1_PTR) that was previously reported in poplar (Jurka 2010). Fluorescence in situ hybridization (FISH) analysis confirmed that Pt45 is located in the centromeres of every *P. trichocarpa* chromosome (Fig. 1B).

Celine, the most abundant LINE family in poplar

We used Pt45 as an anchor sequence and identified a full-length LINE family, named Celine (Centromeric LINE), in the *P. trichocarpa* genome. An example of a full-length Celine element is 6,114 bp in length and encodes 2 overlapping open reading frames (ORFs) of 1,617 and 3,765 bp, respectively. The ORFs are preceded by a 635 bp untranslated region (UTR) at the 5' end and an 189 bp UTR at the 3' end and terminated by a poly(A) tail of 12 bases (Fig. 2A). The function of the protein encoded by ORF1 is unknown, but its sequence possesses a domain DUF4283 (<https://www.ncbi.nlm.nih.gov/Structure/cdd/PF14111>) (Fig. 2A) that is conserved among LINEs from diverse plant species, suggesting its importance for LINE replication and/or transposition. ORF2 encodes the domains typical for all autonomous LINEs: an endonuclease and a reverse transcriptase (RT). Given that the transcripts are not spliced, the presence of the 2 ORFs suggests that Celine may use noncanonical strategies to translate both ORFs from a single

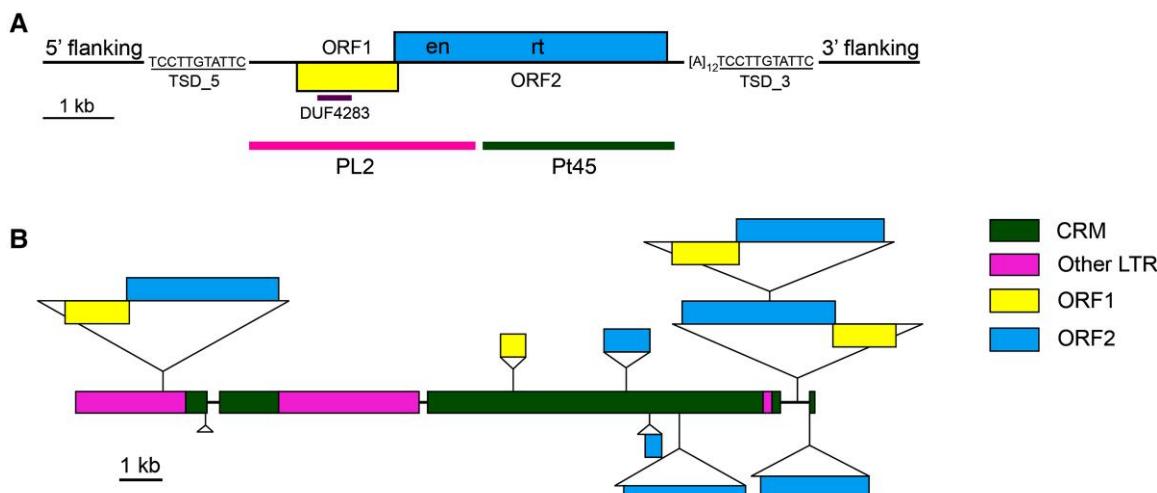


Figure 2. Structure and organization of *Celine* elements. **A**) Structure of a full-length *Celine* element from *P. trichocarpa*. This element is located on chromosome 11 (10,103,867 to 10,109,980 bp, minus strand, element ID *Celine_full034*). ORF1 contains the DUF4283 domain conserved among LINEs from diverse plant species. ORF2 contains the endonuclease (en) and RT domains. Underlined sequences represent TSD. The 12 bp poly[A] tail is depicted as [A]₁₂. The positions of the 2 FISH probes are indicated. **B**) The organization of *Celine* elements in a single Nanopore read from the *P. simonii* genome. The *Celine* elements are depicted as triangles with coding and noncoding sequences as diagrammed in **A**).

transcript (Gupta and Bansal 2020). Translation reinitiation and an internal ribosome entry site may mediate the initiation of translation, which was described for the *L1* elements in human and mouse (Alisch et al. 2006; Li et al. 2006).

In addition to *Celine*, additional LINE families were identified in the *P. trichocarpa* genome. Using protein sequences from known LINE elements, we retrieved a total of 18 LINE families of all LINEs from *P. trichocarpa*, including *Celine* and 2 families from the Repbase database (Supplementary Table S2). DNA sequences associated with these 18 LINE families account for only 0.85% of the genome. Non-LTR retrotransposons (mainly LINEs) usually occupy a small portion of the plant genomes, but they amplify to a moderate degree in some species. In a previous study, we collected 87 plant genomes with an estimated fraction of non-LTR retrotransposons (Cerbin et al. 2022). Among them, 11 (13%) harbor over 6% of non-LTR elements (range from 6% to 22%), and 17 (20%) contain <0.85%. As a result, the amplification of LINE elements in poplar is limited compared with the majority of other plants. *Celine* is the most abundant LINE element in the *P. trichocarpa* genome (Supplementary Table S2).

Celine colonized in poplar centromeres

To further characterize the location of *Celine* elements in poplar centromeres, we conducted immunofluorescence using the anti-CENH3 antibody (Fig. 3B) followed by FISH using Pt45 (Fig. 3C) on meiotic pachytene chromosomes prepared from *P. simonii*, which diverged from *P. trichocarpa* ~4.36 million years ago (Mya) (Wu et al. 2020). Pachytene chromosomes have superior cytological resolution compared with the small and highly condensed mitotic metaphase chromosomes of poplar (Xin et al. 2018). The immunofluorescence signals nearly completely overlapped with the FISH signals (Fig. 3A), suggesting that the Pt45 sequence is highly enriched in the CENH3-associated functional centromeres. We observed a similar size and intensity of the immunofluorescence signals in different centromeres (Fig. 3B), suggesting a similar size of the centromeres from different chromosomes. However, the size and intensity of the FISH signals varied significantly among different chromosomes (Fig. 3C), suggesting variable copy numbers of *Celine* in different centromeres.

To further confirm the centromeric localization of the *Celine* elements, we conducted pachytene FISH using another *Celine*-related DNA probe PL2, a 3,334 bp DNA fragment that is immediately adjacent to Pt45 within the full-length *Celine* element (Fig. 2A). The FISH signals generated from Pt45 and PL2 overlapped completely and were confined in the centromeres (Fig. 3, D to F).

Organization of the *Celine* elements in centromeres

We mapped the 36.1 M CENH3 ChIP-seq reads to the *P. trichocarpa* reference genome (see Materials and methods). The distribution of unique ChIP-seq reads was displayed in 1 kb windows along the 19 poplar chromosomes. Significant sequence enrichment was observed in the centromeres of most poplar chromosomes, except for Cen13 and Cen14 (Supplementary Fig. S1). The sizes of the 17 CENH3-binding domains averaged 633 kb, ranging from 427 to 1,267 kb among the 17 chromosomes (Supplementary Table S3). The centromeres of chromosomes 13 and 14 are likely composed of highly repetitive satellite repeats, which may not be included in the current reference genome. A similar phenomenon was previously reported in potato centromeres (Gong et al. 2012). To validate this hypothesis, we performed FISH analysis of all top 9 most abundant repeats identified in the *P. trichocarpa* centromeres (Supplementary Table S1). Interestingly, we discovered that repeat Pt7 hybridized to the 4 centromeres of chromosomes 4 and 13. Repeat Pt20 hybridized to 3 centromeres of both copies of chromosomes 14 and 1 copy of chromosome 5 (Supplementary Fig. S2).

We conducted dual-color FISH on DNA fibers prepared from *P. simonii* using Pt45 (green) and PL2 (magenta) as probes. The 2 probes generated long contiguous fiber-FISH signals (Fig. 4A). We collected many long fiber-FISH signals and selected 12 high-quality signals for measurement (Supplementary Fig. S3). These signals appeared to be intact and spanned $215.1 \pm 63 \mu\text{m}$ ($n = 12$), representing an average of $690.5 \pm 202.2 \text{ kb}$ using a $3.21 \text{ kb}/\mu\text{m}$ conversion rate (Cheng, Buell, et al. 2002). Thus, the total amount of centromeric sequences from the 19 *P. simonii* chromosomes was estimated to be $13.11 \pm 3.8 \text{ Mb}$. A significant proportion of the

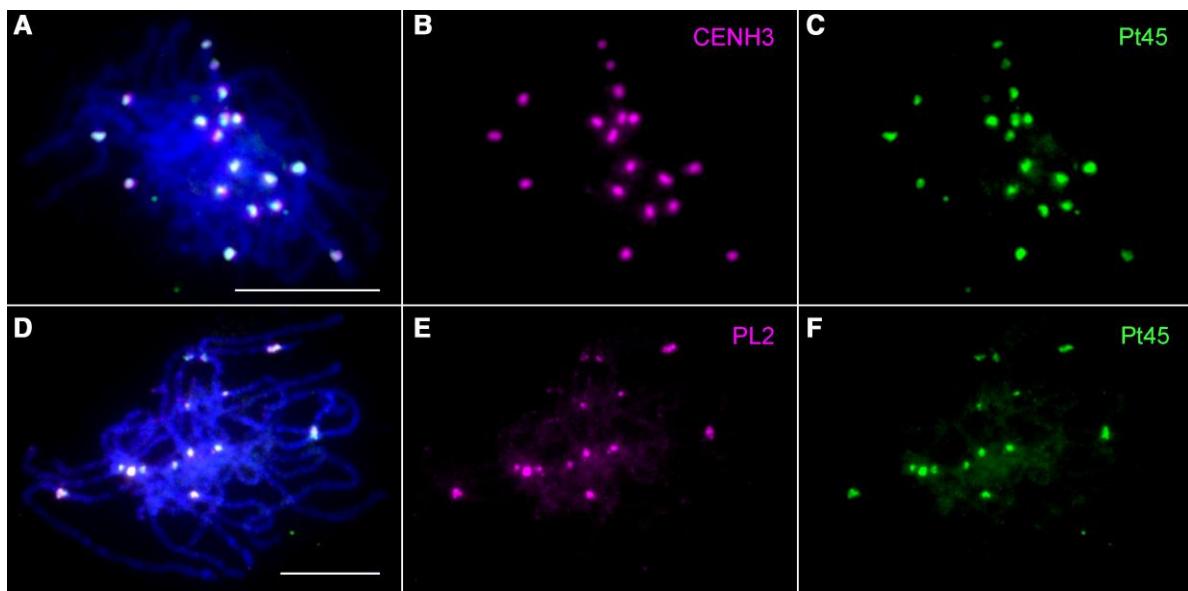


Figure 3. Mapping of CENH3 and *Celine* on pachytene chromosomes of *P. simonii*. **A)** Combined immunofluorescence assay of CENH3 and FISH using Pt45. **B)** Immunofluorescence signals that were digitally separated from **A**). **C)** FISH signals that were digitally separated from **A**). **D)** FISH mapping of PL2 and Pt45. **E)** FISH signals of PL2 that were digitally separated from **D**). **F)** FISH signals of Pt45 that were digitally separated from **D**). Bars = 10 μ m.

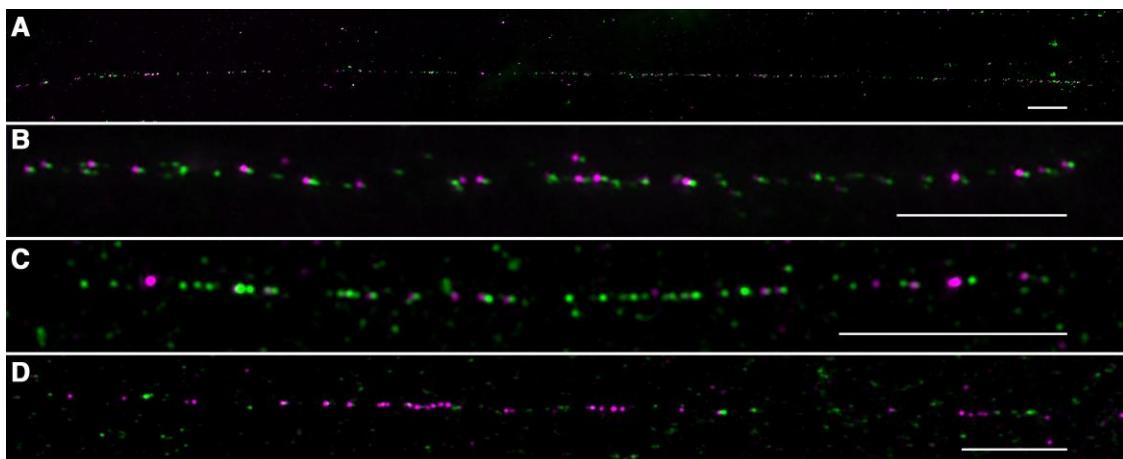


Figure 4. Dual-color fiber-FISH analysis of *Celine* using probe Pt45 (green) and PL2 (magenta) on DNA fibers prepared from *P. simonii*. **A)** A 273.6 μ m fiber-FISH signal, which represents ~878 kb of DNA and likely represents an intact centromere. **B)** A representative fiber-FISH signal with adjacent green and magenta signal dots. **C)** A representative fiber-FISH signal with contiguous green signal dots. **D)** A representative fiber-FISH signal with contiguous magenta signal dots. Bars = 10 μ m.

fiber-FISH signals were composed of adjacent green and magenta dots (Fig. 4B), indicating that these *Celine* elements contain both sequences and are likely full-length or near full-length elements. However, we observed contiguous green (Fig. 4C) or magenta signals (Fig. 4D), which were as long as 25.2 μ m (~81 kb). The *Celine* elements associated with these single-color signals are likely truncated. These clustered single-color signals were possibly derived from nested insertions or from regional duplication/amplification events.

Recent amplification and short life span of *Celine* elements

TEs can be grouped into autonomous elements, which code the proteins required for transposition, or nonautonomous elements, which rely on their cognate autonomous elements for movement

within the genome. A total of 58 full-length or nearly full-length *Celine* elements (<100 bp truncation at the 5' end and <500 bp internal deletion) were identified in the latest assembly of the *P. trichocarpa* genome (Pop_tri_v4) (Supplementary Data Set 1 and Supplementary Fig. S4). Among the full-length elements, 14 (24%) harbor both intact ORF1 and ORF2 (Supplementary Data Set 1 and Supplementary Fig. S4), suggesting that they have the potential to encode the functional transposition machinery. Among them, 12 overlap with the CENH3-binding domains, and the 2 additional elements are located within 20 kb of the CENH3-binding domains. As a result, virtually all the putative autonomous elements are buried in the functional centromeres. For the remainder (44) of the elements, the ORFs are disrupted by insertions, deletions, or point mutations (Supplementary Fig. S4). Thus, these elements are likely nonautonomous despite their sizes. The overall nucleotide level pairwise identity among these

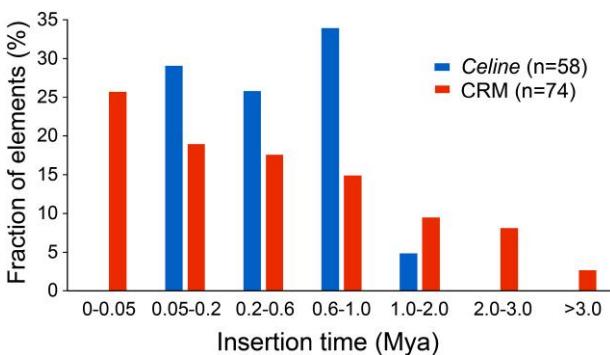


Figure 5. Estimated insertion time of full-length Celine elements and intact centromeric retrotransposon of maize (CRM) elements in *P. trichocarpa*. The x axes show the insertion time in Mya.

elements ranged from 93.8% to 99.7%. Because each individual insertion was derived from its immediate ancestral copy, an approximate distribution of elements over time can be estimated through the highest pairwise similarity of elements in the genome. Using an “all versus all match,” the highest pairwise identity for each element varied from 95.2% to 99.7%. This analysis revealed the presence of many recent elements and a few old elements with a median identity of 99.1%. Assuming a mutation rate of $\mu = 1.3 \times 10^{-8}$ per bp per year (Ma and Bennetzen 2004), 55 out of the 58 (95%) of the Celine elements were inserted into the genome within 1 million years, with the 3 other elements inserted within 1 to 2 million years, suggesting a recent amplification of the Celine family (Fig. 5). Nevertheless, we did not identify 2 identical full-length elements, suggesting a lack of current or extremely recent transposition activity.

As a comparison, we identified 74 intact CRM elements from the *P. trichocarpa* genome and used the sequence identity of the 2 LTRs to estimate the insertion time as described (SanMiguel et al. 1998). The median LTR identity is 99.2%, which is slightly higher than the pairwise identity of Celine elements (99.1%). However, the CRM elements and Celine elements have distinct amplification spectra (Fig. 5). The oldest CRM element was inserted 3.3 Mya, which is close to the detection limit. Meanwhile, 14 CRM elements have identical LTRs, suggesting current or very recent activity. Among them, only 6 are located in the CENH3-binding domains or within 20 kb flanking regions. The other 8 elements are located on chromosomal arms, with 5 of them harboring intact ORFs. These results suggest that CRM elements have been active from the trackable past. In other words, CRM elements have a much longer life span than the Celine elements.

Targeting specificity of Celine

Both CRM and Celine elements are highly enriched in the CENH3-binding domains. However, CRM elements are also present throughout individual chromosomes (Supplementary Fig. S5). A majority (84%) of the Celine elements was detected in the CENH3-binding domains; in contrast, only 51% of the CRM elements were bound to CENH3 domains (Fig. 6A). Celine elements outside of centromeres are relatively rare compared with CRM elements (Fig. 6B). Among the 58 full-length Celine elements identified in *P. trichocarpa*, 52 inserted into other TEs; most of these insertions were into Gypsy-like LTR retrotransposons, including 30 into CRM elements.

In humans, L1 elements recognize specific target sequences through its endonuclease domain, which generates staggered

nicks prior to retrotransposition (Jurka 1997). To investigate whether Celine has any sequence specificity, we retrieved target site duplication (TSD) sequences as well as 10 bp sequences flanking the TSDs. We detected TSDs (10 bp or longer) in 43 of the 58 full-length elements (Supplementary Data Set 1) and in 100 truncated Celine elements (Supplementary Data Set 2). The length of TSDs ranged from 10 to 30 bp, with an average of 15 bp. The frequent insertions of Celine into other elements raise the question whether the amplification of Celine is due to the transposition of the target element also carrying Celine. If that is the case, one would expect duplicated TSD and flanking sequences among individual Celine elements. However, each Celine element has a unique TSD sequence (Supplementary Data Sets 1 and 2). Moreover, all of the 14 CRM elements with identical LTRs do not contain Celine sequences or other nested insertions, suggesting elements with nested insertions are unlikely competent for further transposition. Those observations indicate that Celine was amplified through their own transposition machinery instead of piggybacking on other TEs including CRM elements.

The average GC content of these retrieved sequences is 35.5%, which is slightly higher than the 33.8% GC in *P. trichocarpa* genome. We examined the base occurrence in each individual position. Base bias is most significant around the junction between the 5' flanking sequence and TSD. The first nucleotide of the TSD (position 1) is G or T (90%); the nucleotide immediately upstream of the TSD (position 0) and the 2nd nucleotide of the TSD (position 2) are mostly A (66% and 58%, respectively), and the 2nd nucleotide upstream of TSD (position -1) is also biased toward A (57% of occurrences) (Fig. 6C). However, there is no significant bias at the 3' junction site (Fig. 6D). This suggests that Celine primarily targets AAGA/AATA or its variants as the 5' nicking site. In contrast, the target sequences of CRM elements contain a few AT-rich sites (Fig. 6E); hence, the 2 elements have distinct specificity at the sequence level.

For many CRM retrotransposons, a chromodomain (CHDCR) is present at the C-terminus of integrase, and this domain was assumed to direct CRM elements to centromeres (Neumann et al. 2011). Analysis of intact Celine sequences using DANTE failed to detect any CRM-related domains (Neumann et al. 2019).

Celine elements in *P. simonii*

To further investigate the genomic distribution and organization of Celine elements, we examined the presence of Celine elements in our recently developed reference genome of *P. simonii*, which was sequenced using Oxford Nanopore long-read methodology. This reference genome contains 413 Mb of sequences and is comprised of 2,814 contigs assembled using wtdbg2 (Ruan and Li 2020). Sequences similar to Celine were found in 245 contigs, accounting for 21.1 Mb of the genome. Celine elements represent 11% of the genomic sequences in the 245 contigs. However, 10 of the 245 contigs only contain <2% of Celine-related sequences, suggesting that these contigs were likely derived from the boundary regions between centromere and pericentromeric region. If these 10 contigs are excluded, the remaining 235 contigs may represent the core centromeric regions containing 15.2 Mb, which is close to the estimation of 13.1 Mb based on fiber-FISH measurements (see above). The Celine elements account for ~15% of the DNA in the 235 contigs. Similar to what is observed in *P. trichocarpa*, most Celine elements (~65%) in *P. simonii* inserted into LTR retrotransposons, and some of them (~7%) were nested within themselves. For example, a single nanopore read of 46.5 kb was found to harbor 9 Celine elements, which accounts for over half of the sequence (Fig. 2B).

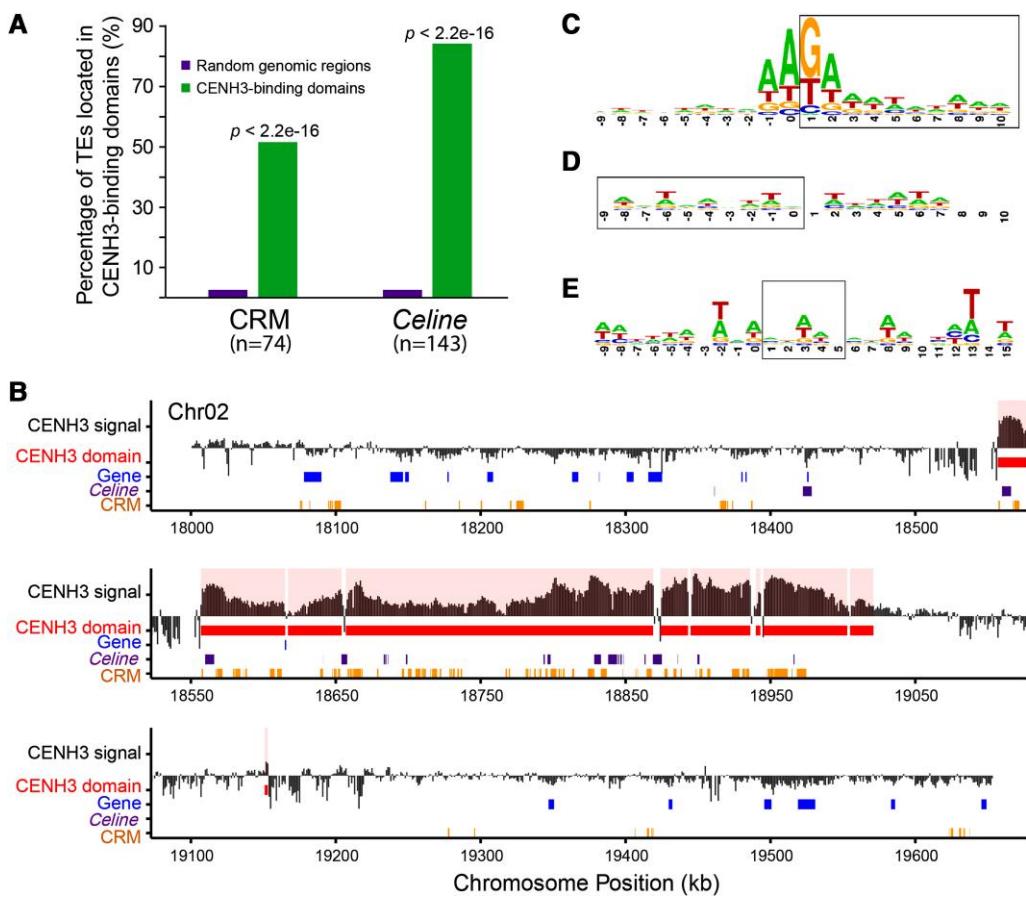


Figure 6. Distribution and targeting specificity of Celine and centromeric retrotransposon of maize (CRM) elements. **A)** Percentages of CRM and Celine elements located in the CENH3-binding domains. We computationally generated 100 times of random genomic regions that is equivalent to the length of total CENH3-binding domains. The average percentage of CRM and Celine within with the random genomic regions is shown in the y axis. **B)** Distribution of CRM and Celine elements in the centromere of pericentromeric region of chromosome 2 of *P. trichocarpa*. The x axes show the chromosome position. CENH3 signal track in the y axes indicates the relative enrichment of CENH3 ChIP-seq reads. The CENH3-binding domains are highlighted. Genes, CRM, and Celine elements are shown in different tracks. Note: Celine elements are rare outside of the CENH3-binding domains. **C)** A pictogram illustrating the sequence at the 5' end junction of Celine. **D)** A pictogram illustrating the sequence at the 3' end junction of Celine. **E)** Target sequence of CRM elements. Sequences inside the boxes represent TSD.

Celine-like LINE elements in plants

We selected a set of 4 diploid *Populus* species from different sections and a diploid willow species for FISH mapping using Pt45 and PL2. Both probes produced robust centromeric signals on all chromosomes from *Populus deltoides* (section Aigeiros), *Populus lasiocarpa* (section Leucoides), and *Populus euphratica* (section Turanga) (Fig. 7, A to C). However, the FISH signals were significantly weaker on chromosomes of *Populus tomentosa* (section Leuce), which diverged from *P. trichocarpa* 13.4 Mya (An et al. 2022) (Fig. 7D). Unambiguous FISH signals were not detected in the centromeres of chromosomes prepared from shrub willow (*Salix suchowensis*), which diverged from poplar 65 Mya (Tuskan et al. 2006).

Previous studies reported 3 LINE elements located in the centromeric regions in plants, including the Nanica element in banana (D'Hont et al. 2012; Belser et al. 2021), HaCEN-LINE in sunflower (Nagaki et al. 2015), and LINE-CL3 in the parasitic and holocentric *Cuscuta europaea* (Oliveira et al. 2020; Vondrák et al. 2021). To further test whether Celine-like elements are present in additional plant species, we searched for related elements in the National Center for Biotechnology Information (NCBI) nonredundant database and available plant genomes in phytozome (<https://phytozome.jgi.doe.gov/>).

The RT domain from the recovered sequences and those from the 3 known centromeric LINE elements were used to build a phylogenetic tree (Fig. 8). LINEs from plant species fall into 2 clades: the L1 clade and the RTE clade. Of the 7 subclades within the L1 clade (Heitkam et al. 2014), the L1-CS subclade contains LINEs associated with centromeres. We identified 2 putative centromeric LINEs based on their phylogenetic relationship with Celine (Fig. 8). The first element, L1-01_Cs, was identified in the tea plant (*Camellia sinensis*). FISH analysis using a probe developed from L1-01_Cs revealed dispersed signals on most chromosomes. However, punctuated signals were observed in the centromeric regions of several chromosomes (Fig. 7E). The second element, Celine_Fp, was identified in the green ash tree *Fraxinus pennsylvanica*. Distinct centromeric FISH signals were observed on 2 chromosomes (Fig. 7F). Examination of the green ash genome identified 15 Celine_Fp elements on 10 (out of 23) chromosomes, accounting for 0.015% of the genome. Only 4 chromosomes harbor 2 or more elements. The highest pairwise identity of those elements range from 84% to 96% (corresponding to 6.2 to 1.5 million years), suggesting limited recent activity of the Celine-like LINE in green ash with much older elements when compared with *P. trichocarpa*.

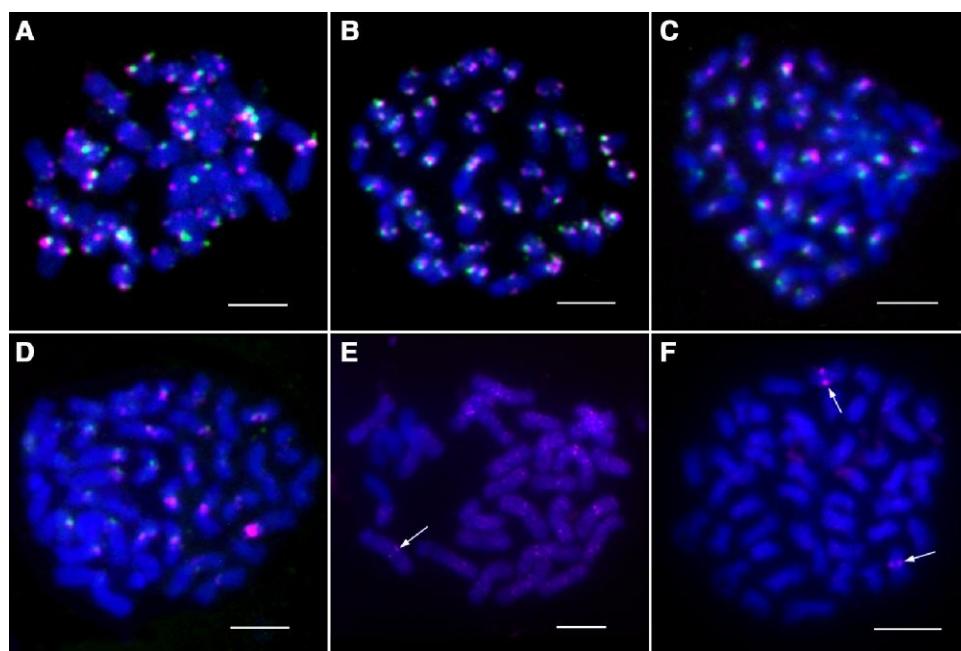


Figure 7. FISH mapping of *Celine* and *Celine*-related LINE elements in different plant species. **A** to **D**) Dual-color FISH of Pt45 (magenta) and PL2 (green) on metaphase chromosomes from *P. deltoides* (A), *P. euphratica* (B), *P. lasiocarpa* (C), and *P. tomentosa* (D). **E**) FISH mapping of a *Celine*-related LINE element in *C. sinensis*. The arrow indicates one of the punctuated centromeric signals. **F**) FISH mapping of a *Celine*-related LINE element in *F. pennsylvanica*. A pair of centromeric signals is indicated by arrows. Bars = 5 μ m.

Discussion

The centromere targeting specificity of CRM and *Celine* elements

The centromeric specificity of CRM elements has been one of the most intriguing mysteries in plant centromere biology. The integrase of the CRM elements contains a distinct chromodomain compared with other chromoviruses. This chromodomain has been speculated to play a role in the centromeric specificity of CRM elements (Kordis 2005; Gao et al. 2008). However, extensive analysis of CRM elements from a large number of plant species has yet to reveal a key domain or motif in the integrase that would be required for their centromeric specificity (Neumann et al. 2011). CRM elements intermingle with the centromeric satellite repeat CentO in rice (Cheng, Dong, et al. 2002) and CentC in maize (Jin et al. 2004). Nevertheless, CRM elements do not appear to target satellite repeats in rice and maize. CRM elements become the major centromeric DNA component in plant species lacking a dominant centromeric satellite repeat(s) such as in wheat (Liu et al. 2008; Li et al. 2013).

LINEs have different structures compared with LTR retrotransposons. Proteins encoded by the 2 ORFs of *Celine* do not contain a domain similar to the chromodomain of CRM elements. In general, LINE elements have distinct niches from Gypsy-like LTR elements (Cerbin et al. 2022), so it is intriguing to observe that *Celine*-like elements colocalize with CRM elements. The Nanica element was extensively intermingled with CRM elements in banana centromeres (Belser et al. 2021). Similarly, *Celine* elements are frequently nested with CRM elements in poplar (see *Results*). Despite the colocalization of these 2 elements, it is clear they target different sequences (Fig. 6, C to E). Moreover, *Celine* is more specifically located in centromeric regions than CRM elements in poplar (Fig. 6B, *Supplementary Fig. S5*). As a result, the frequent association of *Celine* with CRM elements is because they are both enriched in centromeric regions, not because they share targeting

mechanisms. At this stage, it is unclear how *Celine* targets centromeres. Our analysis indicates it preferentially targets AAGA or AATA sequence motifs. Nevertheless, these combinations of nucleotides are common in the genome. Thus, it is unlikely that these motifs are sufficient to determine its chromosomal locations. It is possible that both CRM and *Celine* elements target a component or different components associated with CENH3 nucleosomes but with different affinities.

Evolution of *Celine*-like retrotransposons

TEs are major components of eukaryotic genomes. The success of a TE relies on the genetic and epigenetic environments of the genome and the presence of other TE families. TEs are more dynamic and variable than genes due to their ability to amplify and that most of them are dispensable to the host organisms. Most TE families experience a full life cycle of birth, amplification, and extinction (Blumenstiel 2019; Liu et al. 2022). From an evolutionary point of view, individual families of transposons are only transiently present in the genome. LINEs represent the most abundant TE in mammalian genomes (Lander et al. 2001). Whereas in most plant genomes, LINEs only account for a few percent or less (Cerbin et al. 2022). The underlying mechanism for the low abundance of LINEs in plants is not well understood. LINEs were recently found to preferentially insert in introns in sacred lotus (*Nelumbo nucifera*) (Cerbin et al. 2022). Concordantly, the average intron size is 1,988 bp in lotus, which is significantly larger than the average intron sizes in other model plant species, such as *Arabidopsis thaliana* (170 bp) (Kaul et al. 2000) (TAIR10), rice (447 bp) (Kawahara et al. 2013) (IRGSP-1.0), and poplar (<400 bp) (Tuskan et al. 2006). Intriguingly, large introns are a well-known characteristic associated with mammalian genomes. *Celine* is the most abundant LINE element in poplar (*Supplementary Table S2*). Thus, centromeres may serve as a “safe haven” for *Celine* to survive and thrive, similar to the

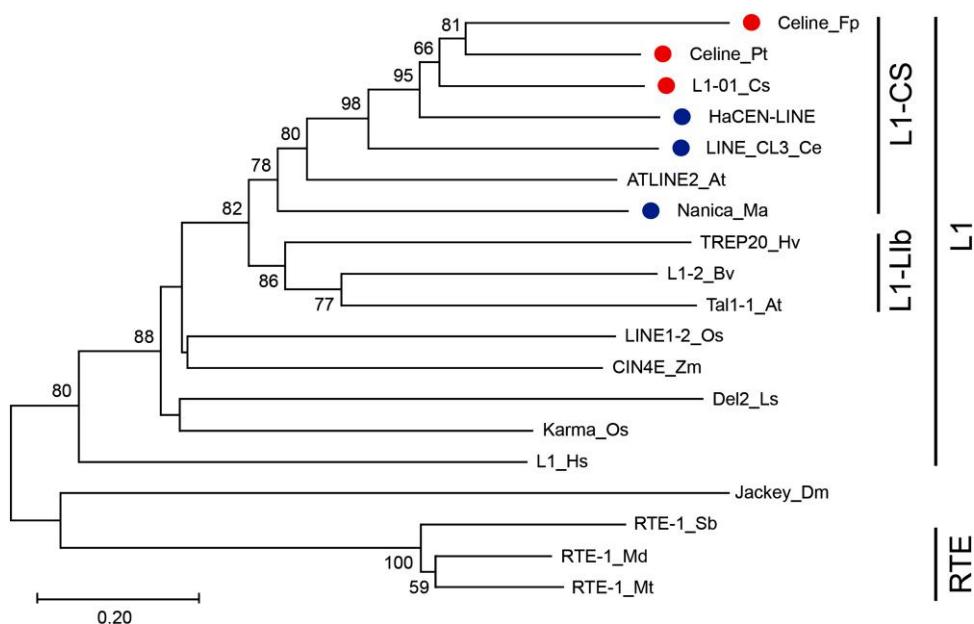


Figure 8. The phylogeny of core RT domain of selected LINE elements. The evolutionary history was inferred using the minimum evolution method with branch length corresponding to the evolutionary distance. Numbers next to the branch indicate the % bootstrap support (1,000 replicates, 50% cutoff). Elements with a red dot are centromere-located elements identified in this study; elements with a blue dot are centromere-located elements identified previously. Abbreviations for species: At, *A. thaliana*; Bv, *Beta vulgaris*; Ce, *C. europaea*; Cs, *C. sinensis*; Dm, *Drosophila melanogaster*; Fp, *F. pennsylvanica*; Ha, *Helianthus annuus*; Hv, *Hordeum vulgare*; Hs, *Homo sapiens*; Ls, *Lilium speciosum*; Ma, *Musa acuminata*; Md, *Malus x domestica*; Mt, *Medicago truncatula*; Os, *O. sativa*; Pt, *P. trichocarpa*; Sb, *Sorghum bicolor*; Zm, *Zea mays*. Abbreviations for elements: L1, human non-LTR retrotransposon LINE-1; L1-Lib, sweet potato LINE Lib; L1-CS, a LINE-like retrotransposon in a sex chromosome of the dioecious plant *Cannabis sativa*; RTE, the RTE-1 (retrotransposable element) element, which was first identified in *Caenorhabditis elegans*.

large introns housing for LINEs in lotus and mammalian genome.

Besides the Nanica element from banana (a monocot plant), all the other 5 Celine-like elements were found in eudicots. Among them, 4 plants (*C. europaea*, sunflower, tea, and *F. pennsylvanica*) are asterids (but in different orders), whereas *Populus* belongs to rosids (Moore et al. 2010). Interestingly, Celine is phylogenetically related to elements from Asterids (Fig. 8). Since asterids diverged from rosids about 125 Mya (Zeng et al. 2017), it suggests that there was either an ancient horizontal transfer event or Celine diverged into multiple groups before the divergence of dicots. The presence of Celine-like elements in distantly related species supports that Celine has an ancient origin. If so, it raises the question of why a Celine-like element is absent in most of the sequenced plant genomes while CRM elements are widely present in plants. This is likely attributed to the unique transposition mechanism of LINEs and the high specificity of Celine. Upon insertion into the genome, most LTR elements are intact elements, and it is common for an autonomous element to give birth to another autonomous element. In contrast, most LINE elements are truncated at the 5' end upon insertion, representing nonautonomous elements (Hancks and Kazazian 2016). In this scenario, the consequence of the high specificity of Celine leads to the high density of elements in the centromeric regions, with elements nested with each other. This elevates the frequency of truncated elements due to the interruption of the autonomous elements. Even if the element remains intact, the formation of heterochromatin around the centromere may prevent active transcription, resulting in loss of transposition activity and eventual extinction. In contrast, the targeting of CRM elements is not as specific as Celine, and multiple putative autonomous elements are found in chromosomal arms, allowing the continuous activity of this family of elements (Fig. 5) and likely contributing to the prevalence of

CRM in many plant genomes. Again, this demonstrates the importance of targeting specificity and transposition mode to the success of TEs. From this point of view, the activity of Celine is transient on an evolutionary scale, and this explains why it is only detected in a few plant species among thousands of sequenced plant genomes. This model would predict that in a few million years, no FISH signal of Celine will be detected in green ash and signals will be detected on approximately half of the poplar chromosomes. As a result, the high centromere specificity of Celine represents an evolutionary “dead end.” Meanwhile, new Celine elements may evolve from elements with lower centromere specificity or be introduced through horizontal transfer.

Materials and methods

Plant materials

Six poplar species ($2n=2x=38$) were used in the present study, including *P. trichocarpa*, *P. simonii*, *P. deltoides*, *P. euphratica*, *P. lasiocarpa*, and *P. tomentosa*. Three additional nonpoplar species were also used for the presence of Celine-like elements, including willow (*S. suchowensis*, $2n=2x=38$), tea plant (*C. sinensis*, $2n=2x=30$), and green ash tree (*F. pennsylvanica*, $2n=2x=46$).

Immunofluorescence assay, FISH and fiber-FISH

A CENH3 antibody was developed as a rabbit polyclonal antiserum and raised against the synthesized peptide of the 20 most N-terminal amino acid sequence (MARTKHPVARKRARSPKRSD) of the CENH3 protein of *P. trichocarpa*. Immunofluorescence was performed according to previously published protocols using the poplar anti-CENH3 antibody (Zhang et al. 2005). For the immunofluorescence combined with FISH assay, after recording of the immunostaining signals, the cytological preparations were washed

and followed with a sequential FISH procedure as previously described (Xin et al. 2020).

Preparation of mitotic and meiotic chromosomes was performed according to the protocols described in our previous studies (Xin et al. 2018; Xin et al. 2020). DNA probes specific to the Pt45 and PL2 sequences were amplified via PCR using *P. trichocarpa* DNA as a template. DNA probes of *Celine*-like elements were amplified from the genomic DNA of *C. sinensis* and *F. pennsylvanica* using specific primers (Supplementary Table S4). These amplified DNA fragments were excised from agarose gel, purified, and labeled by nick translation with either digoxigenin-dUTP or biotin-dUTP. FISH and fiber-FISH were performed according to published protocols (Jackson et al. 1998; Xin et al. 2020). Cytological measurements of the fiber-FISH signals were converted into kilobases using a 3.21 kb/μm conversion rate (Cheng, Buell, et al. 2002).

ChIP-seq and mapping of CENH3-binding domains

ChIP was performed as previously described (Nagaki et al. 2003). Young leaf tissue of *P. trichocarpa* was used to extract chromatin for ChIP assays. Approximately 30 ng of ChIP and input DNA were used for library preparation and sequenced by Illumina HiSeq 2000 platform with 125 bp paired reads. The sequence reads from ChIP and input were mapped to genome v4.0 of *P. trichocarpa* (<http://www.phytozome.net/poplar>) by Bowtie2 (Langmead et al. 2009). We allowed a 1 bp mismatch threshold between each sequence read and the reference genome. Only the reads mapped to a unique site in the poplar genome were used for further analysis. We divided each poplar chromosome into 1 kb windows and calculated the unique read number per base pair mappable region. Read density was presented as the number of unique reads in a 1 kb window per the length of mappable region in the same window. The final read density was adjusted using the input sequence read data to reduce background signals.

We used SICER2 (Zang et al. 2009) to identify CENH3-binding domains in each poplar centromere within 1 kb windows. We set a mapping stringency that the false discovery rate (FDR) value of a CENH3-subdomain was <0.001 and the fold change of normalized reads number ChIP/input was >5. Identification of centromeric repeats was performed based on the similarity-based clustering method (Novak et al. 2013). Briefly, 5 million reads from the input were used to perform graph-based clustering using the RepeatExplorer web server (<https://repeatexplorer.elixir.cerit-sc.cz/>). Repeats were identified and classified based on their sequence similarity as individual repeat clusters. To identify repeats enriched in centromeres, ChIP and input reads were mapped to the repeat clusters using BLAT (Kent 2002). The CENH3 enrichment for each repeat was determined as described previously (Gong et al. 2012).

Identification of *Celine* and CRM elements in the poplar genome and estimation of abundance

To search for full-length *Celine* elements, the initial Pt45 sequence (see *Results*), which is 2,816 bp in length, was used to search the latest poplar genome assembly ((Tuskan et al. 2006), Pop_tri_v4) using BLASTN ($E < 10^{-10}$) (Altschul et al. 1990). Pop_tri_v4 was derived from a 133.2x of PACBIO coverage sequences as well as a high-density poplar map (https://phytozome-next.jgi.doe.gov/info/Ptrichocarpa_v4.1). Sequences matching Pt45 as well as 4 kb flanking sequences on each side were retrieved and aligned with MUSCLE using default parameters (Edgar 2004). The alignment was manually examined for the presence of TSDs flanking the

boundary of the alignment. Only TSDs that are 10 bp or longer are considered high confidence and included in Supplementary Data Sets 1 and 2. One of the longest elements, located on chromosome 11 (10,103,867 to 10,109,980 bp; Fig. 2A), harbors 2 apparently intact ORFs when compared with known LINEs in Repbase (Bao et al. 2015) and was considered a representative *Celine* element. Using this element as a standard, *Celine* with a similar length or truncated <100 bp at the 5' end were considered full-length elements, while other elements were considered truncated elements. For CRM elements, poplar LTR elements in Repbase were extracted. Additional LTR element sequences in poplar were collected using LTR_retriever (Ou and Jiang 2018). All LTR elements were classified using DANTE (Neumann et al. 2019), and those containing CRM domains are considered CRM elements. The sequences of CRM elements and *Celine* were included in a repeat library to mask the poplar genome using RepeatMasker (<https://www.repeatmasker.org/>), with the abundance of each element assessed based on the length of each element masked. For *Celine* elements, both full-length and 5' end truncated elements are considered a copy. Fragments (without either end) are not included in copy number estimations.

Target specificity of *Celine* and CRM elements

For chromosomal level distribution of *Celine* and CRM elements, each chromosome was divided into bins that are 100 kb in length. The fraction of *Celine* or CRM elements was estimated based on the length they covered in each bin. Thereafter, the relative abundance value of the bin with the highest fraction was set to 100 and used to normalize the values in other bins. For the sequence preference at the junction sites of *Celine*, sequences at the 5' junction site (10 bp upstream of TSD plus 10 bp into TSD) were retrieved and the sequence logo was generated using WebLogo (Crooks et al. 2004). Sequences at the 3' junction site (the last 10 bp of the TSD plus 10 bp downstream of TSD) were processed similarly. The sequence logo was generated for CRM elements with the 5 bp TSD plus 10 bp flanking sequences on each side. To evaluate the number of *Celine* elements inserted into CRM elements, 150 bp sequence downstream of each *Celine* element was masked using a CRM element library. If the flanking sequence was masked, the relevant *Celine* was considered to be inserted into a CRM element.

Phylogenetic analysis

The RT domain of *Celine* elements and other LINEs were aligned with MUSCLE using default parameters (Edgar 2004). The origin of sequences used in phylogenetic analysis is listed in Supplementary Tables S5 and S6. The RT domain was defined based on comparison with ORF2 of L1 from the human genome (511 to 773 amino acids, GenBank: AAA51622.1). Phylogenetic trees were generated using a neighbor-joining method with MEGA (Kumar et al. 2018). Support for the internal branches of the phylogeny was assessed using 1,000 bootstrap replicates.

Acknowledgments

We thank Dr. Pingdong Zhang (Beijing Forestry University) and Dr. Xiaobiao Zhu (Anhui Agriculture University) for collecting and providing the plant materials for this research.

Author contributions

J.J. and M.X. conceived the development of the manuscript. H.X., W.Z., and Y.N. conducted the chromatin and cytogenetic

experiments. B.Y. and T.Z. developed the *P. simonii* reference genome. H.X., Yi.W., Yu.W., P.N., N.J., J.J., and M.X. analyzed the data. H.X., N.J., J.J., and M.X. wrote the manuscript.

Supplementary data

The following materials are available in the online version of this article.

Supplementary Figure S1. Distribution and density of CENH3 ChIP-seq reads on each chromosome of *P. trichocarpa*.

Supplementary Figure S2. FISH mapping of centromeric repeats Pt7 and Pt20 in *P. trichocarpa*.

Supplementary Figure S3. A set of 12 high-quality long contiguous fiber-FISH signals derived from Pt45 (green) and PL2 (magenta).

Supplementary Figure S4. Structure of additional 57 full-length or nearly full-length *Celine* elements in *P. trichocarpa*.

Supplementary Figure S5. Distribution and density of CENH3, *Celine*, and CRM on each chromosome of *P. trichocarpa*.

Supplementary Table S1. Nine most abundant repeat clusters identified in the centromeres of *P. trichocarpa* chromosomes.

Supplementary Table S2. Amplification of LINE elements in *P. trichocarpa*.

Supplementary Table S3. Size of CENH3-binding chromosomal domains in *P. trichocarpa*.

Supplementary Table S4. PCR primers used in the study.

Supplementary Table S5. Elements used in phylogenetic analysis (from NCBI).

Supplementary Table S6. Elements used in phylogenetic analysis (from other resources).

Supplementary Data Set 1. List of full-length *Celine* elements in *P. trichocarpa*.

Supplementary Data Set 2. List of truncated *Celine* elements with TSD in *P. trichocarpa*.

Funding

This research is supported by the United States National Science Foundation grant IOS-1740874, the United States Department of Agriculture National Institute of Food and Agriculture and AgBioResearch at Michigan State University (Hatch grant M1CL2707) to N.J., the United States National Science Foundation grant ISO-2029959 to J.J., and grant 31670603 from the National Natural Science Foundation of China to M.X.

Conflict of interest statement. None declared.

Data availability

ChIP-seq data sets have been submitted to NCBI under the BioProject accession number PRJNA1021681. The genome assembly of *P. simonii* has been deposited to Genome Warehouse in the National Genomics Data Center (NGDC) under accession number GWHDUDC00000000 (<https://ngdc.cncb.ac.cn/gwh/>).

References

Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV. Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev.* 2006;20(2):210–224. <https://doi.org/10.1101/gad.1380406>

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)

An XM, Gao K, Chen Z, Li J, Yang X, Yang XY, Zhou J, Guo T, Zhao TY, Huang S, et al. High quality haplotype-resolved genome assemblies of *Populus tomentosa* Carr., a stabilized interspecific hybrid species widespread in Asia. *Mol Ecol Resour.* 2022;22(2):786–802. <https://doi.org/10.1111/1755-0998.13507>

Aragon-Alcaide L, Miller T, Schwarzacher T, Reader S, Moore G. A cereal centromeric sequence. *Chromosoma* 1996;105(5):261–268. <https://doi.org/10.1007/BF02524643>

Bao WD, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* 2015;6(1):11. <https://doi.org/10.1186/s13100-015-0041-9>

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet.* 2011;12(1):187–215. <https://doi.org/10.1146/annurev-genom-082509-141802>

Belser C, Baurens FC, Noel B, Martin G, Cruaud C, Istace B, Yahiaoui N, Labadie K, Hribova E, Dolezel J, et al. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Commun Biol.* 2021;4(1):1047. <https://doi.org/10.1038/s42003-021-02559-3>

Blumenstiel JP. Birth, school, work, death, and resurrection: the life stages and dynamics of transposable element proliferation. *Genes (Basel).* 2019;10(5):336. <https://doi.org/10.3390/genes10050336>

Cerbin S, Ou SJ, Li Y, Sun YN, Jiang N. Distinct composition and amplification dynamics of transposable elements in sacred lotus (*Nelumbo nucifera* Gaertn.). *Plant J.* 2022;112(1):172–192. <https://doi.org/10.1111/tpj.15938>

Cheng ZK, Buell CR, Wing RA, Jiang JM. Resolution of fluorescence in-situ hybridization mapping on rice mitotic prometaphase chromosomes, meiotic pachytene chromosomes and extended DNA fibers. *Chromosome Res.* 2002;10(5):379–387. <https://doi.org/10.1023/A:1016849618707>

Cheng ZK, Dong FG, Langdon T, Ouyang S, Buell CB, Gu MH, Blattner FR, Jiang JM. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell.* 2002;14(8):1691–1704. <https://doi.org/10.1105/tpc.003079>

Contreras-Galindo R, Kaplan MH, He S, Contreras-Galindo AC, Gonzalez-Hernandez MJ, Kappes F, Dube D, Chan SM, Robinson D, Meng F, et al. HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. *Genome Res.* 2013;23(9):1505–1513. <https://doi.org/10.1101/gr.144303.112>

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004;14(6):1188–1190. <https://doi.org/10.1101/gr.849004>

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 2012;488(7410):213–217. <https://doi.org/10.1038/nature11241>

de Sotero-Caio CG, Cabral-de-Mello D, Calixto MD, Valente G, Martins C, Loreto V, de Souza MJ, Santos N. Centromeric enrichment of LINE-1 retrotransposons and its significance for the chromosome evolution of Phyllostomid bats. *Chromosome Res.* 2017;25(3-4):313–325. <https://doi.org/10.1007/s10577-017-9565-9>

Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 2004;5(1):113. <https://doi.org/10.1186/1471-2105-5-113>

Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 2008;18(3):359–369. <https://doi.org/10.1101/gr.7146408>

Gong ZY, Wu YF, Koblizkova A, Torres GA, Wang K, Iovene M, Neumann P, Zhang WL, Novak P, Buell CR, et al. Repeatless and repeat-based centromeres in potato: implications for centromere

evolution. *Plant Cell*. 2012;24(9):3559–3574. <https://doi.org/10.1105/tpc.112.100511>

Gorinsek B, Gubensek F, Kordis D. Evolutionary genomics of chromoviruses in eukaryotes. *Mol Biol Evol*. 2004;21(5):781–798. <https://doi.org/10.1093/molbev/msh057>

Gupta A, Bansal M. RNA-mediated translation regulation in viral genomes: computational advances in the recognition of sequences and structures. *Brief Bioinform*. 2020;21(4):1151–1163. <https://doi.org/10.1093/bib/bbz054>

Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. *Mob DNA*. 2016;7(1):9. <https://doi.org/10.1186/s13100-016-0065-9>

Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, Alonso A, Warburton PE, Black BE. The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nat Struct Mol Biol*. 2013;20(6):687–695. <https://doi.org/10.1038/nsmb.2562>

Heitkam T, Holtgrawe D, Dohm JC, Minoche AE, Himmelbauer H, Weisshaar B, Schmidt T. Profiling of extensively diversified plant LINEs reveals distinct plant-specific subclades. *Plant J*. 2014;79(3):385–397. <https://doi.org/10.1111/tpj.12565>

Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*. 2001;293(5532):1098–1102. <https://doi.org/10.1126/science.1062939>

Jackson SA, Wang ML, Goodman HM, Jiang JM. Application of fiber-FISH in physical mapping of *Arabidopsis thaliana*. *Genome*. 1998;41(4):566–572. <https://doi.org/10.1139/g98-093>

Jiang JM, Birchler JA, Parrott WA, Dawe RK. A molecular view of plant centromeres. *Trends Plant Sci*. 2003;8(12):570–575. <https://doi.org/10.1016/j.tplants.2003.10.011>

Jiang JM, Nasuda S, Dong FG, 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(24):14210–14213. <https://doi.org/10.1073/pnas.93.24.14210>

Jin WW, Melo JR, Nagaki K, Talbert PB, Henikoff S, Dawe RK, Jiang JM. Maize centromeres: organization and functional adaptation in the genetic background of oat. *Plant Cell*. 2004;16(3):571–581. <https://doi.org/10.1105/tpc.018937>

Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A*. 1997;94(5):1872–1877. <https://doi.org/10.1073/pnas.94.5.1872>

Jurka J. L1 elements from black cottonwood. *Repbase Rep*. 2010;10:157–157.

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY, et al. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*. 2000;408(6814):796–815. <https://doi.org/10.1038/35048692>

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu JZ, Zhou SG, et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*. 2013;6(1):4. <https://doi.org/10.1186/1939-8433-6-4>

Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res*. 2002;12(4):656–664. <https://doi.org/10.1101/gr.229202>

Kordis D. A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. *Gene*. 2005;347(2):161–173. <https://doi.org/10.1016/j.gene.2004.12.017>

Kumar A, Bennetzen JL. Plant retrotransposons. *Ann Rev of Genet*. 1999;33(1):479–532. <https://doi.org/10.1146/annurev.genet.33.1.479>

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35(6):1547–1549. <https://doi.org/10.1093/molbev/msy096>

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921. <https://doi.org/10.1038/35057062>

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25. <https://doi.org/10.1186/gb-2009-10-3-r25>

Li BC, Choulet F, Heng YF, Hao WW, Paux E, Liu Z, Yue W, Jin WW, Feuillet C, Zhang XY. Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. *Plant J*. 2013;73(6):952–965. <https://doi.org/10.1111/tpj.12086>

Li PWL, Li JF, Timmerman SL, Krushel LA, Martin SL. The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. *Nucleic Acids Res*. 2006;34(3):853–864. <https://doi.org/10.1093/nar/gkj490>

Liu P, Cuerda-Gil D, Shahid S, Slotkin RK. The epigenetic control of the transposable element life cycle in plant genomes and beyond. *Ann Rev Genet*. 2022;56(1):63–87. <https://doi.org/10.1146/annurev-genet-072920-015534>

Liu Z, Yue W, Li DY, Wang RRC, Kong XY, Lu K, Wang GX, Dong YS, Jin WW, Zhang XY. Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. *Chromosoma*. 2008;117(5):445–456. <https://doi.org/10.1007/s00412-008-0161-9>

Ma JX, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A*. 2004;101(34):12404–12410. <https://doi.org/10.1073/pnas.0403715101>

Miller JT, Dong FG, Jackson SA, Song J, Jiang JM. Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. *Genetics*. 1998;150(4):1615–1623. <https://doi.org/10.1093/genetics/150.4.1615>

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci U S A*. 2010;107(10):4623–4628. <https://doi.org/10.1073/pnas.0907801107>

Nagaki K, Cheng ZK, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang JM. Sequencing of a rice centromere uncovers active genes. *Nat Genet*. 2004;36(2):138–145. <https://doi.org/10.1038/ng1289>

Nagaki K, Neumann P, Zhang D, Ouyang S, Buell CR, Cheng ZK, Jiang JM. Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Mol Biol Evol*. 2005;22(4):845–855. <https://doi.org/10.1093/molbev/msi069>

Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang JM. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics*. 2003;163(3):1221–1225. <https://doi.org/10.1093/genetics/163.3.1221>

Nagaki K, Tanaka K, Yamaji N, Kobayashi H, Murata M. Sunflower centromeres consist of a centromere-specific LINE and a chromosome-specific tandem repeat. *Front Plant Sci*. 2015;6:912. <https://doi.org/10.3389/fpls.2015.00912>

Neumann P, Navratilova A, Koblizkova A, Kejnovsky E, Hribova E, Hobza R, Widmer A, Dolezel J, Macas J. Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob DNA*. 2011;2(1):4. <https://doi.org/10.1186/1759-8753-2-4>

Neumann P, Novak P, Hostakova N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob DNA*. 2019;10(1):1. <https://doi.org/10.1186/s13100-018-0144-1>

Novak P, Neumann P, Macas J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*. 2010;11(1):378. <https://doi.org/10.1186/1471-2105-11-378>

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*. 2013;29(6):792–793. <https://doi.org/10.1093/bioinformatics/btt054>

Oliveira L, Neumann P, Jang TS, Klemme S, Schubert V, Koblizkova A, Houben A, Macas J. Mitotic spindle attachment to the holocentric chromosomes of *Cuscuta europaea* does not correlate with the distribution of CENH3 chromatin. *Front Plant Sci.* 2020;10:1799. <https://doi.org/10.3389/fpls.2019.01799>

Ou SJ, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of LTR retrotransposons. *Plant Physiol.* 2018;176(2):1410–1422. <https://doi.org/10.1104/pp.17.01310>

Presting GG, Malysheva L, Fuchs J, Schubert I. A Ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant J.* 1998;16(6):721–728. <https://doi.org/10.1046/j.1365-313x.1998.00341.x>

Robledillo LA, Neumann P, Koblizkova A, Novak P, Vrbova I, Macas J. Extraordinary sequence diversity and promiscuity of centromeric satellites in the legume tribe Fabeae. *Mol Biol Evol.* 2020;37(8): 2341–2356. <https://doi.org/10.1093/molbev/msaa090>

Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods.* 2020;17(2):155–158. <https://doi.org/10.1038/s41592-019-0669-3>

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. *Nat Genet.* 1998;20(1):43–45. <https://doi.org/10.1038/1695>

Su HD, Liu YL, Liu C, Shi QH, Huang YH, Han FP. Centromere satellite repeats have undergone rapid changes in polyploid wheat subgenomes. *Plant Cell.* 2019;31(9):2035–2051. <https://doi.org/10.1105/tpc.19.00133>

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombaut S, Salamov A, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006;313(5793):1596–1604. <https://doi.org/10.1126/science.1128691>

Vondrak T, Oliveira L, Novak P, Koblizkova A, Neumann P, Macas J. Complex sequence organization of heterochromatin in the holocentric plant *Cuscuta europaea* elucidated by the computational analysis of nanopore reads. *Comput Struct Biotechnol J.* 2021;19: 2179–2189. <https://doi.org/10.1016/j.csbj.2021.04.011>

Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, Shi JH, Gao Z, Han FP, Lee H, Xu RH, et al. Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genet.* 2009;5(11): e1000743. <https://doi.org/10.1371/journal.pgen.1000743>

Wu HN, Yao D, Chen YH, Yang WG, Zhao W, Gao H, Tong CF. De novo genome assembly of *Populus simonii* further supports that *Populus simonii* and *Populus trichocarpa* belong to different sections. *G3 (Bethesda)*. 2020;10(2):455–466. <https://doi.org/10.1534/g3.119.400913>

Xin HY, Zhang T, Han YH, Wu YF, Shi JS, Xi ML, Jiang JM. Chromosome painting and comparative physical mapping of the sex chromosomes in *Populus tomentosa* and *Populus deltoides*. *Chromosoma* 2018;127(3):313–321. <https://doi.org/10.1007/s00412-018-0664-y>

Xin HY, Zhang T, Wu YF, Zhang WL, Zhang PD, Xi ML, Jiang JM. An extraordinarily stable karyotype of the woody *Populus* species revealed by chromosome painting. *Plant J.* 2020;101(2):253–264. <https://doi.org/10.1111/tpj.14536>

Yang XM, Zhao HN, Zhang T, Zeng ZX, Zhang PD, Zhu B, Han YH, Braz GT, Casler MD, Schmutz J, et al. Amplification and adaptation of centromeric repeats in polyploid switchgrass species. *New Phytol.* 2018;218(4):1645–1657. <https://doi.org/10.1111/nph.15098>

Zang CZ, Schones DE, Zeng C, Cui KR, Zhao KJ, Peng WQ. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*. 2009;25(15):1952–1958. <https://doi.org/10.1093/bioinformatics/btp340>

Zeng LP, Zhang N, Zhang QA, Endress PK, Huang J, Ma H. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytol.* 2017;214(3):1338–1354. <https://doi.org/10.1111/nph.14503>

Zhang HQ, Koblizkova A, Wang K, Gong ZY, Oliveira L, Torres GA, Wu YF, Zhang WL, Novak P, Buell CR, et al. Boom-bust turnovers of megabase-sized centromeric DNA in *Solanum* Species: rapid evolution of DNA sequences associated with centromeres. *Plant Cell.* 2014;26(4):1436–1447. <https://doi.org/10.1105/tpc.114.123877>

Zhang T, Talbert PB, Zhang WL, Wu YF, Yang ZJ, Henikoff SG, Henikoff S, Jiang JM. The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. *Proc. Natl. Acad. Sci. USA.* 2013;110(50):E4875–E4883. <https://doi.org/10.1073/pnas.1319548110>

Zhang WL, Yi CD, Bao WD, Liu B, Cui JJ, Yu HX, Cao XF, Gu MH, Liu M, Cheng ZK. The transcribed 165-bp CentO satellite is the major functional centromeric element in the wild rice species *Oryza punctata*. *Plant Physiol.* 2005;139(1):306–315. <https://doi.org/10.1104/pp.105.064147>

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, Birchler JA, Jiang JM, Dawe RK. Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell.* 2002;14(11):2825–2836. <https://doi.org/10.1105/tpc.006106>