Assembly of the 81.6 Mb centromere of pea chromosome 6 elucidates the structure and evolution of metapolycentric chromosomes

Jiří Macas[1,*], Laura Ávila Robledillo[1], Jonathan Kreplak[2], Petr Novák[1], Andrea Koblížková[1], Iva Vrbová[1], Judith Burstin[2], and Pavel Neumann[1]

[1]Biology Centre, Czech Academy of Sciences, Institute of Plant Molecular Biology, Branišovská 31, České Budějovice, CZ-37005, Czech Republic

[2]Agroécologie, AgroSup Dijon, INRA, Univ. Bourgogne, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France

[*]Corresponding author: e-mail: macas@umbr.cas.cz; phone: +420 387775516

Keywords: centromere evolution; CENH3 chromatin; satellite DNA; chromosome painting; *Pisum sativum*

Author contributions: J.M. designed the study; J.M., L.A.R., A.K., and I.V. performed the experiments; J.M., J.K., P. No., J.B., and P. Ne. analyzed the data; J.M. wrote the paper with input from all co-authors.

Competing interest statement: The authors declare no competing interest.

Abstract

Centromeres in the legume genera *Pisum* and *Lathyrus* exhibit unique morphological characteristics, including extended primary constrictions and multiple separate domains of centromeric chromatin. These so-called metapolycentromeres resemble an intermediate form between monocentric and holocentric types, and therefore provide a great opportunity for studying the transitions between different types of centromere organizations. However, because of the exceedingly large and highly repetitive nature of metapolycentromeres, highly contiguous assemblies needed for these studies are lacking. Here, we report on the assembly and analysis of a 177.6 Mb region of pea (*Pisum sativum*) chromosome 6, including the 81.6 Mb centromere region (CEN6) and adjacent chromosome arms. Genes, DNA methylation profiles, and most of the repeats were uniformly distributed within the centromere, and their densities in CEN6 and chromosome arms were similar. The exception was an accumulation of satellite DNA in CEN6, where it formed multiple arrays up to 2 Mb in length. Centromeric chromatin, characterized by the presence of the CENH3 protein, was predominantly associated with arrays of three different satellite repeats; however, five other satellites present in CEN6 lacked CENH3. The presence of CENH3 chromatin was found to determine the spatial distribution of the respective satellites during the cell cycle. Finally, oligo-FISH painting experiments, performed using probes specifically designed to label the genomic regions corresponding to CEN6 in *Pisum*, *Lathyrus*, and *Vicia* species, revealed that metapolycentromeres evolved via the expansion of centromeric chromatin into neighboring chromosomal regions and the accumulation of novel satellite repeats. However, in some of these species, centromere evolution also involved chromosomal translocations and centromere repositioning.

Significance

Despite their conserved function, plant centromeres exhibit considerable variation in their morphology and sequence composition. For example, centromere activity is restricted to a single region in monocentric chromosomes, but is distributed along the entire chromosome length in holocentric chromosomes. The principles of centromere evolution that led to this variation are largely unknown, partly due to the lack of high-quality centromere assemblies. Here, we present an assembly of the pea metapolycentromere, a unique type of centromere that represents an intermediate stage between monocentric and holocentric organizations. This study not only provides a detailed insight into sequence organization, but also reveals possible mechanisms for the formation of the metapolycentromere through the spread of centromeric chromatin and the accumulation of satellite DNA.

Introduction

Centromeres are chromosomal regions that facilitate faithful chromosome segregation during cell division by serving as an anchor point for the assembly of the kinetochore, a protein complex that connects centromeric chromatin to spindle microtubules (Musacchio and Desai 2017). In most species, the position of the centromere on chromosomes is determined epigenetically by the presence of the centromere-specific histone variant CENH3 (also called CENP-A) and other proteins comprising the constitutive centromere-associated network (Hara and Fukagawa 2017). Despite their conserved function, eukaryotic centromeres are highly variable in size, structure, and sequence composition, a phenomenon called the centromere paradox (Henikoff et al. 2001).

Centromeres exhibit two distinct types of organization, which influence the overall morphology of chromosomes (Schubert et al. 2020). They are either restricted to a single specific region that forms a primary constriction during mitosis (monocentric chromosomes) or distributed along the entire chromosome length (holocentric chromosomes). Species with monocentric chromosomes are more common and presumably ancestral. Several phylogenetic lineages of animals and plants have independently transitioned to holocentricity (Melters et al. 2012). Recently, another type of centromere organization has been described in the legume genera *Pisum* and *Lathyrus* (Neumann et al. 2012; Neumann et al. 2015). These species possess "metapolycentric" chromosomes characterized by extended primary constrictions, which account for up to one-third of the chromosome length in metaphase and contain multiple domains of centromeric chromatin characterized by the presence of CENH3. These CENH3 domains are located along the outer periphery of the primary constriction and interact with the mitotic spindle; however, the interior of the constriction consists of CENH3-free chromatin. This morphology, together with the distribution of certain histone phosphorylation marks (Neumann et al. 2016), strongly resembles chromatin organization on holocentric chromosomes, suggesting that metapolycentric chromosomes may represent an intermediate state between monocentric and holocentric chromosomes (Neumann et al. 2016; Schubert et al. 2020). Thus, metapolycentric chromosomes provide a unique opportunity for studying the changes associated with the transition between different centromere organizations.

The molecular and evolutionary mechanisms leading to centromere variation remain poorly understood, because of difficulties in sequencing and assembling centromeric regions (Peona et al. 2018). Deciphering the complete nucleotide sequence of centromeres in plants is complicated by the large size of these genome regions and their accumulation of highly repetitive DNA sequences such as long-terminal repeat (LTR)-retrotransposons and satellite DNA (satDNA) (Hartley and O'Neill 2019). In particular, satDNA is a major obstacle to the gapless assembly of centromeres because it is arranged in megabase-sized arrays of almost identical, tandemly arranged monomers. At the same time, satDNA is of particular interest because it is known to be a key sequence component that interacts with CENH3 proteins in many centromeres (Talbert and Henikoff 2020).

Recent advances in sequencing, computational, and cytogenetic techniques have ushered in a new era of centromere research. In this regard, the so-called long-read sequencing technologies, which include the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) platforms, have provided a real breakthrough by offering the ability to generate "ultralong" reads that can efficiently resolve satellite repeats. The utility of these technologies, together with novel scaffolding and computational approaches specifically tailored to repeat-rich genomic regions, was best demonstrated by the completion of the gapless assembly of all human centromeres (Altemose et al. 2022; Nurk et al. 2022). Complete centromere assemblies have also been recently reported for several species of higher plants, including maize (*Zea mays*) (Liu et al. 2020; Hufford et al. 2021), Arabidopsis (*Arabidopsis thaliana*) (Naish et al. 2021; Wang et al. 2022), and rice (*Oryza sativa*) (Song et al. 2021), while near-complete assemblies have been achieved in additional species such as tomato (*Solanum lycopersicum*) (Rengs et al. 2022). Despite these advances, the number of species with centromere assemblies is still very limited and does not reflect centromere variation in higher plants.

In this study, we constructed the centromere assembly of garden pea (*Pisum sativum* L. cv. Cameor), a species with metapolycentric chromosomes. In addition to their exceptional organization, the centromeres of pea are populated with a large number of different satellite repeats (Neumann et al. 2012; Ávila Robledillo et al. 2020), which is in contrast to plant species studied previously, which showed only one or few satellites occupying the centromeres of all chromosomes. Although the first genome draft of the same pea genotype is available (Kreplak et al. 2019), it lacks most of the repeat-rich centromeric regions because of the inherent limitations of the short-read sequencing technology used to generate this assembly. To overcome this limitation, we used long-read sequencing technologies to generate new sequence data, which were assembled and verified using a combination of bioinformatics and cytogenetic approaches. We selected the centromere of pea chromosome 6 (CEN6) for this study because this chromosome has the largest primary constriction (estimated at 70–100 Mb) carrying multiple satellite repeats associated with CENH3 chromatin (Neumann et al. 2012). The assembly was used to address the following: (1) how CEN6 differs in sequence composition and long-range organization from its neighboring chromosome arms and from the centromeres of other plant species, (2) how the linear sequence of metapolycentromere transforms into the specific three-dimensional structure observed on pea metaphase chromosomes; and (3) whether metapolycentromeres arise from regional centromeres by spreading of CENH3 chromatin to neighboring chromosomal regions or by expansion due to the accumulation of repetitive DNA.

Results

Assembly of pea CEN6

We performed long-read sequencing, together with extensive manual curation and assembly verification by cytogenetic mapping, to obtain a highly contiguous and reliable sequence of CEN6 (*SI Appendix*, Fig. S1). First, we optimized the protocol for generating long nanopore reads from pea. This resulted in 119.6 Gb (27.8× coverage) of sequence data represented by reads ranging 30–801 kb in length (N50 = 83.8 kb). A portion of the ultralong reads (>120 kb, 8.5× coverage, N50 = 171.7 kb) were then used to create scaffolds, starting with reads containing single-copy marker sequences mapped cytogenetically or genetically to CEN6 or with reads containing CEN6-specific satellite repeats. These "seed" reads were gradually extended by repeated semiautomated identification of terminally overlapping ultralong reads in both directions until scaffolds from adjacent seeds were merged. This procedure was relatively laborious because of the manual curation involved, but it allowed us to obtain verified scaffolds free of structural misassemblies that often affect repeat-rich regions. In the next step, contigs generated from highly accurate PacBio HiFi reads (73.1 Gb; 17× coverage) using two alternative assemblers (HiCanu and Hifiasm) were compared with the nanopore scaffolds. With the exception of two missing duplications (306 kb and 5,243 kb), there were no large structural discrepancies between the HiFi contigs and the nanopore scaffolds, with identical long-range structures of several satDNA arrays of up to 2 Mb in length. Moreover, some highly homogenized satDNA arrays that could not be scaffolded with nanopore reads were fully assembled from the HiFi reads. This result justified the use of HiFi contigs for scaffolding the remaining regions not covered by nanopore scaffolds (*SI Appendix*, Fig. S1) and for using HiFi reads to polish the entire assembly. During and after the scaffolding process, the assembly was verified by multicolor fluorescence in situ hybridization (FISH) mapping of selected satellite repeats and single-copy markers on pea chromosome 6 at different levels of condensation (pachytene, prometaphase, and metaphase). This approach resulted in a 177,603,725 bp-long assembly of the entire CEN6 and its adjacent chromosomal regions, with only a single gap located in one of the FabTR-10 satellite arrays (Fig. 1A,B).

Structure and sequence composition of CEN6

The assembly was annotated with respect to all major types of genomic sequences, including genes, tandem repeats, and various groups of transposable elements. We also generated chromatin immunoprecipitation-sequencing (ChIP-seq) reads using antibodies for both variants of the pea CENH3 protein to analyze the distribution of centromeric chromatin along the CEN6 sequence. This revealed multiple distinct regions of CENH3 accumulation up to ~1 Mb in length (Fig. 1C). Because the transition of primary constriction to chromosome arms on metaphase chromosome 6 is marked by the positions of the outermost CENH3 loci (Fig. 1A), the positions of the first and last CENH3 peaks were used to define an 81.6 Mb region in the assembly corresponding to the primary constriction (Fig. 1B). Mapping the molecular marker sequences from the pea genetic map (Tayeh et al. 2015) onto the assembly revealed that the annotated constriction overlapped with the nonrecombining region of the linkage group LGII, further confirming its correct placement in the assembly (*SI Appendix*, Fig. S1).

The locations showing the highest accumulation of CENH3, which appeared as peaks in the ChIP-seq analysis track, were always associated with satDNA arrays (Fig. 1C,D). These arrays included FabTR-10 repeats, which were located at multiple positions in CEN6, and FabTR-48 and FabTR-49, each of which occupied only a single locus. By contrast, other large satellites in CEN6, such as FabTR-85, -106, and -107, with arrays up to 2 Mb in size, were free of CENH3. Pea contains two variants of the CENH3 protein that differ in sequence and can be distinguished with specific antibodies (Neumann et al. 2016). The use of these two antibodies in ChIP-seq experiments revealed that the distribution patterns of the two CENH3 variants were identical (*SI Appendix*, Fig. S2).

The primary constriction showed no significant difference in sequence composition when compared with the adjacent assembly regions representing the proximal parts of the short and long arms of chromosome 6, except for the accumulation of satDNA (Fig. 1E). LTR-retrotransposons, including the lineage of Ty3/gypsy Ogre elements, a dominant repeat in the pea genome, showed uniform distribution along the entire assembly. Similar distributions were exhibited by Ty1/copia elements and DNA transposons. The lineage of Ty3/gypsy CRM elements, known to target plant centromeres (Neumann et al. 2011), was found partially enriched in the constriction; however, these elements occur in the pea genome only in hundreds of copies and therefore have no significant effect on centromere composition. Annotation of the centromeric DNA revealed 602 genes, which were supported by the RNA-seq data, indicating that these genes were transcriptionally active. The gene density in the centromere was 7.4/Mb (or 8.3/Mb, excluding regions with satDNA arrays), which was lower than that in the adjacent chromosome arms (12.0/Mb).

Since the tools for analyzing DNA methylation in nanopore reads have recently become available (Ni et al. 2021), we examined the frequencies of cytosine methylation in all three contexts known from higher plants. DNA methylation profiles were generally similar between the centromere and chromosome arms, and were characterized by strong cytosine methylation in CG and CHG contexts, and mostly unmethylated CHH motifs in both regions (Fig. 1F and *SI Appendix*, Figs. S3A and S3D). However, there were some notable exceptions, such as a portion of the satDNA arrays, which were hypomethylated compared with the average patterns. This was most evident in the CHG motifs in FabTR-10 and FabTR-106, and in the CHH motifs in FabTR-107 (*SI Appendix*, Fig. S3B,C). In the case of FabTR-10, variation was detected among arrays located at different parts of the centromere, with arrays located near the centromere-chromosome arm junction being the most hypomethylated. Apart from these large blocks of satDNA, detailed inspection of methylation profiles along the assembly revealed smaller regions of reduced methylation, with a part of these regions overlapping with or adjacent to the genes. This finding was also reflected in the gene methylation frequency histograms, which showed hypomethylation of a substantial proportion of CG and CHG motifs, and high levels of methylation in the remaining motifs, resulting in a bimodal histograms (*SI Appendix*, Fig. S3D). No difference was observed between the methylation patterns of genes located within the centromere and those located in chromosome arms.

Homogenization patterns of satDNA arrays

Similarities among monomers within individual satDNA arrays and between multiple arrays of the same repeat are shown in Fig. 2. The major satellite repeat of CEN6, FabTR-10, consisted of eight arrays (a1–a8; 230–893 kb in length), all of which were associated with CENH3 chromatin (Fig. 1C,D). The pea genome contains two main families of FabTR-10, FabTR-10-PST-A and FabTR-10-PST-B, which differ in monomer length (459 and 1,975 bp, respectively) (Ávila Robledillo et al. 2020). Although there was some variation in monomer lengths in FabTR-10 (not shown), all CEN6 arrays could be assigned to the FabTR-10-PST-A family. Additionally, dot plots of sequence similarity showed that homogenization of FabTR-10 monomers mainly occurred within individual arrays or their parts, resulting in sequence divergence between arrays at different loci (Fig. 2). The only exception was the high sequence similarity between the adjacent arrays a7 and a8, indicating that these arrays originated following a recent duplication and inversion event. The orientation of monomers was uniform within each array, except in a2, which contained an inversion of a portion of the array. However, the monomers showed no preferred orientation throughout the centromere. Interestingly, the binding to CENH3 was relatively uniform across the arrays, regardless of the degree of sequence homogenization and methylation or the presence of particular sequence variants of FabTR-10 (*SI Appendix*, Fig. S4).

Each of the remaining six satellites analyzed occupied a single locus in CEN6. Only two of these satellites, FabTR-48 and FabTR-49, were associated with CENH3. No major differences were observed in array homogenization patterns between CENH3-associated satellites, including FabTR-10, -48 and -49, and non-CENH3 satellites, as both groups showed patchy dot-plot patterns indicative of regions within the arrays with increased local sequence homogenization. In general, there were no trends of higher sequence homogenization at the center of the arrays. The FabTR-107 and FabTR-85 arrays showed patterns of long parallel lines, indicating segmental duplications of large portions of these arrays (Fig. 2).

Spatial arrangement of CEN6 during mitosis and interphase

We employed FISH with satDNA probes as cytogenetic landmarks to examine how the primary sequence of CEN6 transforms into the three-dimensional structure of the metapolycentromere during mitosis. The results showed that satDNA arrays associated with CENH3 domains are located along the outer periphery of the primary constriction, as required for the interaction of CENH3 chromatin with the kinetochore and mitotic spindle (Fig. 3A). Each of the FabTR-48- and FabTR-49-specific probes produced a single fluorescent spot, corresponding to their respective single loci in the assembly. The probe for the major CENH3-associated repeat, FabTR-10, generated signals along the entire length of the constriction; however, the number of signals did not exactly match the number of FabTR-10 arrays in the assembly, indicating the fusion of signals from proximally positioned arrays. In contrast to the CENH3-associated repeats, the arrays of the other large satellites (FabTR-85, -106, and -107) were observed predominantly within chromatids, often near the chromosome axis, or as linear signals across the chromatid width (Fig. 3B). This may be because chromatin is packed into megaloops, with CENH3 domains driven to the periphery of the constriction and the non-CENH3 chromatin constituting its interior.

Simultaneous detection of CENH3 and satellite repeats by immuno-FISH in nuclei showed that, in contrast to their multidomain structure on metaphase chromosomes, all CENH3 domains aggregated into a single spot per interphase chromosome, resulting in 14 CENH3 spots per nucleus (Fig. 3C). Consequently, FISH signals from CENH3-associated satellites overlapped with these spots (data not shown). However, FISH signals from satellite repeats not associated with CENH3, such as FabTR-85, -106, and -107, were found relatively far from the CENH3 spots, suggesting that these satellites were located on decondensed chromatin loops emanating from the densely packed CENH3 domains (Fig. 3D). Overall, these experiments revealed that the spatial arrangement and condensation of different parts of the centromere sequence during the cell cycle differ, depending on their association with CENH3 chromatin.

Elucidation of CEN6 evolution in Fabeae using oligo-FISH painting probes

Taking advantage of the CEN6 assembly, we designed a set of FISH painting probes based on oligo pools derived from single-copy regions in the assembly (Fig. 4A). Two probes were designed for the primary constriction, covering either its entire length (probe PS6-C; 8,915 oligos) or a specific 3.7 Mb region within the constriction (probe PS6-C1.8; 1,800 oligos). The third probe was designed to label the regions of both the long and short arms of chromosome 6 directly adjacent to the constriction (probe PS6-A; 19,250 oligos). Despite the low average density of hybridizing oligos (0.12 oligos/kb in PS6-C and 0.26 oligos/kb in PS6-A), the probes produced relatively uniform and specific signals at their target regions (Fig. 4B,C and *SI Appendix*, Fig. S5).

To elucidate the evolution of metapolycentric chromosomes, we used the painting probes to identify the regions homoeologous to pea CEN6 in the chromosomes of selected Fabeae species (Fig. 4C). In *Pisum fulvum*, the species most closely related to pea, the PS6-C probe labeled the entire constriction on one chromosome pair, with signal extending into the short arm. The signal from the PS6-A probe was correspondingly shifted, confirming that the region corresponding to the *P. sativum* CEN6 constriction short-arm junction was within the short arm of *P. fulvum* chromosome 6. This observation of the shorter constriction, based on chromosomal morphology, was confirmed by CENH3 immunolabeling (*SI Appendix*, Fig. S5B).

We then examined representatives of the genus *Lathyrus*, which is known to share metapolycentric chromosome morphology with *Pisum*, although the size of the primary constriction varies considerably among *Lathyrus* species (Neumann et al. 2015). In *L. clymenum*, which has chromosomes with relatively short constrictions, the painting probes hybridized to a single chromosome pair, although signal intensity was weaker than that observed in *Pisum*. The probes produced the expected pattern, i.e., a single green band (PS6-C) located between two red bands (PS6-A), one on either side; however, this pattern was shifted from the centromere (as observed in *P. sativum*) into the long chromosome arm (Fig. 4C). The same results were obtained for the closely related *L. ochrus*. By contrast, *L. sativus*, which has extremely elongated centromeres, showed signals that overlapped with primary constrictions on a pair of chromosomes. However, the PS6-C signal did not cover the entire constriction, leaving out the region adjacent to the short arm, and contained a large unlabeled gap within the constriction. Considering the signal of the PS6-A probe and simultaneous hybridization with the FabTR-2 probe, which marks the positions of CENH3 chromatin in *L. sativus* (Ávila Robledillo et al. 2020), we concluded that the constriction on this chromosome extends into the region corresponding to the short arm of pea chromosome 6. In addition, further experiments using *L. sativus* satDNA probes developed previously (Vondrak et al. 2020) revealed that the gap in the PS6-C signal was caused by the amplification of the FabTR-54 repeat, which is not present in *P. sativum* (Fig. 4D).

To complement our study with related Fabeae species that possess monocentric chromosomes, we applied the *P. sativum* oligo-FISH probes to two *Vicia* species: *V. tetrasperma*, which is phylogenetically closely related to the *Pisum*/*Lathyrus* clade, and *V. faba* (Fig. 4C). The signals from the probes were more difficult to detect. In *V. faba*, the green signal (PS6-C) was completely absent, probably because it is the most distant to *P. sativum* and has a larger genome, and only weak red signals (PS6-A) were detected in the long- and short-arm regions surrounding the centromere of chromosome 3. In *V. tetrasperma*, the probes labeled centromeric regions of two chromosome pairs, indicating chromosomal rearrangements such as the reciprocal translocation of short arms.

Discussion

Centromeres represent the final frontiers of genome projects because of their high contents of satellite repeats, which in principle are extremely difficult to assemble. However, the recent introduction of accurate long-read sequencing technologies and advanced assembly strategies has led to gapless assemblies of several complex genomes, ushering in a new era in centromere research. In plants, complete centromere assemblies have been constructed only for monocentric species to date, including maize (Liu et al. 2020; Hufford et al. 2021), rice (Song et al. 2021) and *Arabidopsis thaliana* (Naish et al. 2021; Wang et al. 2022). In addition, high-quality assemblies of three holocentric species belonging to the *Rhynchospora* genus recently became available (Hofstatter et al. 2022). Here, we report the assembly of a genomic region representing yet another type of centromere organization, namely metapolycentromere, in the pea cultivar Cameor. Except a single gap in one of the satDNA arrays, the assembly is without gaps, providing the most detailed sequence information lacking in previous studies of metapolycentromeres, which mainly used cytogenetic approaches (Neumann et al. 2012; Neumann et al. 2015; Neumann et al. 2016; Ávila Robledillo et al. 2020). Similar to the previously reported complete assemblies of human and plant genomes, the contiguity of CEN6 assembly was enabled by the use of highly accurate long reads (PacBio HiFi), which enabled the reconstruction of most satDNA arrays, and by combining the assembly with physically localized cytogenetic markers. A unique feature of our study was the use of ultralong nanopore reads for creating manually curated scaffolds for most of the assembly, since the repetitive and complex structure of pea centromeres makes them prone to misassemblies. This makes our CEN6 assembly superior in completeness and contiguity even to the novel high-quality genome assembly of the pea cultivar ZW6 (Yang et al. 2022) (data not shown), which was published during preparation of this manuscript.

It has been speculated that metapolycentromeric chromosomes represent an intermediate state between monocentric and holocentric chromosomes (Neumann et al. 2012; Neumann et al. 2015). Monocentric chromosomes are generally characterized by an uneven distribution of genomic features along their length, with centromeric and pericentromeric regions showing greater repetitive DNA accumulation, lower gene density, and different epigenetic profiles than the chromosome arms. By contrast, holocentric chromosomes show a more homogeneous distribution of repeats, genes, and histone modifications (Hofstatter et al. 2022). For example, during mitosis, histone H2A phosphorylation at Thr120 (H2AT120ph) is detected across almost the entire length of holocentric chromosomes but is restricted to the (peri)centromeres in monocentric chromosomes (Schubert et al. 2020). In this respect, pea CEN6 is more similar to holocentromeres, as we did not detect significant differences in the distribution of genes and most repeats between the constriction and neighboring chromosome arms. It is also noteworthy that H2AT120ph and histone H3 phosphorylation marks H3T3ph, H3S10ph, and H3S28ph have been shown to extend throughout the entire constrictions of *P. sativum* and *L. sativus* metapolycentric chromosomes (Neumann et al. 2016). On the other hand, several satDNA families accumulate in CEN6, forming long arrays, some of which are associated with CENH3. Arrays of centromeric satellites up to several megabasepairs in length are typical of monocentric chromosomes, whereas holocentric chromosomes either lack CENH3-associated satellites (Heckmann et al. 2013) or have them distributed as multiple short arrays across their length (Hofstatter et al. 2022).

Although information on the long-range structure, methylation profiles, and CENH3-binding ability of centromeric satellites along the fully assembled arrays is still sparse, several common features have been reported for human alpha satellites, *Arabidopsis* CEN180, and rice CentO, including (1) the presence of chromosome-specific variants of centromeric satellites; (2) homogenization of satellite sequences within each array, often resulting in the highest similarity at the centers of arrays; (3) nonuniform binding of CENH3 along the arrays; and (4) hypomethylation of array regions associated with CENH3 (Naish et al. 2021; Song et al. 2021; Altemose et al. 2022; Gershman et al. 2022; Wang et al. 2022). On the other hand, CENH3 chromatin is largely restricted to the centromeric satellite arrays in humans and *Arabidopsis*, whereas this association is not as tight in rice, where most of the CENH3 is located outside the CentO arrays in some centromeres (Song et al. 2021). The centromeres of maize differ even more substantially; several chromosomes lack the centromeric satellite CentC, and CENH3 shows no preferential binding to CentC or to other repeats (Liu et al. 2020), suggesting that these limited observations cannot be generalized.

Our characterization of pea CEN6 provides further evidence for the diversity in plant centromeres. Instead of a single type of satellite repeat, the pea genome contains multiple distinct satellite sequences, three of which are associated with CENH3. Unlike the above-mentioned species (*Arabidopsis*, rice, human), we observed no evidence of preferential sequence homogenization in the centers of satDNA arrays in pea, regardless of their association with CENH3. Moreover, CENH3 enrichment profiles in pea were relatively uniform along the arrays, despite their sequence variation. These observations suggest that, unlike human or *Arabidopsis* centromeres, the association of CENH3 with pea centromeric satellites is not determined by their sequence. The occurrence of multiple centromeric satellites and their rapid turnover is common in Fabeae species (Ávila Robledillo et al. 2020), implying that their evolution cannot be explained by the centromere drive model (Henikoff et al. 2001), which requires the presence of a single centromeric satellite. The question of what features make some of the pea CEN6 satellites competent for CENH3 binding remains unanswered, even considering their variation in cytosine methylation patterns (Fig. 1 and *SI*

Appendix, Fig. S3), because we could not detect any methylation profiles that would consistently distinguish between arrays associated with CENH3 from those not associated with CENH3. For example, only some of the CENH3-binding FabTR-10 arrays were hypomethylated, but hypomethylation was also detected in some CENH3-less satellites such as FabTR-106 and FabTR-107.

One of the most intriguing questions that could be addressed, owing to the availability of the centromere assembly, is the origin and evolution of metapolycentric chromosomes. We approached this problem by developing oligo-pool FISH painting probes to identify regions orthologous to pea CEN6 in related Fabeae species. These experiments revealed the highly dynamic nature of centromere evolution in Fabeae, characterized by centromere shifts, chromosome translocations, and the expansion (and perhaps contraction) of primary constrictions. Our results support the view that the expansion of metapolycentromeres is facilitated mainly by the spreading of CENH3 chromatin from the centromere into adjacent chromosome arms. However, the factor(s) triggering this process and the molecular mechanisms involved remain to be elucidated.

Insights into the possible mechanisms involved in metapolycentromere formation could be obtained from centromere shifts reported in monocentric chromosomes (see (Montenegro et al. 2022) and references therein). These centromere shifts are explained either by chromosomal rearrangements such as translocations or inversions or by the repositioning of centromeric chromatin to a new location without disrupting the linear order of chromosomes (Schubert 2018). Uncovering the exact mechanisms, especially in the case of centromere repositioning, depends on the availability of gapless genome assemblies of related genotypes that differ in centromere position, as defined by their CENH3 distribution. Such efforts have been initiated in the pangenome studies of maize and wheat (*Triticum aestivum*), where centromere shifts have been detected in some of the genotypes examined (Walkowiak et al. 2020; Hufford et al. 2021). In addition, Xue and colleagues conducted a detailed investigation of the formation of a new centromere domain on rice chromosome 8 (Xue et al. 2022), and showed that the formation of this domain was triggered by the deletion of a part of the existing centromere including the CentO array. The new domain arose in a nearby genomic region, which contained increased amounts of CENH3 in the wild-type genotype. Thus, this mechanism can generate centromeres with multiple CENH3 domains, similar to metapolycentric chromosomes. However, compared with rice, the CENH3 domains in the pea CEN6 metapolycentromere are much more widely spaced and are all confined to satDNA arrays. Another mechanism, based on the mobilization of CENH3-associated centromeric satellite Tyba by Helitron elements, has been proposed to facilitate the spread of centromeric chromatin in holocentric *Rhynchospora* species (Hofstatter et al. 2022). However, this is unlikely to occur in pea centromeres because CENH3-associated satellites in the pea genome are organized in a few large arrays, unlike the centromeric satellites of *Rhynchospora*, which exist as a large number of scattered and much shorter loci that may be embedded in functional Helitron elements.

The only mechanism we have identified thus far that may favor the propagation of CENH3 domains in metapolycentromeres and is supported by our sequence data is that of segmental duplications, which are frequent in some plant centromeres (Ma and Jackson 2006). The larger of the two segmental duplications identified in pea CEN6 originated from the region between simple sequence repeat (SSR)-like arrays and FabTR-10 arrays, and contained portions of these arrays in the duplicated sequence. Because FabTR-10 repeats are associated with CENH3, a new but relatively small (73 kb) CENH3 domain was generated 5.2 Mb downstream of the original array. However, this mechanism cannot explain the origin of other CENH3 loci because no traces of sequence duplications were detectable at these loci. Thus, segmental duplication could be just one of several synergistic forces driving the evolution of metapolycentric chromosomes.

To gain further insight into the rapid and divergent evolution of centromeres in Fabeae, several research directions are conceivable. A new improved version of the whole-genome sequence of pea cv. Cameor, based on the sequence data and methods described in this study, is currently under construction and is expected to provide near-complete assemblies of the remaining six centromeres. Sequence comparison of these centromeres with CEN6 (described here) will enable the identification of common features of evolutionary or functional significance. This approach will be further strengthened by the inclusion of the highly contiguous genome assemblies of related species, such as *L. sativus* (metapolycentric) and *V. faba* (monocentric), which are also in progress (Jayakodi et al. 2022). In addition to the investigation of centromere properties, these assemblies should also be used for the comparative analysis of kinetochore genes to reveal any differences in kinetochore composition among species with different centromere organization. The rationale for this approach stems from the finding that the transition to holocentricity in some groups of organisms is accompanied by the loss or multiplication of CENH3 or other kinetochore genes (Drinnenberg et al. 2014; Cortes-Silva et al. 2020; Oliveira et al. 2020), similar to the duplication and diversification of CENH3 genes in *Pisum* and *Lathyrus* (Neumann et al. 2015).

Materials and methods

Genomic DNA preparation and sequencing

High molecular weight (HMW) DNA was prepared from the nuclei extracted, and subsequently purified, from the young leaves of pea (*Pisum sativum* L. cv. 'Cameor') seedlings, as described previously (Vondrak et al. 2020). The quality of DNA preparations was checked using field inversion gel electrophoresis (FIGE) to ensure that the DNA fragment size was >100 kb. Then, 3–40 µg of input HMW DNA was subjected to 20 runs of nanopore sequencing on the MinION sequencer (Oxford Nanopore Technologies) using the following library preparation kits, according to the manufacturer's instructions: SQK-LSK109 (13 runs), SQK-LSK110 (1 run), SQK-RAD004 (3 runs), and SQK-ULK001 (3 runs). Raw nanopore reads were basecalled using Oxford Nanopore basecaller Guppy (ver. 3.6.0 and 4.5.4). Quality-filtering of the resulting FastQ reads and their conversion to FASTA format were performed with BBDuk (part of BBTools, https://jgi.doe.gov/data-and-tools/bbtools/) using the quality cutoff parameter maq = 8. Reads shorter than 30 kb were discarded. PacBio HiFi reads were generated from the same input HMW DNA by DNA Sequencing Center of the Brigham Young University (UT, USA) using four SMRT Cells on a PacBio Sequel II instrument by running the Circular Consensus Sequencing (CCS) protocol for 30 h.

CEN6 scaffolding and assembly

A fraction of the ultralong nanopore reads (>160 kb) was used to create scaffolds covering most of the assembled region. The scaffolding process was initiated by identifying "seed" nanopore reads, which contained sequences of genetic markers located in the nonrecombining region of linkage group LGII in the pea high-density genetic map (Tayeh et al. 2015). A portion of these marker sequences were also detected on metaphase chromosomes with the highly sensitive FISH protocol, which was used to determine their exact physical location (*SI Appendix*, Fig. S1). Additional physically localized seed reads were derived from the edges of the arrays of satellite repeats, FabTR-48, -49, and -50, which were previously shown to be specific to CEN6 (Neumann et al. 2012; Ávila Robledillo et al. 2020). Next, the seed reads were extended in both 5' and 3' directions by searching the database of ultralong reads using BLASTN (Altschul et al. 1997) and minimap2 (Li 2018) for similarities with their 60 kb terminal regions. The identified read overlaps were verified by sequence similarity dot plots automatically generated using Gepard (Krumsiek et al. 2007) and by manual inspection, ensuring that the extending read sequence was confirmed by at least one other overlapping read. Eventually, if the extending or confirming reads could not be obtained from the longest fraction, collections of reads shorter than 160 kb were searched. The verified extending reads were then merged with the seed reads to form initial scaffolds. This process was then iterated using the end regions of scaffolds as queries in the next round of similarity searches and extensions until two adjacent scaffolds were merged. Alternatively, the extensions were stopped when the scaffolds reached highly homogenized regions of some satellite repeats that prevented the reliable selection of overlapping reads, because of the relatively high error rate of nanopore reads. On the other hand, higher sequence variation and the presence of mobile element insertions in satellite arrays allowed them to be reliably scaffolded with long nanopore reads.

The assembly of HiFi reads was performed using Hifiasm assembler (Cheng et al. 2021) version 0.16.1, with default parameters. Alternatively, HiCanu (Nurk et al. 2020) version 2.1.1 was used with the options “genomeSize=4.2G useGrid=false -pacbio-hifi”. Contigs from the HiFi assemblies were used to cover the regions that were not scaffolded using nanopore reads (mostly within the long arm of chromosome 6, *SI Appendix*, Fig. S1). The HiFi contigs were also used to fill gaps in the nanopore scaffolds corresponding to satDNA arrays. With the exception of the a7 array of satellite FabTR-10, which was not fully represented in any HiFi contig, all satDNA arrays were fully assembled and were therefore used to represent these regions in the assembly. Finally, the assembly was polished with HiFi reads using Racon version 1.4.20 (Vaser et al. 2017).

Assembly annotation

Annotation of repetitive sequences was performed using a combination of different tools available on the RepeatExplorer Galaxy Server (https://repeatexplorer-elixir.cerit-sc.cz/). Transposable element sequences encoding conserved protein domains were identified based on their similarities to the REXdb database (Neumann et al. 2019) using DANTE (https://github.com/kavonrtep/dante). Full-length LTR-retrotransposon sequences were annotated using the DANTE_LTR tool (https://github.com/kavonrtep/dante_ltr), which combines the results of DANTE with similarity- and structure-based identification of LTR-retrotransposon signatures such as LTRs, primer binding sites (PBSs), and target site duplications (TSDs). The identified full-length LTR-retrotransposons were also used to create a reference database for similarity-based annotation of repeats in the assembly. The database was also enriched with consensus sequences of repeats obtained from the RepeatExplorer analysis of Fabeae genomes (Macas et al. 2015) and with a collection of Fabeae satDNA sequences compiled on the basis of our previous studies (Macas et al. 2015; Ávila Robledillo et al. 2020; Vondrak et al. 2020). In parallel with similarity-based detection, tandem repeats were identified, based on their genomic organization, with Tandem Repeats Finder ver. 4.09 (Benson 1999) using the parameters "2 5 7 80 10 500 2000". The output of the search was parsed and converted to GFF format using TRAP (Sobreira et al. 2006).

Gene annotation was performed by launching FINDER (Banerjee et al. 2021) on the CEN6 assembly supplemented with unscaffolded HiFi contigs representative of the remaining parts of the genome. Briefly, 30 RNA-seq libraries (Alves-Carvalho et al. 2015; Henriet et al. 2019) were mapped to the assembly by STAR, and assembled with psi-class (Song et al. 2019). Next, the mapped data were processed by braker2 (Brůna et al. 2021) to perform a de novo annotation of genes. To improve the quality of annotation, Ryūtō (Gatter and Stadler 2021) was run twice on the mapping results, once for the stranded library and the second time for the unstranded library. The results of Ryūtō and psi-class were combined using Mikado (Venturini et al. 2018) to obtain a high-quality (HQ) annotation dataset. A low-quality (LQ) dataset was built by filtering braker2 results as follows. First, genes overlapping a repeat annotation were removed. Then, only the genes with at least one hit in the eggNOG protein database were retained. Functional annotation of these genes was performed using TRAPID with the PLAZA Dicots 4.0 database.

CENH3 ChIP-seq analysis

ChIP experiments were performed with native chromatin as described previously (Neumann et al. 2012), using custom antibodies that specifically recognize one of the two variants of pea CENH3 proteins. DNA fragments were purified from the immunoprecipitated samples, and the corresponding control samples (Input; digested chromatin not subjected to immunoprecipitation) were sequenced on the Illumina platform (Admera Health, NJ, USA) in paired-end, 150 bp mode. Duplicate experiments, including independent chromatin preparations, were performed for each CENH3 variant using either one antibody (P23 for CENH3-2) or two different antibodies (P22 and P43 for CENH3-1); both anti-CENH3-1 antibodies were raised against an identical peptide in rabbit (P22) and chicken (P43), and tested previously (Neumann et al. 2012). The resulting reads were quality-filtered and trimmed using Trimmomatic (Bolger et al. 2014) (minimum allowed length = 100 nt), yielding 122–211 million reads per sample, which were mapped onto the assembly using Bowtie 2 version 2.4.2 (Langmead and Salzberg 2012), with options -p 64 -U. Subsequent analysis was performed on full output from Bowtie2 program and on output where all multimapped reads were filtered out. Filtering of multimapped reads was performed using Sambamba version 0.8.1 (Tarasov et al. 2015) with options "-F [XS] == null and not unmapped and not duplicate". Regions with statistically significant ChIP/Input enrichment ratio were identified by comparing ChIP and Input mapped reads using the epic2 program (Stovner and Sætrom 2019), with the parameter "--bin-size 200". Alternative identification of enrichment was performed using MACS2 (Zhang et al. 2008) version 2.1.1.20160309, with default settings. The ChIP/Input ratio was calculated for plotting purposes using bamCompare (version 3.5.1) from the deepTools package (Ramírez et al. 2016). The program was run with the parameter "–binSize 200" to calculate the log2 ratio for the 200 nt window size. The resulting data were plotted using the rtracklayer package of R (Lawrence et al. 2009).

Methylation analysis

Cytosine methylation was analyzed in all three contexts (CG, CHG, and CHH) by detecting the frequency of 5-methyl cytosine (5mC) in nanopore reads, which were aligned to the CEN6 assembly using DeepSignal-plant ver. 0.1.4 (Ni et al. 2021) with the model "model.dp2.CNN.arabnrice2-1_120m_R9.4plus_tem.bn13_sn16.both_bilstm.epoch6.ckpt". Prior to the analysis, nanopore reads were rebasecalled using the latest version of Guppy (ver. 6.0.1) and resquiggled using Tombo ver. 1.5.1. Methylation frequencies were calculated for each cytosine position in the assembly, based on the number of methylated and methyl-free cytosines detected in the aligned nanopore reads. The methylation analysis pipeline was run on a Linux server equipped with 126 GB RAM, 24 CPUs, and the NVIDIA GeForce GTX 3060 graphics card.

Bioinformatics analysis

Unless stated otherwise, all data handling and bioinformatic analyses were implemented using custom Python, Perl, and R scripts, and executed on a Linux-based server equipped with 256 GB RAM and 48 CPUs.

Centromere painting probe design and FISH

The painting probes were designed on the basis of unique 45 nt oligos, which were selected from specific regions of the CEN6 assembly using the Chorus2 program (Zhang et al. 2021). The probes were then synthesized by Daicel Arbor Biosciences (Ann Arbor, MI, USA) either as myTags Custom Labeled Probes (PS6-C, labeled with Alexa Fluor 488; PS6-A, labeled with ROX) or as myTags Custom Immortal Probe PS6-C1.8, which was subsequently labeled with biotin-16-dUTP, as described previously (Braz et al. 2020). The satDNA-based probes were either synthesized as an oligo-pool probe (oPools$^{\mathrm{TM}}$ Oligo Pools, IDT) or cloned and labeled with Alexa Fluor 568 or 488 (Thermo Fisher Scientific, Waltham, MA, USA) via nick translation (Kato et al. 2006). The cloned probes for single-copy expressed sequence tag (EST)-based genetic markers were labeled with Alexa Fluor 488 or Alexa Fluor 568 (Thermo Fisher Scientific, Waltham, MA, USA) using nick translation.

Mitotic chromosomes used for cytogenetic analyses were prepared from synchronized root apical meristems (Neumann et al. 2015). After cell cycle synchronization, chromosome preparations were obtained using different protocols, depending on their end use: single-copy FISH targets and centromere painting probes (Aliyeva-Schnorr et al. 2015), satDNA-based probes (Ávila Robledillo et al. 2020), or CENH3 immunolabeling (Neumann et al. 2002; Ávila Robledillo et al. 2020). Pachytene chromosomes were extracted from anthers as described previously (Zhong et al. 1996), with some modifications. Flower buds (3–5 mm in size) were collected, fixed in Carnoy's solution

(3:1 ethanol: acetic acid) overnight at room temperature, and then transferred to 70% ethanol and incubated at 4°C until needed for further analysis. After rinsing with distilled water for 5 min, the flower buds were washed twice with 1× citrate buffer for 5 min each time. Finally, the flower buds were dissected, and the anthers were removed and placed on a microscope slide in a drop of 60% acetic acid, where they were squashed under a coverslip.

FISH using painting probes and satDNA-based probes was performed as described previously (Macas et al. 2007), with hybridization and washing temperatures adjusted to account for the probe AT/CG content. Hybridization stringency was modified to allow for 10% mismatches (when hybridized to *P. sativum* chromosomes) or 20–30% mismatches (when hybridized to the chromosome preparations of other species). When performing FISH using painting probes, 3–10 pmol of the probe was used per slide; post-hybridization washes were conducted in 0.1× SSC instead of 50% formamide/2× SSC; and the biotin-labeled PS6-C1.8 probe was detected using streptavidin-Alexa Fluor 488 (Jackson Immunoresearch). FISH using satDNA oligo-pool probes was performed according to the method described previously (Fields et al. 2019), with some modifications. Briefly, after rinsing in 2× SSC, the chromosome preparations were fixed in 45% acetic acid for 4 min, postfixed in 2× SSC containing 4% formaldehyde for 10 min, and washed in 2× SSC for 10 min after each fixation. Following dehydration in an ethanol series (50%, 70%, and 96%), 20 µl of the hybridization mix (50% [v/v] formamide, 10% dextran sulfate in 2× SSC, and 30–100 pmol of the oligo-pool probe) was applied to each slide with chromosome preparations, which was then incubated at 84°C for 3 min to induce DNA denaturation. After 20 h of hybridization, all washes were performed at 37°C. Single-copy FISH was performed as described previously (Aliyeva-Schnorr et al. 2015).

To perform multicolor FISH, up to two rounds of rehybridization were performed. To remove the previously hybridized probes, the slides were washed at room temperature in 4× SSC/0.2% Tween 20 for at least 30 min and twice in 2× SSC for 5 min, then in 50% formamide/2× SSC for 10 min at 55°C, and finally in 2× SSC for 10 min at room temperature. Samples were postfixed before proceeding with the next hybridization. Immunolabeling, combined with FISH, was conducted as described previously (Ávila Robledillo et al. 2020).

Data availability

Raw data used for scaffolding, sequence assembly, and ChIP-seq analysis are available from the European Nucleotide Archive (study accession no. PRJEB54858). The final CEN6 sequence and its annotation are available from the Czech National Repository (DOI: 10.48700/datst.8t29q-nfr77) and from the interactive genome browser JBrowse (http://w3lamc.umbr.cas.cz/lamc/jbrowse.php).

ACKNOWLEDGMENTS

We thank V. Tetourová and J. Látalová for technical assistance, and D. Beránková and E. Hřibová (Institute of Experimental Botany, Olomouc, Czech Republic) for help with labeling myTags Custom Immortal Probe. We also thank the ELIXIR CZ Research Infrastructure Project

(LM2018131) for providing computing and data-storage facilities. This work was funded by the Czech Science Foundation (grant no. GACR 20-24252S).

References

Aliyeva-Schnorr L, Beier S, Karafiátová M, Schmutzer T, Scholz U, Doležel J, Stein N, Houben A. 2015. Cytogenetic mapping with centromeric bacterial artificial chromosomes contigs shows that this recombination-poor region comprises more than half of barley chromosome 3H. *Plant J.* 84:385–394.

Altemose N, Logsdon GA, Bzikadze A V., Sidhwani P, Langley SA, Caldas G V., Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, et al. 2022. Complete genomic and epigenetic maps of human centromeres. *Science* 376, 6588.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Alves-Carvalho S, Aubert G, Carrère S, Cruaud C, Brochot A-L, Jacquin F, Klein A, Martin C, Boucherot K, Kreplak J, et al. 2015. Full-length de novo assembly of RNA-seq data in pea (*Pisum sativum* L.) provides a gene expression atlas and gives insights into root nodulation in this species. *Plant J.* 84:1–19.

Ávila Robledillo L, Neumann P, Koblížková A, Novák P, Vrbová I, Macas J. 2020. Extraordinary sequence diversity and promiscuity of centromeric satellites in the legume tribe Fabeae. *Mol. Biol. Evol.* 37:2341–2356.

Banerjee S, Bhandary P, Woodhouse M, Sen TZ, Wise RP, Andorf CM. 2021. FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinformatics* 22:205.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27:573–580.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.

Braz GT, Yu F, do Vale Martins L, Jiang J. 2020. Fluorescent In Situ Hybridization Using Oligonucleotide-Based Probes. In: In Situ Hybridization Protocols. Methods in Molecular Biology, vol 2148. p. 71–83.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinf.* 3:lqaa108.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18:170–175.

Cortes-Silva N, Ulmer J, Kiuchi T, Hsieh E, Cornilleau G, Ladid I, Dingli F, Loew D, Katsuma S, Drinnenberg IA. 2020. CenH3-independent kinetochore assembly in Lepidoptera requires CCAN, including CENP-T. *Curr. Biol.* 30:561-572.e10.

Drinnenberg IA, DeYoung D, Henikoff S, Malik HS. 2014. Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. *Elife* 3:e03676.

Fields BD, Nguyen SC, Nir G, Kennedy S. 2019. A multiplexed DNA FISH strategy for assessing genome architecture in *Caenorhabditis elegans*. *Elife* 8:e42823.

Gatter T, Stadler PF. 2021. Ryūtō: improved multi-sample transcript assembly for differential transcript expression analysis and more. *Bioinformatics* 37:4307–4313.

Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, Shumate A, Razaghi R, Koren S, et al. 2022. Epigenetic patterns in a complete human genome. *Science* 376:2021.05.26.443420.

Hara M, Fukagawa T. 2017. Critical Foundation of the Kinetochore: The Constitutive Centromere-Associated Network (CCAN). In: Centromeres and Kinetochores. Vol. 112. Springer, Cham. p. 29–57.

Hartley G, O'Neill R. 2019. Centromere repeats: Hidden gems of the genome. *Genes (Basel).* 10:223.

Heckmann S, Macas J, Kumke K, Fuchs J, Schubert V, Ma L, Novák P, Neumann P, Taudien S, Platzer M, et al. 2013. The holocentric species *Luzula elegans* shows interplay between centromere and large-scale genome organization. *Plant J.* 73:555–565.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293:1098–1102.

Henriet C, Aimé D, Térézol M, Kilandamoko A, Rossin N, Combes-Soia L, Labas V, Serre R-F, Prudent M, Kreplak J, et al. 2019. Water stress combined with sulfur deficiency in pea affects yield components but mitigates the effect of deficiency on seed globulin composition. *J. Exp. Bot.* 70:4287–4304.

Hofstatter PG, Thangavel G, Lux T, Neumann P, Vondrak T, Novak P, Zhang M, Costa L, Castellani M, Scott A, et al. 2022. Repeat-based holocentromeres influence genome architecture and karyotype evolution. *Cell* 185:3153-3168.e18.

Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, et al. 2021. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* 373:655–662.

Jayakodi M, Golicz AA, Kreplak J, Fechete LI, Angra D, Bednář P, Bornhofen E, Zhang H, Boussageon R, Kaur S, et al. 2022. The giant diploid faba genome unlocks variation in a global protein crop. *bioRxiv*:2022.09.23.509015.

Kato A, Albert PS, Vega JM, Birchler JA. 2006. Sensitive fluorescence *in situ* hybridization signal detection in maize using directly labeled probes produced by high concentration DNA polymerase nick translation. *Biotech Histochem* 81:71–78.

Kreplak J, Madoui M-A, Cápal P, Novák P, Labadie K, Aubert G, Bayer PE, Gali KK, Syme RA, Main D, et al. 2019. A reference genome for pea provides insight into legume genome evolution. *Nat. Genet.* 51:1411–1422.

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 23:1026–1028.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359.

Lawrence M, Gentleman R, Carey V. 2009. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 25:1841–1842.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100.

Liu J, Seetharam AS, Chougule K, Ou S, Swentowsky KW, Gent JI, Llaca V, Woodhouse MR, Manchanda N, Presting GG, et al. 2020. Gapless assembly of maize chromosomes using long-read technologies. *Genome Biol.* 21:121.

Ma J, Jackson SA. 2006. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.* 16:251–259.

Macas J, Neumann P, Navrátilová A. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics* 8:427.

Macas J, Novák P, Pellicer J, Čížková J, Koblížková A, Neumann P, Fuková I, Doležel J, Kelly LJ, Leitch IJ. 2015. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe *Fabeae*. *PLoS One* 10:e0143424.

Melters DP, Paliulis L V, Korf IF, Chan SWL. 2012. Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. *Chromosom. Res.* 20:579–593.

Montenegro C, do Vale Martins L, Bustamante F de O, Brasileiro-Vidal AC, Pedrosa-Harand A. 2022. Comparative cytogenomics reveals genome reshuffling and centromere repositioning in the legume tribe Phaseoleae. *Chromosom. Res.*

Musacchio A, Desai A. 2017. A molecular view of kinetochore assembly and function. *Biology (Basel)* 6:5.

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, Mandáková T, Jamge B, Lambing C, Kuo P, et al. 2021. The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* 374:eabi7489.

Neumann P, Navratilova A, Koblizkova A, Kejnovsky E, Hribova E, Hobza R, Widmer A, Dolezel J, Macas J. 2011. Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob. DNA* 2:4.

Neumann P, Navrátilová A, Schroeder-Reiter E, Koblížková A, Steinbauerová V, Chocholová E, Novák P, Wanner G, Macas J. 2012. Stretching the rules: monocentric chromosomes with multiple centromere domains. *PLoS Genet.* 8:e1002777.

Neumann P, Novák P, Hoštáková N, Macas J. 2019. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA* 10:1.

Neumann P, Pavlíková Z, Koblížková A, Fuková I, Jedličková V, Novák P, Macas J. 2015. Centromeres off the hook: Massive changes in centromere size and structure following duplication of CenH3 gene in Fabeae species. *Mol. Biol. Evol.* 32:1862–1879.

Neumann P, Pozárková D, Vrána J, Dolezel J, Macas J. 2002. Chromosome sorting and PCR-based physical mapping in pea (*Pisum sativum* L.). *Chromosome Res.* 10:63–71.

Neumann P, Schubert V, Fuková I, Manning JE, Houben A, Macas J. 2016. Epigenetic histone marks of extended meta-polycentric centromeres of *Lathyrus* and *Pisum* chromosomes. *Front. Plant Sci.* 7:234.

Ni P, Huang N, Nie F, Zhang J, Zhang Z, Wu B, Bai L, Liu W, Xiao C Le, Luo F, et al. 2021. Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. *Nat. Commun.* 12:1–11.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* 376:44–53.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 30:1291–1305.

Oliveira L, Neumann P, Jang T, Klemme S, Schubert V. 2020. Mitotic spindle attachment to the holocentric chromosomes of *Cuscuta europaea* does not correlate with the distribution of CENH3 chromatin. *Front. Plant Sci.* 10:1799.

Peona V, Weissensteiner MH, Suh A. 2018. How complete are 'complete' genome assemblies? - An avian perspective. *Mol. Ecol. Resour.* 18:1188–1195.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44:W160–W165.

Rengs WMJ, Schmidt MH -W., Effgen S, Le DB, Wang Y, Zaidan MWAM, Huettel B, Schouten HJ, Usadel B, Underwood CJ. 2022. A chromosome scale tomato genome built from complementary PacBio and Nanopore sequences alone reveals extensive linkage drag during breeding. *Plant J.* 110:572-558.

Schubert I. 2018. What is behind "centromere repositioning"? *Chromosoma* 127:229–234.

Schubert V, Neumann P, Marques A, Heckmann S, Macas J, Pedrosa-Harand A, Schubert I, Jang T-S, Houben A. 2020. Super-resolution microscopy reveals diversity of plant centromere architecture. *Int. J. Mol. Sci.* 21:3488.

Sobreira TJP, Durham AM, Gruber A. 2006. TRAP: automated classification, quantification and annotation of tandemly repeated sequences. *Bioinformatics* 22:361–362.

Song J-M, Xie W-Z, Wang S, Guo Y-X, Koo D-H, Kudrna D, Gong C, Huang Y, Feng J-W, Zhang W, et al. 2021. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* 14:1757–1767.

Song L, Sabunciyan S, Yang G, Florea L. 2019. A multi-sample approach increases the accuracy of transcript assembly. *Nat. Commun.* 10:5000.

Stovner EB, Sætrom P. 2019. Epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics* 35:4392–4393.

Talbert PB, Henikoff S. 2020. What makes a centromere? *Exp. Cell Res.* 389:111895.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31:2032–2034.

Tayeh N, Aluome C, Falque M, Jacquin F, Klein A, Chauveau A, Bérard A, Houtin H, Rond C, Kreplak J, et al. 2015. Development of two major resources for pea genomics: the GenoPea 13.2K SNP Array and a high density, high resolution consensus genetic map. *Plant J.* 84:1257–1273.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27:737–746.

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. 2018. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* 7:giy093.

Vondrak T, Ávila Robledillo L, Novák P, Koblížková A, Neumann P, Macas J. 2020. Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *Plant J.* 101:484–500.

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D, et al. 2020. Multiple wheat genomes reveal global variation in modern breeding. *Nature* 588:277–283.

Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, et al. 2022. High-quality *Arabidopsis thaliana* genome assembly with Nanopore and HiFi long reads. *Genomics. Proteomics Bioinformatics* 20:4–13.

Xue C, Liu G, Sun S, Liu X, Guo R, Cheng Z, Yu H, Gu M, Liu K, Zhou Y, et al. 2022. De novo centromere formation in pericentromeric region of rice chromosome 8. *Plant J.* 111:859–871.

Yang T, Liu R, Luo Y, Hu S, Wang D, Wang C, Pandey MK, Ge S, Xu Q, Li N, et al. 2022. Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nat. Genet.* 54:1553–1563.

Zhang T, Liu G, Zhao H, Braz GT, Jiang J. 2021. Chorus2: design of genome-scale oligonucleotide-based probes for fluorescence in situ hybridization. *Plant Biotechnol. J.* 19:1967-1978.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137.

Zhong X-B, de Jong JH, Zabel P. 1996. Preparation of tomato meiotic pachytene and mitotic metaphase chromosomes suitable for fluorescencein situ hybridization (FISH). *Chromosome Res.* 4:24–28.

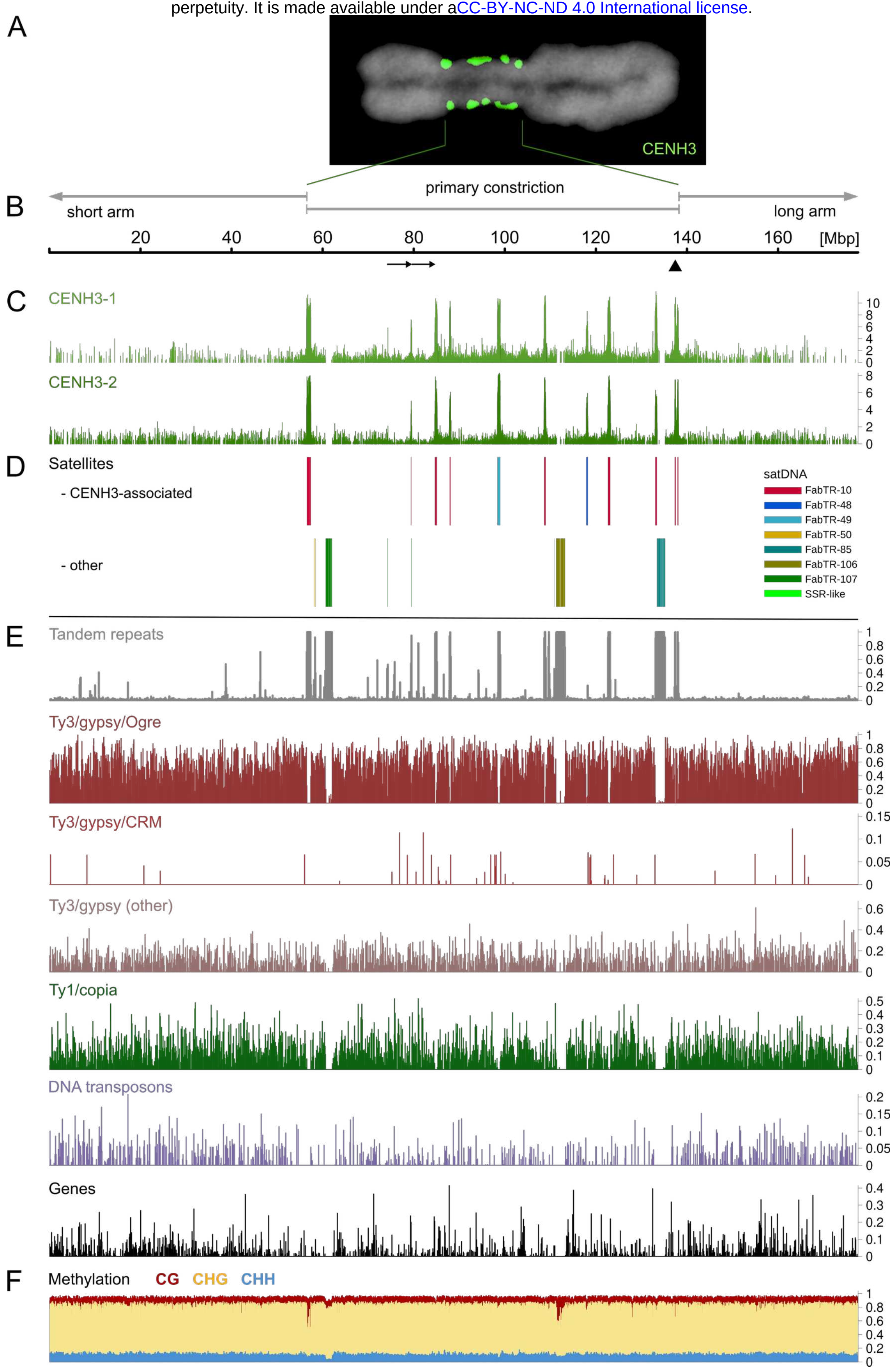

Fig. 1. Features of pea centromere 6 (CEN6). (A) Immunolabeling of CENH3 protein (green) on metaphase chromosome 6 (counterstained with DAPI, gray). **(B)** Position of the primary constriction in the assembly. Arrows below the scale indicate the 5.2 Mb tandem duplication, and the arrowhead shows the position of a single gap in the assembly. **(C)** Distribution of CENH3 chromatin revealed by ChIP-seq experiments using anti-CENH3-1 and anti-CENH3-2 antibodies. Peaks in the graphs correspond to the statistically significant enrichment ratio of ChIP reads to control input reads (see *SI Appendix*, Fig. S2 for full data analysis). **(D)** Positions of large arrays of satellite repeats. Different repeat families are marked by different colors, as indicated in the legend. **(E)** Densities of different types of repetitive DNA sequences and predicted genes calculated in 100 kb windows. **(F)** Cytosine methylation profiles calculated as the ratio of methylated cytosines to all cytosines present in the sequence. Ratios were calculated separately for cytosines in three different contexts (distinguished by plot colors) and averaged for 100 kb windows.

Figure 2

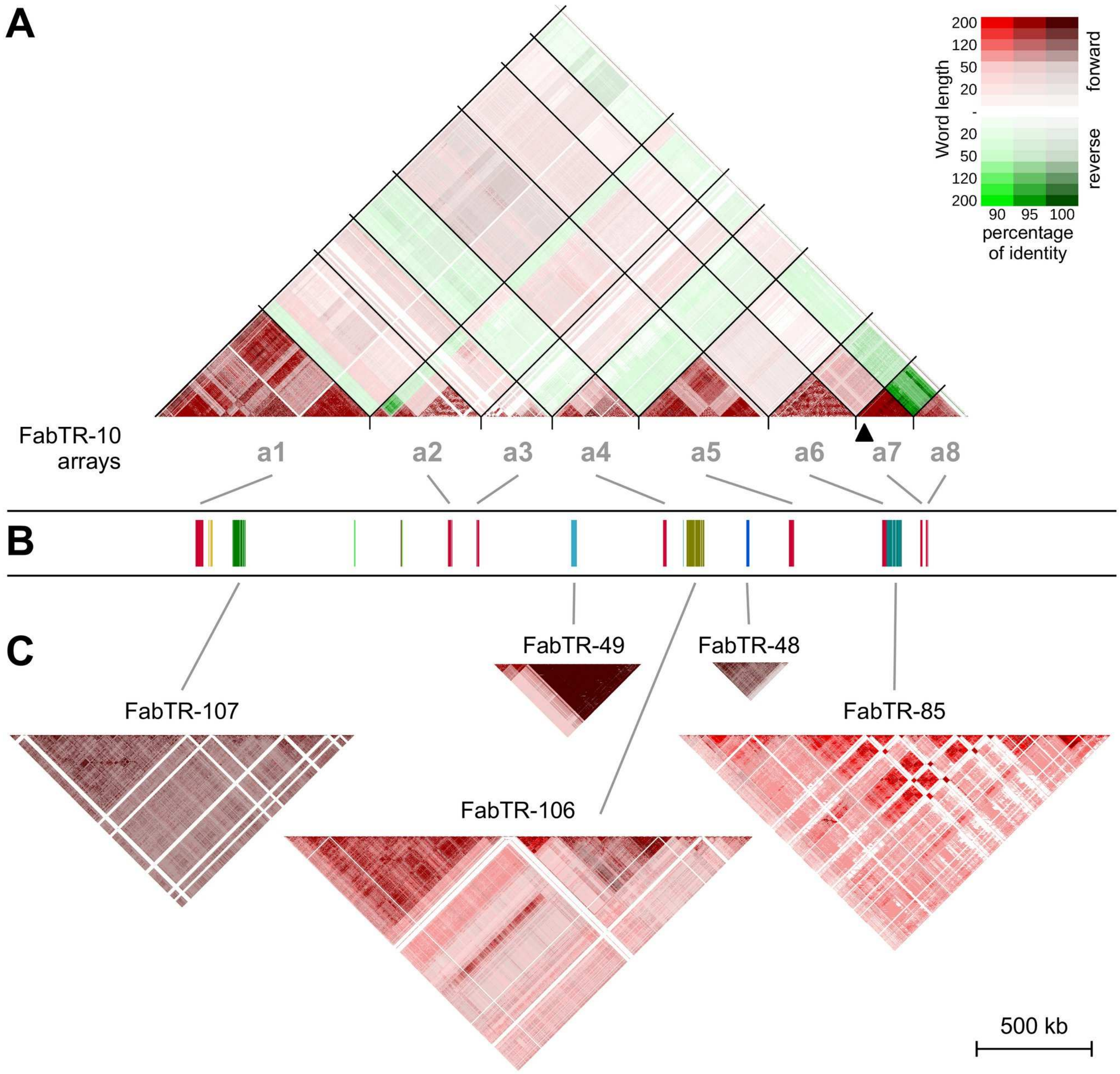

Fig. 2. Sequence homogenization patterns of satellite DNA arrays. Nucleotide sequence similarities were visualized as similarity dot plots of k-mers of different sizes (10–200 nt). The percent identity and mutual orientation of the compared sequences are indicated by the colors shown in the legend. **(A)** Dot-plot of FabTR-10 repeats showing comparison of sequences both within and between arrays located in eight different loci (a1–a8) in CEN6. **(B)** The schematic representation of the array positions in CEN6 (corresponds to Fig. 1D). **(C)** Dot plots of the satellites present in CEN6 as single arrays. Sequence comparisons were performed only within individual arrays for these satellites. All dot plots were calculated identically and drawn to scale to account for differences in sequence homogenization and array lengths. Black arrowhead under the FabTR-10 a7 array shows the position of the gap of unknown length in the assembly.

Figure 3

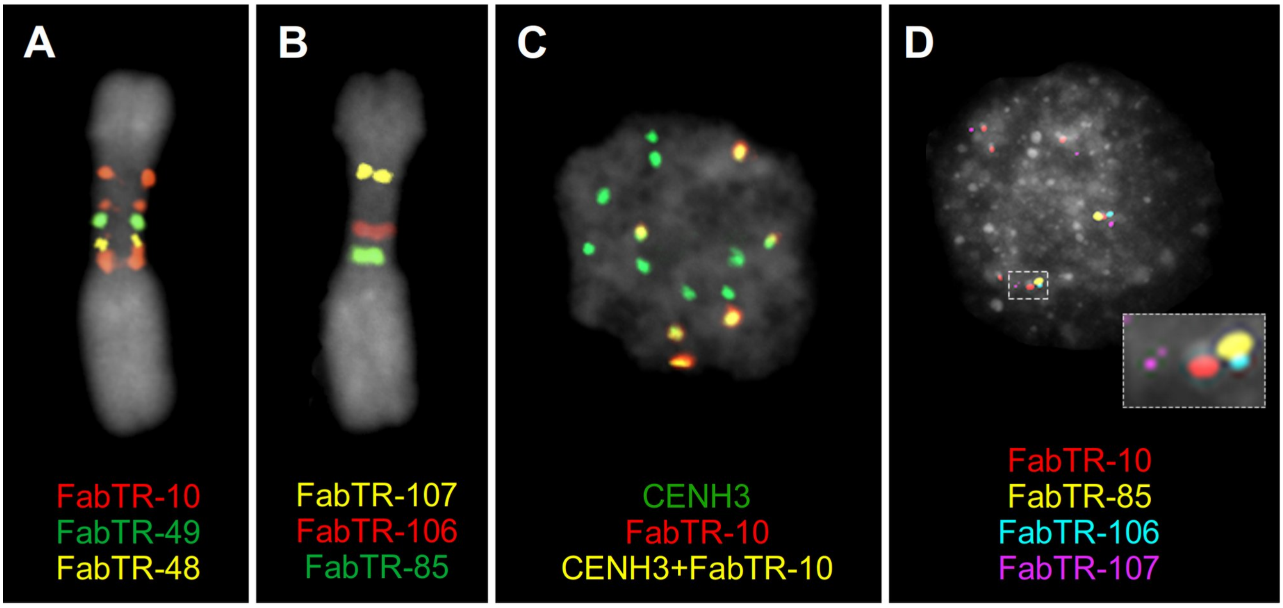

Fig. 3. Association of repeats with CENH3 determines their position on chromosomes and condensation patterns in interphase nuclei. (A-B) Multicolor FISH detection of satellite repeats on metaphase chromosome 6. CENH3-associated satellite repeats are located along the periphery of the primary constriction **(A)**, whereas CENH3-free satellites are embedded within the constriction **(B)**. **(C-D)** Immuno-FISH detection of CENH3 protein and satellite repeats in interphase nuclei. **(C)** All CENH3 loci from each chromosome are condensed into a single spot, along with their associated satellites such as FabTR-10, resulting in 14 CENH3 signals per nucleus (2n = 14). Note that only a part of chromosomes contain FabTR-10. **(D)** CENH3-free satellites are located away from the condensed CENH3 domains of CEN6. The position of CENH3 chromatin is indicated with the FabTR-10 probe. Satellite repeats and CENH3 protein are labeled with different colors as indicated in the figures. Chromosomes and nuclei counterstained with DAPI are shown in gray.

Figure 4

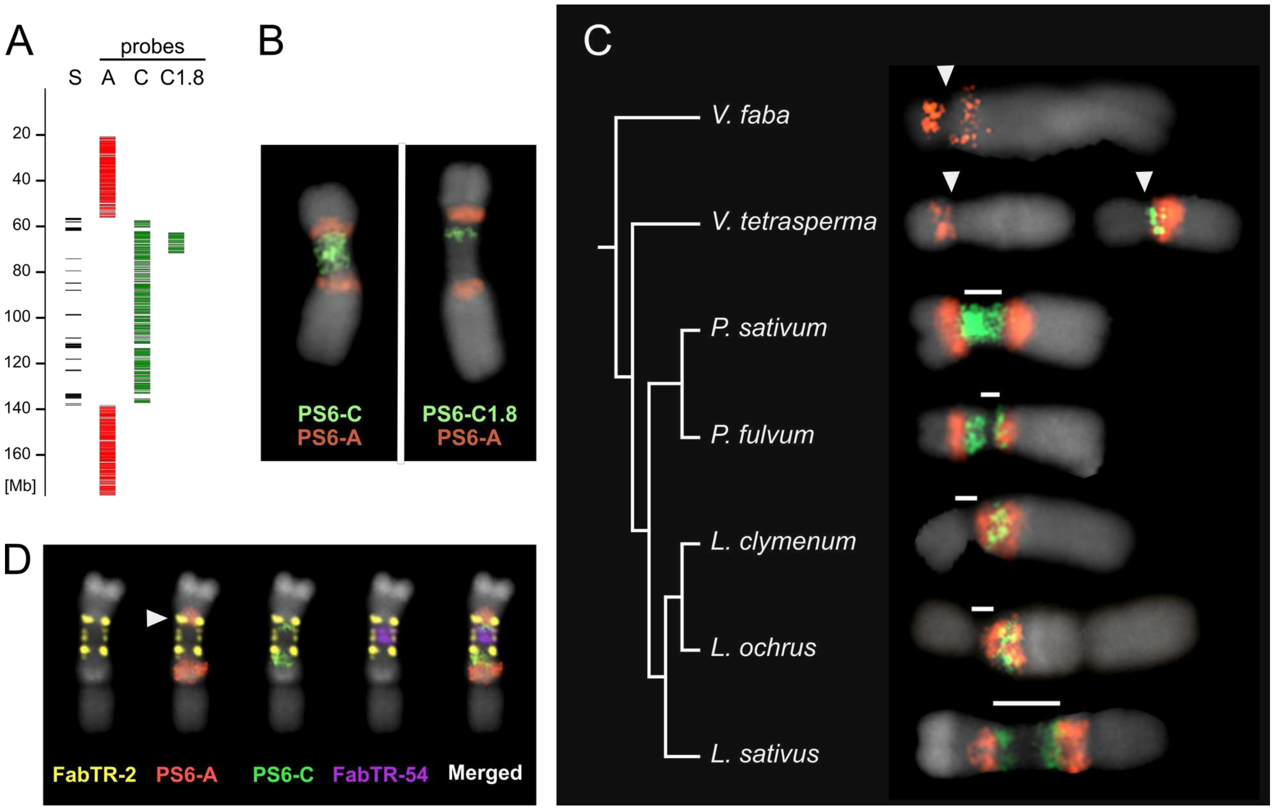

Fig. 4. CEN6 painting probes and their application for the detection of orthologous regions in related species. (A) Positions in the assembly of oligonucleotide sequences used as FISH painting probes. Each column represents different PS6 probes. Column "S" shows the positions of satDNA arrays marking the extent of primary constriction. **(B)** Painting probes applied to *P. sativum* chromosome 6. **(C)** FISH analysis of a set of related *Fabeae* species using PS6-C (green) and PS6-A (red) probes. The phylogenetic tree was adapted from (Ávila Robledillo et al. 2020). Only chromosome(s) that produced hybridization signals are shown. Primary constrictions are marked with white arrowheads (monocentric) or bars (metapolycentric chromosomes). Images of whole chromosome complements can be found in *SI Appendix*, Fig. S5A. **(D)** Multicolor FISH labeling of the *Lathyrus sativus* homeolog of pea chromosome 6 using PS6 painting probes as well as probes for satellite repeats FabTR-54, which fills the gap in the PS6-C signal, and FabTR-2, which is associated with CENH3 chromatin in *L. sativus* (Ávila Robledillo et al. 2020). Arrowhead indicates the overlap of PS6-A and FabTR-2 signals.

Supporting Information Appendix

Fig. S1

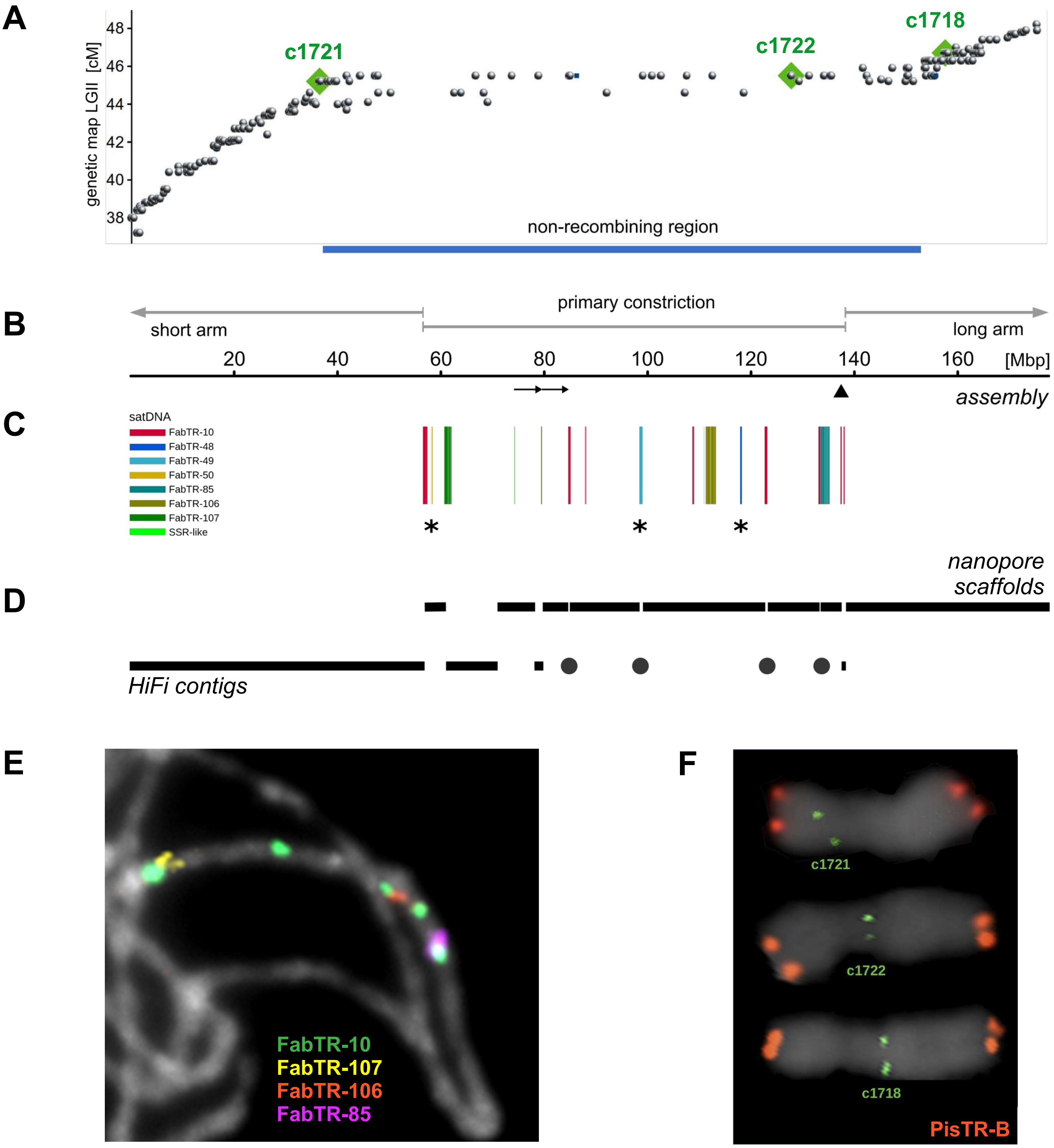

Fig. S1. Assembly construction and verification using genetically and physically localized markers. The nanopore "seed" reads used to initiate CEN6 scaffolding were selected based on the presence of sequences of genetic markers from the nonrecombining region of linkage group LGII or the sequences of CEN6-specific satellite repeats. (**A**) The positions of genetic marker sequences in the assembly (x-axis) compared with their positions on the genetic map. Markers highlighted in green were physically localized on chromosomes (panel F). (**B**) The position of the primary constriction in the assembly. Arrows below the scale indicate the 5.2 Mb tandem duplication, and the arrowhead indicates the position of a single gap in the assembly. (**C**) Positions of the satDNA arrays, with the three CEN6-specific families marked with asterisks. (**D**) Regions of the assembly that were scaffolded with nanopore reads or constructed from HiFi contigs are shown by horizontal bars. Dots mark gaps in nanopore scaffolds corresponding to satDNA arrays that were filled using HiFi contigs. (**E-F**) Examples of assembly verification using FISH. (**E**) Localization of selected satellite repeats on pachytene chromosomes. Note that smaller FabTR-10 signals are not visible due to the short exposure time. (**F**) Sequences of genetic markers (green) detected on metaphase chromosome 6 using the highly-sensitive single-copy FISH protocol. Satellite PisTR-B (red) was used to discriminate chromosomes within the pea karyotype.

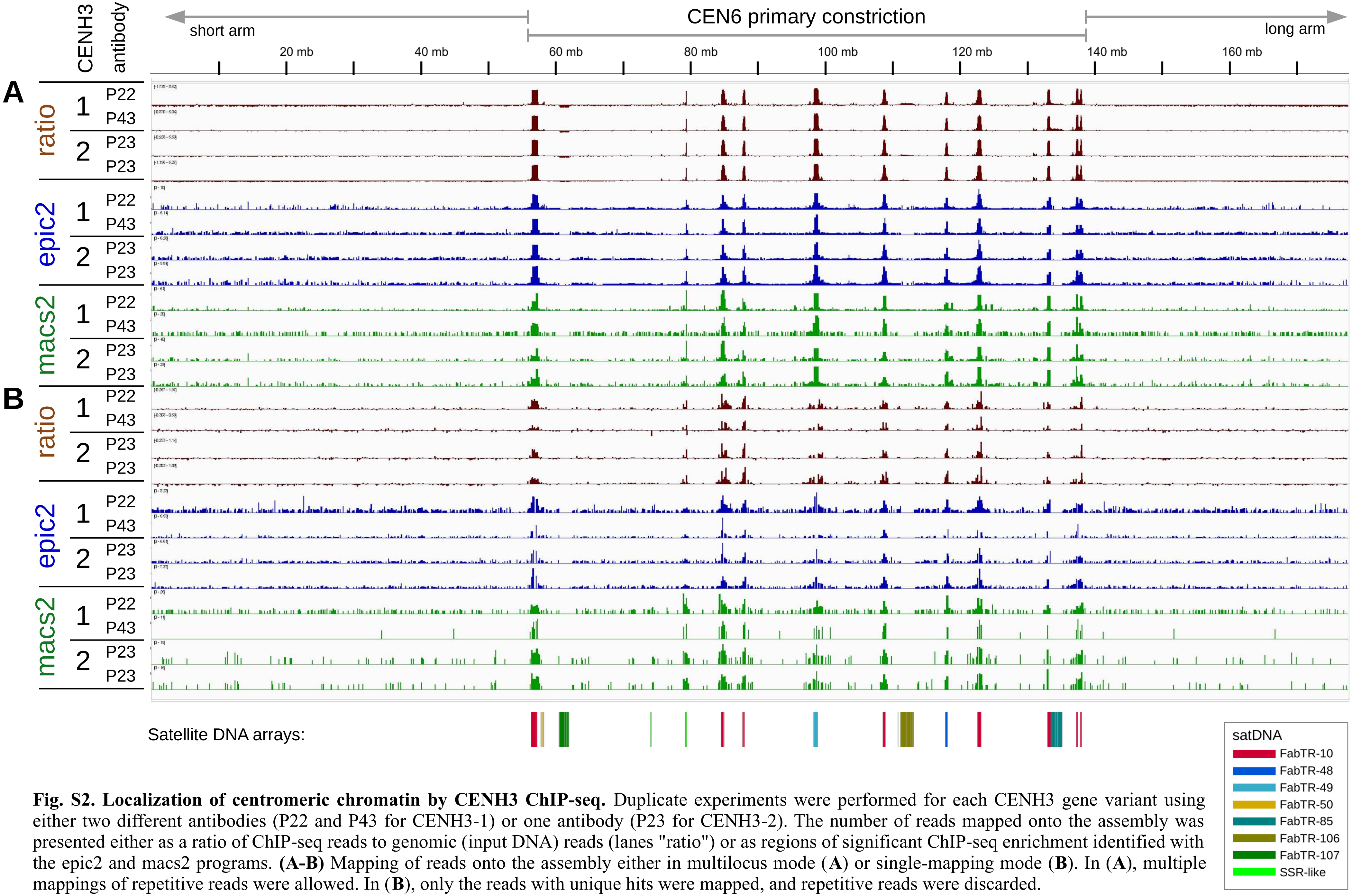

Fig. S2. Localization of centromeric chromatin by CENH3 ChIP-seq. Duplicate experiments were performed for each CENH3 gene variant using either two different antibodies (P22 and P43 for CENH3-1) or one antibody (P23 for CENH3-2). The number of reads mapped onto the assembly was presented either as a ratio of ChIP-seq reads to genomic (input DNA) reads (lanes "ratio") or as regions of significant ChIP-seq enrichment identified with the epic2 and macs2 programs. **(A-B)** Mapping of reads onto the assembly either in multilocus mode (**A**) or single-mapping mode (**B**). In (**A**), multiple mappings of repetitive reads were allowed. In (**B**), only the reads with unique hits were mapped, and repetitive reads were discarded.

Fig. S3 A

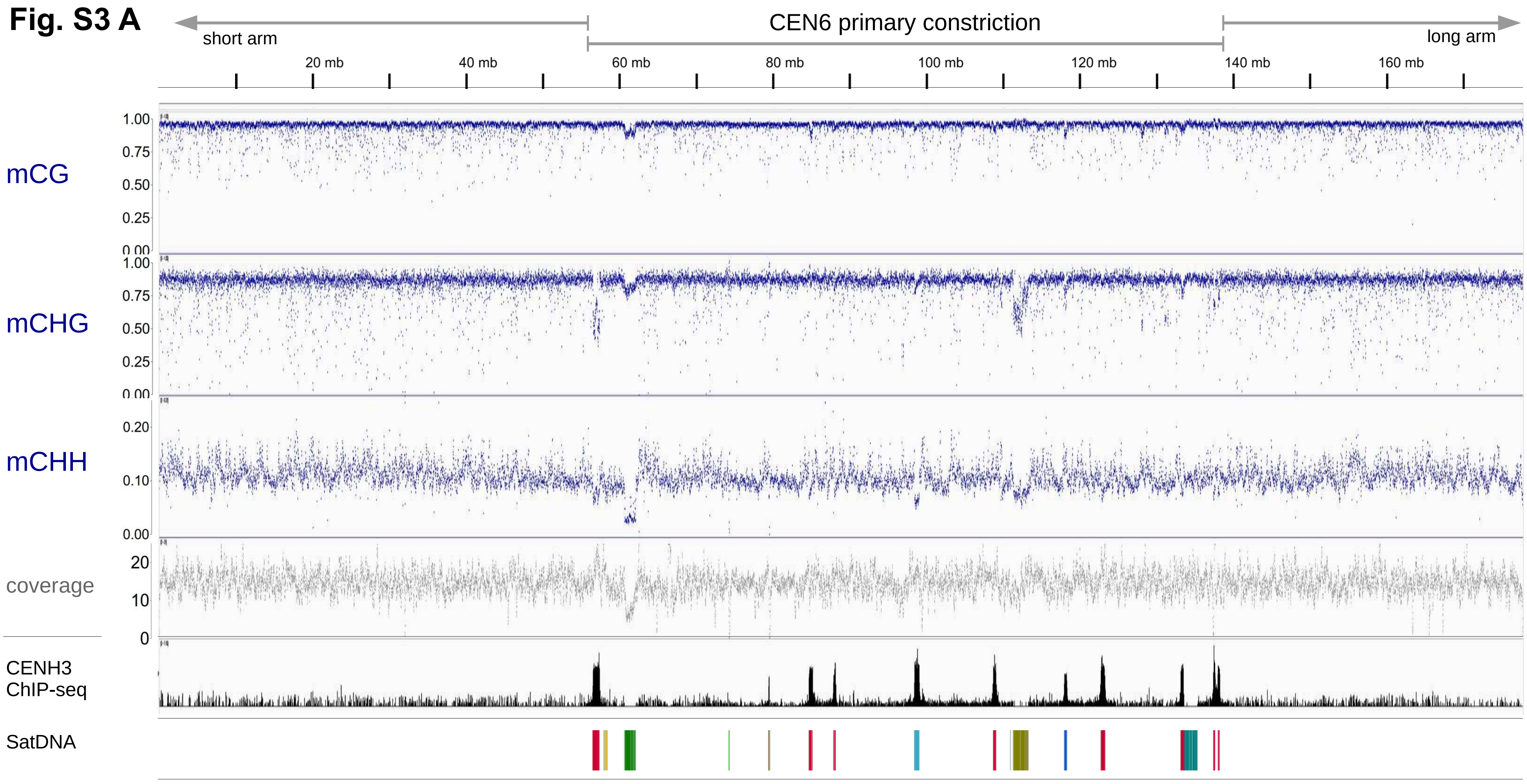

Fig. S3. DNA methylation profile of CEN6. Per-base cytosine methylation frequencies in three sequence contexts known in plants (CpG, CHG, CHH) were obtained by analyzing Oxford Nanopore reads aligned to the assembly using DeepSignal-plant (Ni et al., 2021). **(A)** The plots show the fraction of aligned nanopore reads, in which cytosine was methylated at a given position. The total number of aligned nanopore reads is indicated in the "coverage" plot. The distribution of CENH3 chromatin and annotations of the major families of satDNA are shown for comparison with the methylation profiles.

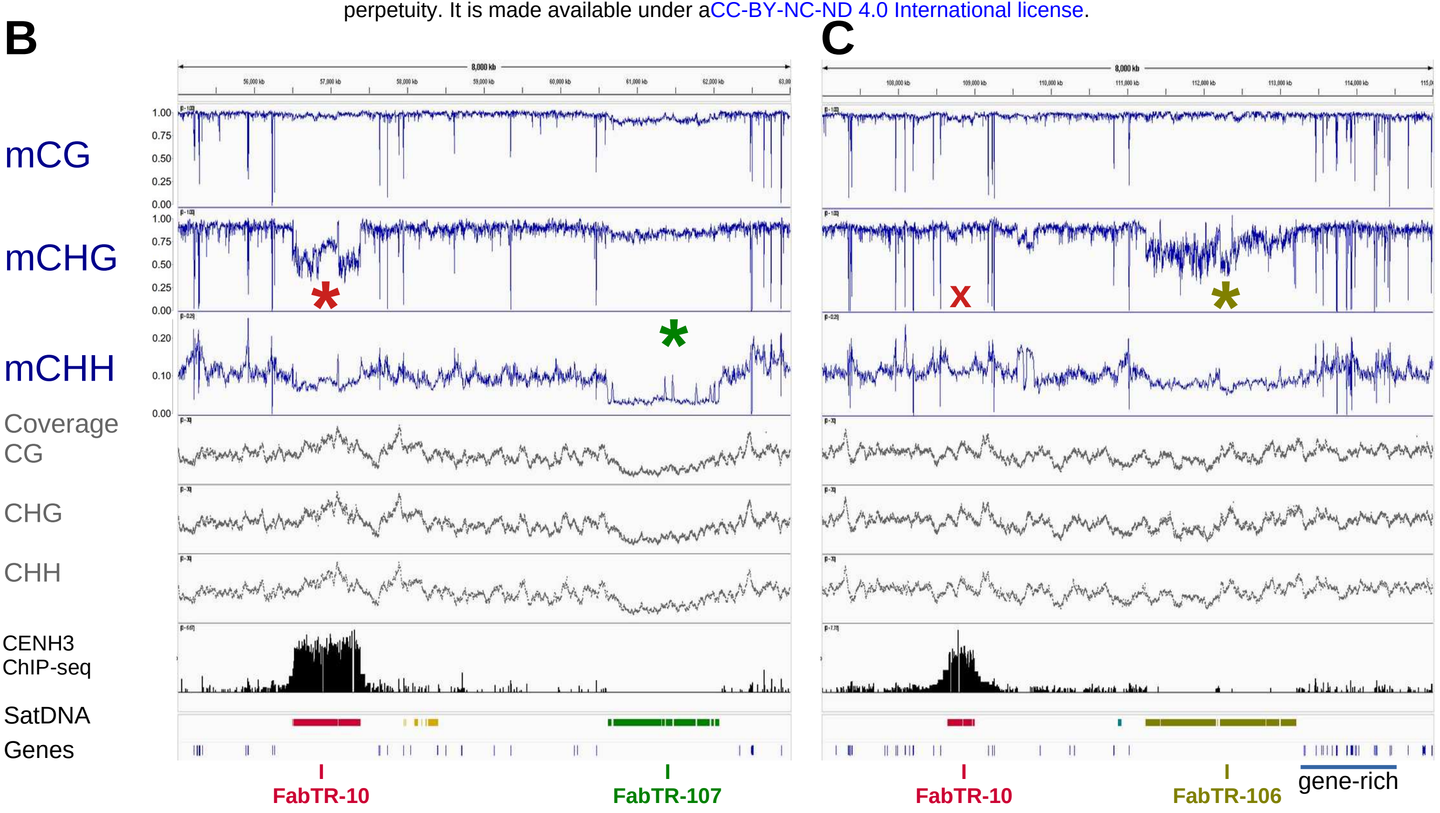

Fig. S3 B,C. Detailed examples of hypomethylated regions. Hypomethylated arrays of satDNA are marked with asterisks. (**B**) Sequence at the short-arm constriction junction contains CHG-hypomethylated FabTR-10, whereas the array of the same repeat within the constriction has a normal methylation level (**C**, marked with "x"). Short hypomethylated islands are best seen in the gene-rich region marked in (C).

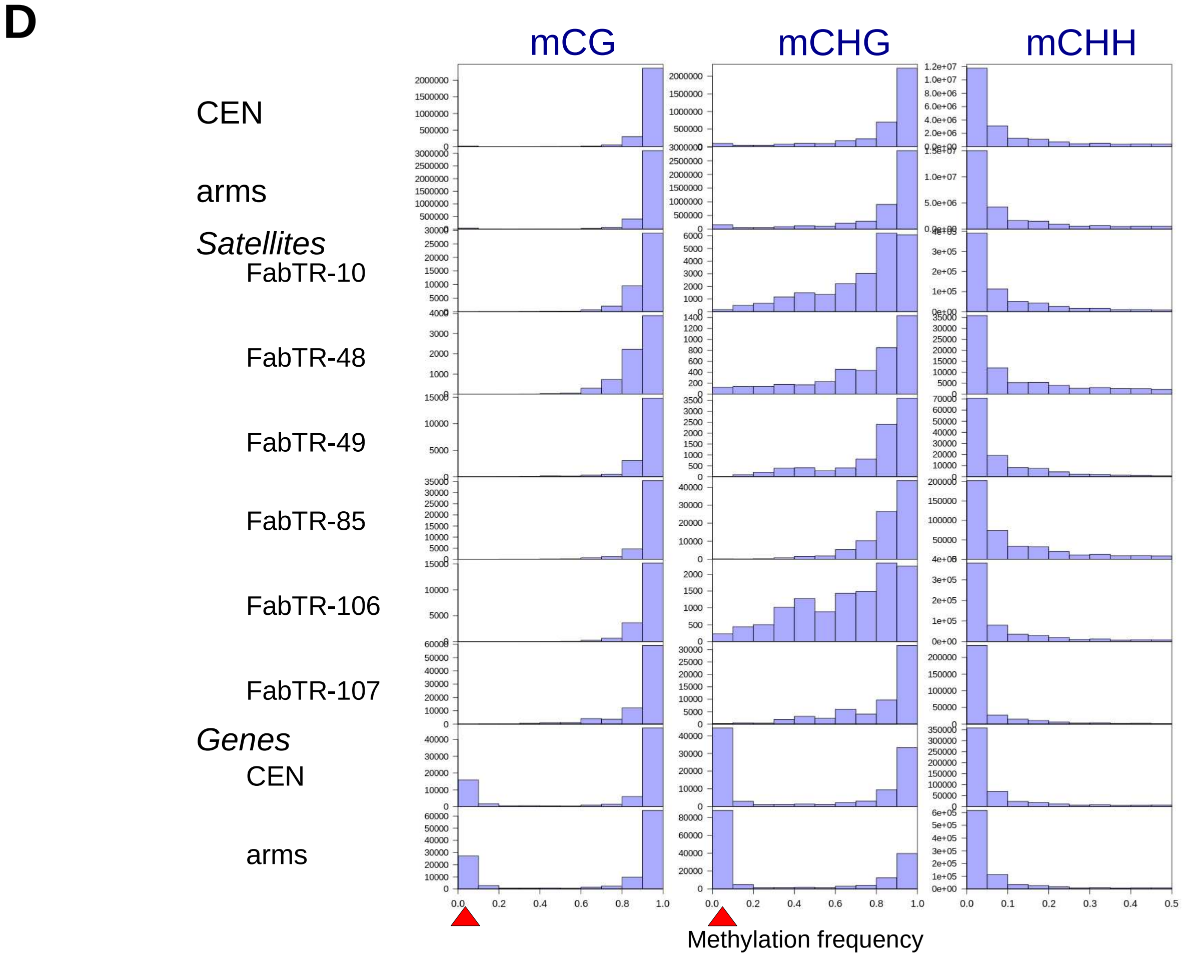

Fig. S3D. Per-base methylation frequency distributions within specific regions or sequence types. Distributions were calculated for the entire primary constriction ("CEN") and chromosome arm ("arms") sequences as well as for specific satellite repeats and genes. Gene sequences occurring in the centromere (CEN) and chromosome arms were analyzed separately. Red arrowheads mark the position of peaks corresponding to hypomethylated genes.

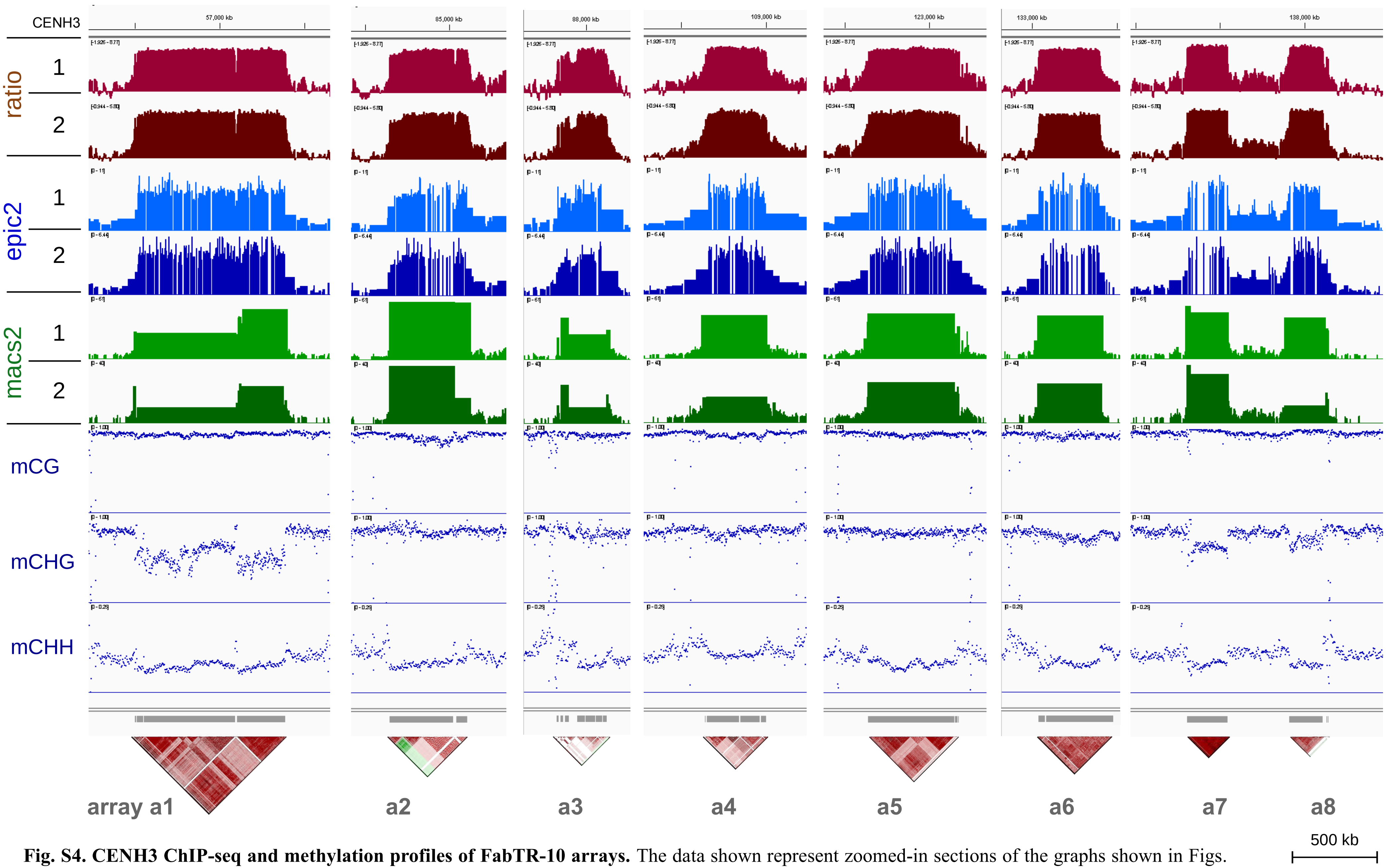

Fig. S4. CENH3 ChIP-seq and methylation profiles of FabTR-10 arrays. The data shown represent zoomed-in sections of the graphs shown in Figs. S2 and S3 corresponding to loci with FabTR-10 arrays. The positions of the arrays are indicated by gray bars below the graphs and are complemented by sequence homogenization dot plots (compiled from Fig. 1).

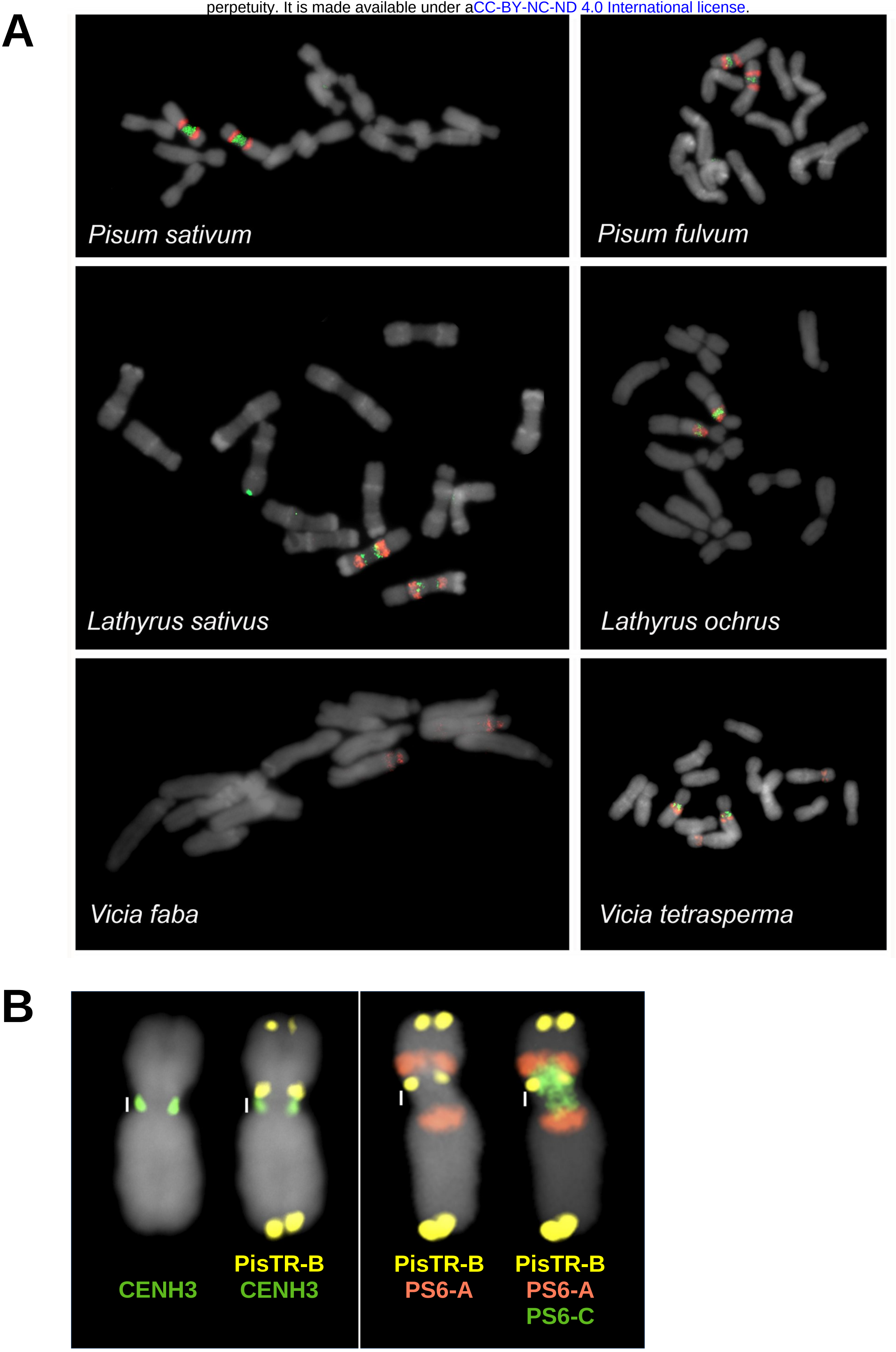

Fig. S5. FISH with CEN6 painting probes. (**A**) Chromosome complements of selected Fabeae species hybridized with PS6-C (green) and PS6-A (red) painting probes. **(B)** Hybridization pattern of CEN6 painting probes on chromosome 6 of *Pisum fulvum*. *Left panel:* extent of the primary constriction (white bar), as revealed by the immunolabeling of CENH3 and the FISH detection of PisTR-B repeats, showing that PisTR-B is located just above the CENH3 signals. *Right panel:* combined FISH detection using the painting probes together with the PisTR-B probe, which was used as a reference for the end of the constriction and shows that the green PS6-C probe extends into the short arm.