The genetic and epigenetic landscape of the Arabidopsis centromeres

Short title: Assembly of the Arabidopsis centromeres.

One-sentence summary: Long-read sequencing and assembly of the Arabidopsis centromeres reveals their genetic and epigenetic topography.

Authors:

Matthew Naish[1]*, Michael Alonge[2]*, Piotr Wlodzimierz[1]*, Andrew J. Tock[1], Bradley W. Abramson[3], Anna Schmücker[4], Terezie Mandáková[5], Bhagyshree Jamge[4], Christophe Lambing[1], Pallas Kuo[1], Natasha Yelina[1], Nolan Hartwick[3], Kelly Colt[3], Lisa Smith[6], Jurriaan Ton[6], Tetsuji Kakutani[7], Robert A. Martienssen[8], Korbinian Schneeberger[9,10], Martin A. Lysak[5], Frédéric Berger[4], Alexandros Bousios[11], Todd P. Michael[3], Michael C. Schatz[2†] and Ian R. Henderson[1†]

Affiliations:

[1] Department of Plant Sciences, Downing Street, University of Cambridge, Cambridge, CB2 3EA, United Kingdom.

[2] Department of Computer Science, Johns Hopkins University, Baltimore, USA.

[3] The Plant Molecular and Cellular Biology Laboratory, Salk Institute for Biological Studies, La Jolla, California, USA.

[4] Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria.

[5] Central European Institute of Technology (CEITEC), Masaryk University, Kamenice 5, Brno 625 00, Czech Republic.

[6] School of Biosciences and Institute for Sustainable Food, University of Sheffield, Sheffield, S10 2TN, United Kingdom.

[7] Department of Biological Sciences, University of Tokyo, Tokyo, Japan.

[8] Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.

[9] Faculty of Biology, LMU Munich, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany.

[10] Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne.

[11] School of Life Sciences, University of Sussex, United Kingdom, UK.

* Equal contribution.

† **Correspondence:** mschatz@cs.jhu.edu and irh25@cam.ac.uk

Abstract:

Centromeres attach chromosomes to spindle microtubules during cell division and, despite this conserved role, show paradoxically rapid evolution and are typified by complex repeats. We used long-read sequencing to generate the Col-CEN *Arabidopsis thaliana* genome assembly that resolves all five centromeres. The centromeres consist of megabase-scale tandemly repeated satellite arrays, which support CENH3 occupancy and are densely DNA methylated, with satellite variants private to each chromosome. CENH3 preferentially occupies satellites that show least divergence and occur in higher-order repeats. The centromeres are invaded by *ATHILA* retrotransposons, which disrupt genetic and epigenetic organization. Centromeric crossover recombination is suppressed, yet low levels of meiotic DSBs occur that are regulated by DNA methylation. We propose that Arabidopsis centromeres are evolving via cycles of satellite homogenization and retrotransposon-driven diversification.

Introduction:

Despite their conserved function during chromosome segregation, centromeres show diverse organization between species, ranging from single nucleosomes to megabase-scale tandem repeat arrays (*1*). Centromere 'satellite' repeat monomers are commonly ~100–200 bp, with each repeat capable of hosting a CENPA/CENH3-variant nucleosome (*1*, *2*). CENPA/CENH3 nucleosomes ultimately assemble the kinetochore and position spindle attachment on the chromosome, allowing segregation during cell division (*3*). Satellites are highly variable in sequence composition and length when compared between species (*2*). The library of centromere repeats present within a genome often shows concerted evolution, yet they have the capacity to change rapidly in structure and sequence within and between species (*1*, *2*, *4*). However, the genetic and epigenetic features that contribute to centromere evolution are incompletely understood, in large part due to the challenges of centromere sequence assembly and functional genomics of highly repetitive sequences.

Genomic repeats, especially long or high-similarity repeats, are notoriously difficult to assemble from fragmented sequencing reads (*5*). As sequencing reads have become longer and more accurate, eukaryotic *de novo* genome assemblies have captured an increasingly complete picture of repetitive elements. Oxford Nanopore Technologies (ONT) long reads have become substantially longer and more accurate (>100 kbp with 95–99% modal accuracy), owing to improved DNA extraction and library preparation, together with advanced machine learning-based basecalling. Additionally, PacBio High-Fidelity (HiFi) reads, while shorter (~15 kbp), are highly accurate (>99%). Using these technologies with new computational methods, researchers have assembled a complete telomere-to-telomere representation of the human genome, including the centromere satellite arrays (*6–8*). This work revealed that ONT and HiFi reads are sufficient to span interspersed unique marker sequences in human centromeres and other complex repeats, suggesting that truly complete genome assemblies for diverse eukaryotes are on the horizon.

Arabidopsis thaliana is a major model plant species and its genome was sequenced in 2000, yet the centromeres, telomeres, and ribosomal DNA repeats have remained unassembled, due to their high

repetition and similarity (*9*). The Arabidopsis centromeres contain millions of base pairs of the *CEN180* satellite, which support CENH3 loading (*10–14*). We used long-read ONT sequencing, followed by polishing with high-accuracy PacBio HiFi reads, to establish the Col-CEN reference assembly, which wholly resolves all five Arabidopsis centromeres. The assembly contains a library of 66,131 *CEN180* satellites, with each chromosome possessing mostly private satellite variants. Chromosome-specific higher-order *CEN180* repetition is prevalent within the centromeres. We identified *ATHILA* retrotransposons that have invaded the satellite arrays and interrupt the genetic and epigenetic organization of the centromeres. By analyzing SPO11-1-oligo data from mutant lines, we demonstrate that DNA methylation epigenetically silences initiation of meiotic DNA double-strand breaks (DSBs) within the centromeres. Our data suggest that satellite homogenization and retrotransposon invasion are driving cycles of centromere evolution in Arabidopsis.

Complete assembly of the Arabidopsis centromeres

We collected Col-0 genomic ONT and HiFi sequencing data comprising a total of 73.6 Gbp (~56× >50 kbp) and 14.6 Gbp (111.3×, 15.6 kbp mean read length), respectively. These data yielded an improved assembly of the Col-0 genome (Col-CEN v1.2), where chromosomes 1, 3 and 5 are wholly resolved from telomere-to-telomere, and chromosomes 2 and 4 are complete apart from the short-arm *45S* rDNA clusters and adjacent telomeres (**Fig. 1**). After telomere patching and repeat-aware polishing with ONT, HiFi and Illumina reads (*15*), the Col-CEN assembly has a quality value (QV) of 45.99 and 51.71 inside and outside of the centromeres, equivalent to approximately one error per 40,000 and 148,000 bases, respectively (**Fig. S1–S2A**, **Table S1**). Additionally, Hi-C and Bionano optical maps validate the large-scale structural accuracy of the assembly (**Fig. S2**). The Col-CEN assembly is highly concordant with TAIR10, showing no large structural differences within the chromosome arms (**Fig. 1B**). 97.5% of Col-0 BAC contigs align to both TAIR10 and Col-CEN with high coverage and identity (>95%), and 99.9% of TAIR10 gene annotations are represented in Col-CEN.

Col-CEN reconstructs all five centromeres spanning 12.6 Mbp of new sequence, 120.0 and 97.6 kbp of *45S* rDNA in the chromosome 2 and 4 nucleolar organizer regions (NORs), and the complete telomeres of the 8 chromosome arms without sub-telomeric NORs (**Fig. 1A–1C, S1–S3**). We found several instances of apparently genuine variation between the Col-0 strains used to generate TAIR10 and Col-CEN (**Fig. S4, Tables S2–S3**). For example, a thionin gene cluster shows a deletion in Col-CEN relative to TAIR10 (**Fig. S4**). In total, 27 TAIR10 genes are missing from Col-CEN due to presence/absence variation, and 13 are present in multiple copies (**Tables S2–S3**). To comprehensively account for variation between Col-0 strains, we aligned ONT, HiFi, and Illumina reads to the Col-CEN assembly and called variants, providing a database of potential allelic differences, including heterozygous variants (https://github.com/schatzlab/Col-CEN). This revealed only 41 and 37 structural variant calls from ONT and HiFi data genome-wide, consistent with very low heterozygosity.

We confirmed chromosome landmarks flanking centromere 1 using fluorescence *in situ* hybridization (FISH), which included labelling a telomeric-repeat cluster located adjacent to the centromere (**Fig. 1D, S5**). To validate centromere structure, we performed *in silico* digestion with *Asc*I and *Not*I and compared the predicted fragments to published physical maps, which validated Col-CEN (**Fig. S6**) (*16*). We also examined our Bionano optical data across the centromeres (**Fig. S7**). The optical contigs are consistent with the structure of Col-CEN *CEN180* arrays, although the low density of centromeric labeling sites prevents full resolution by optical fragments alone (**Fig. S7**).

The centromeres are characterized by a repeated 178-bp satellite repeat (*CEN180*), arranged head-to-tail and organized into higher-order repeats (**Fig. 1D, 2, S8**). We validated the structural and base-level accuracy of the centromeres using techniques from the Human T2T consortium (*6*, *8*), and observed even long-read coverage across the centromeres with few loci showing plausible alternate base signals (**Fig. S1B**). We observed relatively few 'missing' *k*-mers that are found in the assembly but not in Illumina short reads, which are diagnostic of residual consensus errors that remain after polishing (**Fig. S1B**) (*17*). We observed that unique 'marker' sequences are frequent, with a maximum distance between consecutive markers of 41,765 bp within the centromeres, suggesting that our reads can

confidently span these markers and assemble reliably (**Fig. S1C**). The five centromeres are relatively distinct at the sequence level, with each exhibiting chromosome-specific repeats (**Fig. 1E, 2, Tables S4–S5**). Using the Col-CEN sequence, we designed *CEN180* variant FISH probes to label specific centromere arrays (**Fig. 1F**, **S5**). For example, the *CEN180-α*, *CEN180-γ* and *CEN180-δ* probes specifically label arrays within centromere 1 (**Fig. 1F**, **S5**), providing cytogenetic validation for chromosome-specific satellites.

The Arabidopsis *CEN180* satellite repeat library

We performed *de novo* searches for tandem repeats to define the centromere satellite library (**Table S4**). We identified 66,131 *CEN180* satellites in total, with between 11,848–15,613 copies per chromosome (**Fig. 2, S9, Table S4**). The *CEN180* repeats form large tandem arrays, with the satellites within each centromere found predominantly on the same strand, except for centromere 3, which is formed of two blocks on opposite strands (**Fig. 1D, S8**). The distribution of repeat monomer length is constrained around 178 bp (**Fig. 2A, S9**). We aligned all *CEN180* sequences to derive a genome-wide consensus and calculated nucleotide frequencies at each alignment position to generate a position probability matrix (PPM). Each satellite was compared to the PPM to calculate a 'variant distance' by summation of disagreeing nucleotide probabilities. Substantial sequence variation was observed between satellites and the PPM, with a mean variant distance of 20.2 (**Fig. 2A**). Each centromere contains essentially private libraries of *CEN180* monomers, with only 0.3% sharing an identical copy on a different chromosome (**Fig. 1E, Table S4**). In contrast, there is a high degree of *CEN180* repetition within chromosomes, with 57.1–69.0% showing one or more duplicates (**Table S4**). We also observed a minor class of *CEN160* repeats found on chromosome 1 (1,289 repeats, mean length=158.2 bp) (*14*).

We aligned CENH3 ChIP-seq data to the Col-CEN assembly and observed on average 12.9-fold $\log_2$(ChIP/input) enrichment within the *CEN180* arrays, compared to the chromosome arms (**Fig. 1D, S8**) (*10*). CENH3 ChIP-seq enrichment is generally highest within the interior of the main *CEN180* arrays (**Fig. 1D, S8**). We observed a negative relationship between CENH3 ChIP-seq enrichment and

CEN180 variant distance (**Fig. 2D–2E**), consistent with CENH3 nucleosomes preferring to occupy satellites that are closer to the genome-wide consensus. In this respect, centromere 4 is noteworthy, as it consists of two distinct *CEN180* arrays, with the right array showing higher variant distances and lower CENH3 enrichment (**Fig. 1D, 2D, S8**). Together, this is consistent with satellite divergence leading to loss of CENH3 binding, or vice versa.

To define *CEN180* higher-order repeats (HORs), monomers were considered the same if they shared five or fewer pairwise variants. Consecutive repeats of at least two monomers below this variant threshold were identified, yielding 2,408,653 higher-order repeats (**Fig. 2D, Table S5**). Like the *CEN180* monomer sequences, higher-order repeats are largely chromosome-specific (**Table S5**). The mean number of *CEN180* monomers per higher-order repeat was 2.41 (equivalent to 429 bp) (**Fig. 2B, Table S5**), and 95.4% of *CEN180* were a monomer of at least one larger repeat unit. Higher-order repeat block sizes show a negative exponential distribution, and the largest block was formed of 60 monomers (equivalent to 10,689 bp) (**Fig. 2B**). Many higher-order repeats are in close proximity (26% are < 100 kbp apart), although they are dispersed throughout the length of the centromeres. For example, the average distance between higher-order repeats was 380 kbp and the maximum was 2,365 kbp (**Fig. 2B, Table S5**). We also observed that higher-order repeats further apart showed a higher level of variants between the blocks (variants/monomer) (**Fig. 2F**), consistent with satellite homogenization being more effective over repeats that are physically closer. Genome-wide, the *CEN180* quantile with highest CENH3 occupancy correlates with higher-order repetition and elevated CG DNA methylation (**Fig. 2D–2E, 2G**). However, an exception to these trends is centromere 5, which has 6.8–13.4% of higher-order repeats compared to the other centromeres, yet recruits comparable CENH3 (**Fig. 2G, Table S5**).

Invasion of the Arabidopsis centromeres by *ATHILA* retrotransposons

In addition to reduced *CEN180* higher-order repetition, centromere 5 is also disrupted by breaks in the satellite array (**Fig. 2G, S8**). The majority of the main satellite arrays are *CEN180* (92.8%), with only 111 interspersed sequences >1 kbp. Within these breaks, we identified 53 intact and 20 fragmented

ATHILA long terminal repeat (LTR) retrotransposons of the *GYPSY* superfamily (**Fig. 3A–3C, Table S6**) (*18*). The intact *ATHILA* have a mean length of 11.05 kbp, and the majority have similar and paired LTRs, target site duplications, primer binding sites, polypurine tracts and *GYPSY* open reading frames (**Fig. 3C, Table S6**). LTR comparisons indicate that the centromeric *ATHILA* are young, with on average 98.7% LTR sequence identity, which was significantly higher than for *ATHILA* located outside the centromeres (96.9% n=58, Wilcox test P=4.89×10^{-8}) (**Fig. 3D, S10**). We also identified 12 *ATHILA* solo LTRs, consistent with post-integration intra-element homologous recombination (**Table S6**). We observed six instances where centromeric *ATHILA* loci were duplicated on the same chromosome and located between 8.9–538.5 kbp apart, consistent with transposons being copied post-integration, potentially via the same mechanism that generates *CEN180* higher-order repeats. For example, a pair of adjacent *ATHILA5* and *ATHILA6A* elements within centromere 5 has been duplicated within a higher-order repeat (**Fig. S11**). The duplicated elements share target site duplications and flanking sequences and show high identity between copies (99.5% and 99.6%) (**Fig. S11, Table S6**). In contrast, the surrounding *CEN180* show higher divergence and copy number variation between the higher order repeats (94.3–97.3% identity) (**Fig. S11**). This indicates an elevated rate of *CEN180* sequence change compared to the *ATHILA*, following duplication.

We analyzed centromeric *ATHILA* for CENH3 ChIP-seq enrichment and observed a decrease relative to the surrounding *CEN180*, yet higher levels than in *ATHILA* located outside of the centromere (**Fig. 3E**). The *ATHILA* show greater H3K9me2 enrichment compared to all *CEN180* (**Fig. 3E**). We used our ONT reads to profile DNA methylation over the *ATHILA* and observed dense methylation, with higher CHG-context methylation than the surrounding *CEN180* (**Fig. 3F**). Hence, *ATHILA* elements are distinct from the *CEN180* satellites at the chromatin level. We profiled *CEN180* variants around centromeric *ATHILA* loci (n=65) and observed elevated satellite divergence in the flanking regions (**Fig. 3G**), reminiscent of *Nasonia PSR* tandem repeat divergence at the junction with a *NATE* retrotransposon (*19*). This indicates that *ATHILA* insertion was mutagenic on the surrounding satellites, or that transposon insertion influenced the subsequent divergence or homogenization of the adjacent *CEN180*. We also used FISH to cytogenetically validate the presence of *ATHILA6A/6B* and *ATHILA2* sub-

families within the centromeres (**Fig 3H, S5**). Together, this shows that *ATHILA* insertions interrupt the genetic and epigenetic organization of the Arabidopsis *CEN180* arrays.

Epigenetic organization and meiotic recombination within the centromeres

To assess genetic and epigenetic features of the centromeres, we analyzed all chromosome arms along their telomere–centromere axes using a proportional scale (**Fig. 4A**). Centromere midpoints were defined as the point of maximum CENH3 ChIP-seq enrichment (**Fig. S12)**. As expected, *CEN180* satellites are highly enriched in proximity to centromeres, and these regions are relatively GC-rich compared to the AT-rich chromosome arms, at the sequence level (**Fig. 4A**). Gene density drops as the centromeres are approached, whereas transposon density increases, until they are replaced by *CEN180* (**Fig. 4A**). Gene and transposon densities are tracked closely by H3K4me3 and H3K9me2 ChIP-seq enrichment, respectively (**Fig. 4A**). H3K9me2 enrichment is observed within the centromere, although there is a reduction in the center coincident with CENH3 enrichment (**Fig. 4A**), consistent with reduced H3 occupancy caused by CENH3 replacement. A slight increase in H3K4me3 enrichment is observed within the centromeres, relative to the flanking pericentromeres (**Fig. 4A**).

Using our ONT reads with the DeepSignal-plant algorithm (*20*), we observed dense DNA methylation across the centromeres in CG, CHG and CHH contexts (**Fig. 4A–4B**). However, CHG DNA methylation shows relatively reduced centromeric frequency, compared to CG methylation (**Fig. 4A**). This may reflect centromeric depletion of H3K9me2 (**Fig. 4A**), a histone modification that maintains DNA methylation in non-CG contexts (*21*). To further investigate the DNA methylation environment associated with CENH3 deposition, we performed ChIP using either H3K9me2 or CENH3 antibodies and sequenced the immunopurified DNA with ONT. We analyzed methylation frequency in reads that aligned to the centromeres and observed dense CG methylation in both read sets, but depletion of CHG and CHH methylation in the CENH3 reads relative to H3K9me2 (**Fig. S13**). This further supports that H3 replacement by CENH3 causes a decrease in non-CG methylation maintenance within the Arabidopsis centromeres.

To investigate genetic control of centromeric DNA methylation, we analyzed bisulfite sequencing (BS-seq) data from wild type and eight mutants defective in CG and non-CG DNA methylation maintenance (**Fig. S14**) (*21*, *22*). Centromeric non-CG methylation is eliminated in *drm1 drm2 cmt2 cmt3* mutants, and reduced in *kyp suvh5 suvh6*, whereas CG methylation is intact in these backgrounds (**Fig. S14**) (*21*, *22*). In contrast, both CG and non-CG methylation in the centromeres are reduced in *ddm1* and *met1* (**Fig. S14**) (*22*). Hence, centromeric CG-context methylation is relatively high compared with non-CG, and non-CG methylation shows an unexpected dependence on CG maintenance pathways.

We observed pericentromeric ChIP-seq enrichment of the heterochromatic marks H2A.W6, H2A.W7 and H3K27me1, which are relatively depleted within the centromeres (**Fig. 4A**) (*23*, *24*). The Polycomb-group modification H3K27me3 is low in the centromeres and found largely in the chromosome arms (**Fig. 4A**). Enrichment of the euchromatic histone variant H2A.Z is low in the centromeres, but like H3K4me3, shows a slight increase in the centromeres relative to the pericentromeres (**Fig. 4A**), suggesting that the centromeres have a distinct chromatin state relative to neighbouring heterochromatin. We performed immunofluorescent staining of Arabidopsis nuclei for CENH3-GFP and euchromatic and heterochromatic histone modifications (**Fig. 4C, S15, S16**). Quantification of fluorescence intensity confirmed that heterochromatic marks are relatively depleted where CENH3-GFP is enriched (**Fig. 4C, S16**). Hence, the Arabidopsis centromeres show depletion of heterochromatic and enrichment of euchromatic marks relative to the pericentromeres, consistent with a hybrid chromatin state.

Meiotic recombination, including unequal crossover and gene conversion, has been proposed to mediate centromere evolution (*4*, *25*). We mapped 2,080 meiotic crossovers from Col×Ler F_2 sequencing data against the Col-CEN assembly (resolved on average to 1,047 kbp) (**Fig. S17**). As expected, crossovers were suppressed in proximity to the centromeres (**Fig. 4A–4B, S17**). We observed high centromeric ChIP-seq enrichment of REC8-cohesin and ASY1, which are components of the meiotic chromosome axis (**Fig. 4A**) (*26*, *27*). To investigate the potential for meiotic DSB formation within the centromeres, we aligned SPO11-1-oligonucleotides from wild type (*28*). Overall, SPO11-1-oligos are low within the

centromeres, although we observed an increase relative to the pericentromeres, reminiscent of H3K4me3 and H2A.Z ChIP-seq enrichment (**Fig. 4A**). To investigate the role of DNA methylation, we mapped SPO11-1-oligonucleotides from the CG DNA methylation mutant *met1-3* (*28*), which showed a gain of DSBs within the centromeres (**Fig. 4A–4B**). We immunostained meiocytes in early prophase I for CENH3 and V5-DMC1, which is a marker of meiotic interhomolog recombination (**Fig. 4C, S18–S19**). DMC1-V5 foci were observed along the chromosomes and adjacent to the surface of CENH3 foci, but not within them (**Fig. 4C**). Hence, despite suppression of crossovers, we observe evidence for low levels of meiotic recombination initiation within the centromeres, which is influenced by DNA methylation.

CENH3 nucleosomes show a phased pattern of enrichment with the *CEN180*, with relative depletion in spacer regions at the satellite edges (**Fig. 4D**). CENH3 spacer regions also associate with elevated DNA methylation and *CEN180* variants (**Fig. 4D**), consistent with CENH3-nucleosomes influencing epigenetic modification and satellite divergence. We analyzed chromatin and transcription around *CEN180* and *ATHILA* at the fine scale and compared wild type and the DNA methylation mutant *met1-3*. In *met1-3*, CG-context DNA methylation is lost in both *ATHILA* and *CEN180* repeats (**Fig. 4E, S20**) (*29*). However, *met1* RNA-seq and siRNA-seq signals show elevated expression of *ATHILA* transcripts, but not *CEN180* (**Fig. 4E, S20**) (*29*). The greatest RNA and siRNA expression increases in *met1-3* are observed in the *ATHILA* internal 3′ regions (**Fig. 4E, S20**), which correspond to 'TSI' transcripts and easiRNA populations (*30*, *31*). This further indicates that epigenetic regulation of the *CEN180* and *ATHILA* elements are distinct.

DISCUSSION

Leveraging advances in sequencing technology and genome assembly, we have generated the Col-CEN reference genome, which resolves the centromere satellite arrays. By profiling chromatin and recombination within the centromeres, we demonstrate that Col-CEN enables biological insights from existing functional genomics data. Using ONT long-reads we have also resolved patterns of DNA

methylation within the centromeres, highlighting the potential of complete reference assemblies for understanding epigenetic regulation of repeats. The Col-0 centromeres contain interspersed unique sequences that facilitate assembly with modern sequencing reads. However, similar to the human T2T consortium, the Col-CEN assembly required extensive manual processes to polish and curate repetitive loci (*8*, *15*, *32*). We anticipate that as complete genome assembly becomes more automated, researchers will be able to compare centromere sequences across populations and species, ultimately revealing how centromere diversity and evolution impact genome function.

In the centromeres, extensive variation is observed between the *CEN180* and the majority of monomer sequences are private to each centromere. This is consistent with satellite homogenization occurring primarily within chromosomes. The negative correlation between *CEN180* divergence and CENH3 occupancy suggests that centromeric chromatin may promote recombination pathways that lead to homogenization, including DSB formation and repair via homologous recombination. For example, interhomolog strand invasion and non-crossover repair during meiosis, using allelic or non-allelic templates, has the potential to cause *CEN180* gene conversion and structural change (**Fig. S21**). Similarly, repair and recombination using a sister chromatid may also contribute to *CEN180* change, which could occur during mitosis or meiosis (**Fig. S21**). We note that *CEN180* higher-order repeats are on average 432 bp, which is within the size range of Arabidopsis gene conversions (*33*), although we also observe large (10–100 kbp) intra-centromere duplications, for which the origin is less clear. We observe a proximity effect on divergence between *CEN180* higher-order repeats, with repeat blocks further apart showing greater differences. These patterns are reminiscent of human centromeric higher-order repeats, although duplicated blocks of alpha-satellites are longer and occur over greater physical distances (*6*, *34*, *35*). As meiotic crossover repair is suppressed within the centromeres, consistent with patterns across eukaryotes (*25*, *36*), we do not consider unequal crossover to be a major pathway driving Arabidopsis centromere evolution. However, we propose that a recombination-based homogenization process, occurring between allelic or non-allelic locations on the same chromosome, maintains the *CEN180* library close to the consensus that is optimal for CENH3 recruitment (**Fig. S21**).

Aside from homogenizing recombination within the *CEN180*, the centromeres have experienced invasion by *ATHILA* retrotransposons. The ability of *ATHILA* to insert within the centromeres is likely determined by their integrase protein. The Tal1 *COPIA* element from *Arabidopsis lyrata* also shows an insertion bias into *CEN180* when expressed in *A.thaliana (37)*, despite satellite sequences varying between these species (*38*), indicating that epigenetic information may be important for targeting. The majority of the centromeric *ATHILA* elements appear young, based on high LTR identity, and possess many features required for transposition, although the centromeres show differences in the frequency of *ATHILA* insertions, with centromeres 4 and 5 being the most invaded. Compared to *CEN180*, centromeric *ATHILA* have distinct chromatin profiles and are associated with increased satellite divergence in adjacent regions. Therefore, *ATHILA* elements represent a potentially disruptive influence on the genetic and epigenetic organization of the centromeres. However, transposons are widespread in the centromeres of diverse eukaryotes and can directly contribute to repeat evolution (e.g. mammalian CENP-B is derived from a Pogo DNA transposase) (*39*). Therefore, *ATHILA* elements may also beneficially contribute to centromere integrity and stability in Arabidopsis.

The advantage conferred to *ATHILA* by integration within the centromeres is presently unclear, although we speculate that they may be engaged in centromere drive (*40*). Haig-Grafen scrambling via recombination has been proposed as a defense against drive elements within the centromeres (*41*). For example, maize meiotic gene conversion can eliminate centromeric *CRM2* retrotransposons (*25*). Therefore, centromere satellite homogenization may serve as a mechanism to purge *ATHILA*, although in some cases this results in transposon duplication (**Fig. S22**). The presence of *ATHILA* solo LTRs is also consistent with homologous recombination acting on the retrotransposons following integration (**Fig. S22**). Centromere 5 and the diverged *CEN180* array in centromere 4, show both high *ATHILA* density and reduced *CEN180* higher-order repetition. This indicates that *ATHILA* may inhibit *CEN180* homogenization, or that loss of homogenization facilitates *ATHILA* insertion. We propose that each Arabidopsis centromere represents different stages in cycles of satellite homogenization and *ATHILA*-driven diversification. These opposing forces provide a dual capacity for homeostasis and change during centromere evolution. Assembly of centromeres from multiple Arabidopsis accessions, and closely

related species, has the potential to reveal new insights into centromere formation and the evolutionary dynamics of *CEN180* and *ATHILA* repeats.

Methods

Genomic DNA was extracted from *A.thaliana* Col-0 plants and used for ONT and PacBio HiFi long read sequencing, and Bionano optical mapping. ONT reads were used to establish a draft assembly, which was then scaffolded and polished, to generate the Col-CEN v1.2 assembly. ONT reads were used to analyse DNA methylation with the DeepSignal-plant algorithm (*20*). *CEN180* monomers, higher order repeats and *ATHILA* retrotransposons were identified *de novo* using custom pipelines. Short read datasets (**Table S7**) were aligned to Col-CEN to map chromatin and recombination distributions, using standard methods. Cytogenetic analysis of the centromeres was performed using fluorescence *in situ* hybridization and immunofluorescence staining. A full description of all experimental and computational methods can be found in the Supplementary material.

References:

1. H.S. Malik, S. Henikoff, Major evolutionary transitions in centromere complexity. *Cell*. **138**, 1067–1082 (2009).
2. D.P. Melters et al, Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).
3. K.L. McKinley, I.M. Cheeseman, The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16–29 (2016).
4. M.K. Rudd, G.A. Wray, H.F. Willard, The evolutionary dynamics of alpha-satellite. *Genome Res.* **16**, 88–96 (2006).
5. M. Jain et al, Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
6. K.H. Miga, et al, Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. **585**, 79–84 (2020).
7. G.A. Logsdon et al, The structure, function, and evolution of a complete human chromosome 8. *Nature* **593,** 101-107 (2020).
8. S. Nurk et al, The complete sequence of a human genome. *bioRxiv* https://doi.org/10.1101/2021.05.26.445798
9. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant

Arabidopsis thaliana. *Nature*. **408**, 796–815 (2000).

10. S. Maheshwari, T. Ishii, C.T. Brown, A. Houben, L. Comai, Centromere location in Arabidopsis is unaltered by extreme divergence in CENH3 protein sequence. *Genome Res.* **27**, 471–478 (2017).

11. G.P. Copenhaver et al, Genetic definition and sequence analysis of Arabidopsis centromeres. *Science*. **286**, 2468–2474 (1999).

12. P.B. Talbert, R. Masuelli, A.P. Tyagi, L. Comai, S. Henikoff, Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. *Plant Cell*. **14**, 1053–1066 (2002).

13. J.M. Martinez-Zapater, M.A. Estelle, C.R. Somerville, A highly repeated DNA sequence in Arabidopsis thaliana. *Mol. Gen. Genet.* **204**, 417–423 (1986).

14. E.K. Round, S.K. Flowers, E.J. Richards, Arabidopsis thaliana centromere regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7**, 1045–1053 (1997).

15. A.M. McCartney et al, Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies, *bioRxiv* doi:10.1101/2021.07.02.450803.

16. T. Hosouchi, N. Kumekawa, H. Tsuruoka, H. Kotani, Physical map-based sizes of the centromeric regions of Arabidopsis thaliana chromosomes 1, 2, and 3. *DNA Res.* **9**, 117–121 (2002).

17. A. Rhie, B.P. Walenz, S. Koren, A.M. Phillippy, Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

18. D.A. Wright, D.F. Voytas, Athila4 of Arabidopsis and Calypso of soybean define a lineage of endogenous plant retroviruses. *Genome Res.* **12**, 122–131 (2002).

19. B.F. McAllister, J.H. Werren, Evolution of Tandemly Repeated Sequences: What Happens at the End of an Array? *Journal of Molecular Evolution*. **48** (1999), pp. 469–481.

20. P. Ni et al, Genome-wide Detection of Cytosine Methylations in Plant from Nanopore sequencing data using Deep Learning. *bioRxiv* https://doi.org/10.1101/2021.02.07.430077

21. H. Stroud et al, Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nat. Struct. Mol. Biol.* **21**, 64–72 (2014).

22. H. Stroud et al, Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell*. **152**, 352–364 (2013).

23. Y. Jacob et al, ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nat. Struct. Mol. Biol.* **16**, 763–768 (2009).

24. R. Yelagandula et al, The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in Arabidopsis. *Cell*. **158**, 98–109 (2014).

25. J. Shi et al, Widespread gene conversion in centromere cores. *PLoS Biol.* **8**, e1000327 (2010).

26. C. Lambing et al, Interacting Genomic Landscapes of REC8-Cohesin, Chromatin, and Meiotic Recombination in Arabidopsis. *Plant Cell*. **32**, 1218–1239 (2020).

27. C. Lambing et al, ASY1 acts as a dosage-dependent antagonist of telomere-led recombination and mediates crossover interference in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 13647–13658 (2020).

28. K. Choi et al, Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis thaliana transposons and gene regulatory regions. *Genome Res.* **28**, 532–546 (2018).

29. M. Rigal et al, Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in Arabidopsis thaliana F1 epihybrids. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E2083–92 (2016).

30. A. Steimer et al, Endogenous targets of transcriptional gene silencing in Arabidopsis. *Plant Cell*. **12**, 1165–1178 (2000).

31. S. C. Lee et al, Arabidopsis retrotransposon virus-like particles and their regulation by epigenetically activated small RNA. *Genome Res.* **30**, 576–588 (2020).

32. A. Rhie et al, Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737-746 (2021).

33. E. Wijnker et al, The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *Elife*. **2**, e01426 (2013).

34. S.J. Durfy, H.F. Willard, Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: evidence for short-range homogenization of tandemly repeated DNA sequences. *Genomics*. **5**, 810–821 (1989).

35. N. Altemose et al, Complete genomic and epigenetic maps of human centromeres. *bioRxiv* (2021), p. 2021.07.12.452052.

36. M.M. Mahtani, H.F. Willard, Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Res.* **8**, 100–110 (1998).

37. S. Tsukahara et al, Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata. *Genes Dev.* **26**, 705–713 (2012).

38. A. Kawabe, S. Nasuda, Structure and genomic organization of centromeric repeats in Arabidopsis species. *Mol. Genet. Genomics*. **272**, 593–602 (2005).

39. S. J. Klein, R. J. O'Neill, Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Research*. **26** (2018), pp. 5–23.

40. H. S. Malik, The centromere-drive hypothesis: a simple basis for centromere complexity. *Prog. Mol. Subcell. Biol.* **48**, 33–52 (2009).

41. D. Haig, A. Grafen, Genetic scrambling as a defence against meiotic drive. *Journal of Theoretical Biology*. **153** (1991), pp. 531–558.

Acknowledgements:

We thank Isabel Thompson for *ATHILA* analysis, Steve Henikoff for the generous gift of CENH3 antibodies, Alaina Shumate for help with gene Liftoff interpretation, Bettina Fischer for advice on high molecular weight DNA isolation and Milan Pouch for assistance designing FISH probes. **Funding:** This work was supported by BBSRC grants BB/S006842/1, BB/S020012/1 and BB/V003984/1,

European Research Council Consolidator Award ERC-2015-CoG-681987 'SynthHotSpot' and Marie Curie International Training Network 'MEICOM' to IH, Human Frontier Science Program award RGP0025/2021 to TK, MCS and IH, US National Institutes of Health Grant S10OD028632-01, US National Science Foundation grants DBI-1350041 and IOS-1732253 to MCS, Royal Society awards UF160222 and RGF/R1/180006 to AB, the Czech Science Foundation grant no. 21-03909S to TM and MAL, and by the Gregor Mendel Institute (FB), grants Fonds zur Förderung der wissenschaftlichen Forschung (FWF) P26887, P28320, P30802, P32054, and TAI304 to FB and DK, and chromatin dynamics W1238 to AS and BJ, and Leverhulme Trust Research Leadership grant RL-2012-042 to JT. The authors have no competing interests. **Author contributions:** MN sequenced DNA, performed genome assembly and analysis, ChIP-seq, DNA methylation analysis and wrote the manuscript. MA performed genome assembly, polishing, validation, annotation and analysis and wrote the manuscript. PW performed satellite repeat annotation, genome analysis and wrote the manuscript. AJT performed short read alignment, genome analysis and wrote the manuscript. BA sequenced DNA, performed optical mapping and contributed to the assembly. AS performed chromatin immunofluorescence analysis. BJ provided ChIP-seq data. CL and PK performed immunocytology. NE generated the DMC1 epitope-tagged line. NH and KC sequenced DNA and contributed to the assembly. LS, JT and KS performed PacBio sequencing. TK and RM provided intellectual input. TM and MAL performed FISH. FB supervised ChIP-seq, immunofluorescence analysis and wrote the manuscript. AB performed *ATHILA* annotation and genome analysis and wrote the manuscript. TM supervised DNA sequencing, genome assembly and analysis and wrote the manuscript. MCS supervised genome assembly, validation, annotation and analysis and wrote the manuscript. IH supervised DNA sequencing, genome assembly, validation, annotation and analysis and wrote the manuscript. **Competing Interests:** The authors have no competing interests. **Data and materials availability:** The ONT sequencing reads used for assembly are available for download at ArrayExpress accession E-MTAB-10272 (http://www.ebi.ac.uk/arrayexpress/) (Username: Reviewer_E-MTAB-10272 Password: YVJAaVii). The PacBio HiFi reads are available for download at European Nucleotide Archive accession number PRJEB46164 (https://www.ebi.ac.uk/ena/browser/view/PRJEB46164). All data, code and materials are

available in the manuscript or the supplementary materials and at https://github.com/schatzlab/Col-CEN.

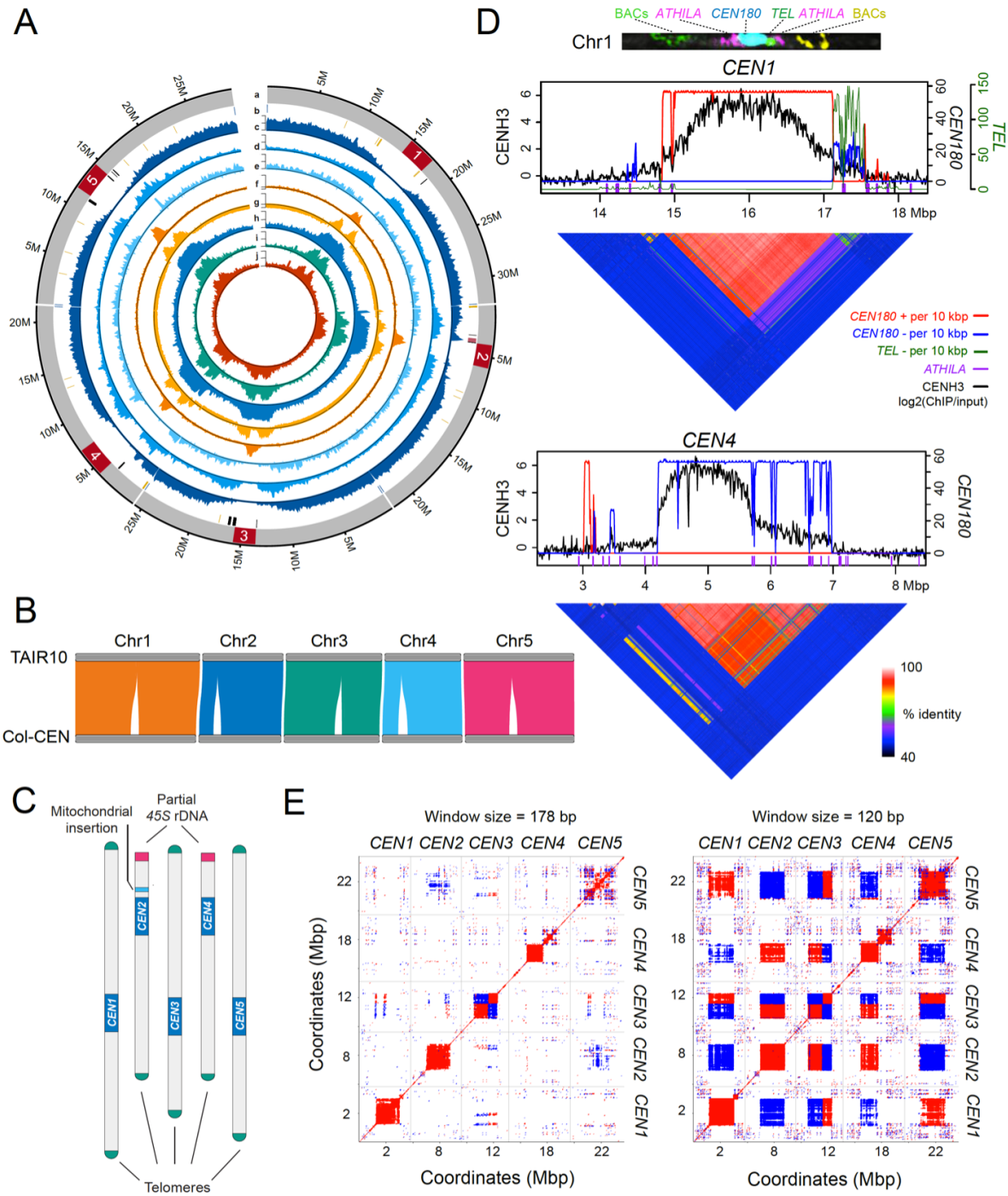

A
B
TAIR10
Col-CEN
Chr1
Chr2
Chr3
Chr4
Chr5
C
Mitochondrial insertion
Partial 45S rDNA
CEN1
CEN2
CEN3
CEN4
CEN5
Telomeres
D
BACs ATHILA CEN180 TEL ATHILA BACs
Chr1
CEN1
CENH3
CEN180
TEL
14
15
16
17
18 Mbp
CEN180 + per 10 kbp
CEN180 - per 10 kbp
TEL - per 10 kbp
ATHILA
CENH3
log2(ChIP/input)
CEN4
3
4
5
6
7
8 Mbp
100
% identity
40
E
Window size = 178 bp
Window size = 120 bp
CEN1 CEN2 CEN3 CEN4 CEN5
Coordinates (Mbp)
F
DAPI
CEN180-α
CEN180-β
Chr1 BACs
Chr1

Figure 1. Complete assembly of the Arabidopsis centromeres. A. Circos plot of the Col-CEN assembly. Quantitative tracks (c-j) are aggregated in 100-kbp bins and independent y-axis labels are given as (low value, mid value, high value, measurement unit): (a) chromosome with centromeres shown in red; (b) telomeres (blue), *45S* rDNA (yellow), *5S* rDNA (black) and the mitochondrial insertion (pink); (c) genes (0, 25, 51, gene number); (d) transposable elements (0, 84, 167, transposable element number); (e) Col×Ler F_2 crossovers (0, 7, 14, crossover number); (f) CENH3 (-0.5, 0, 3, $\log_2$(ChIP/input)); (g) H3K9me2 (-0.6, 0, 2, $\log_2$(ChIP/input)); (h) CG methylation (0, 47, 95, %); (i) CHG methylation (0, 28, 56, %); (j) CHH methylation (0, 7, 13, %). **B.** Syntenic alignments between the TAIR10 and Col-CEN assemblies. **C.** Col-CEN ideogram with annotated chromosome landmarks (not drawn to scale). **D.** CENH3 $\log_2$(ChIP/input) (black) plotted over centromeres 1 and 4 (*10*). *CEN180* per 10-kbp plotted for forward (red) or reverse (blue) strand orientations. *ATHILA* are indicated by purple x-axis ticks. Heatmaps show pairwise sequence identity between all non-overlapping 5-kbp regions. A FISH-stained chromosome 1 at pachytene is shown above, probed with upper-arm BACs (green), *ATHILA* (purple), *CEN180* (blue), the telomeric repeat (green) and bottom-arm BACs (yellow). **E.** Dotplots comparing the five centromeres using a search window of 120 or 178 bp. Red and blue indicate detection of similarity on the same or opposite strands. **F.** Pachytene-stage chromosomes stained with DAPI (black) and *CEN180-α* (red), *CEN180-β* (purple) and chromosome 1 BAC (green) FISH probes. The scale bar represents 10 μM.

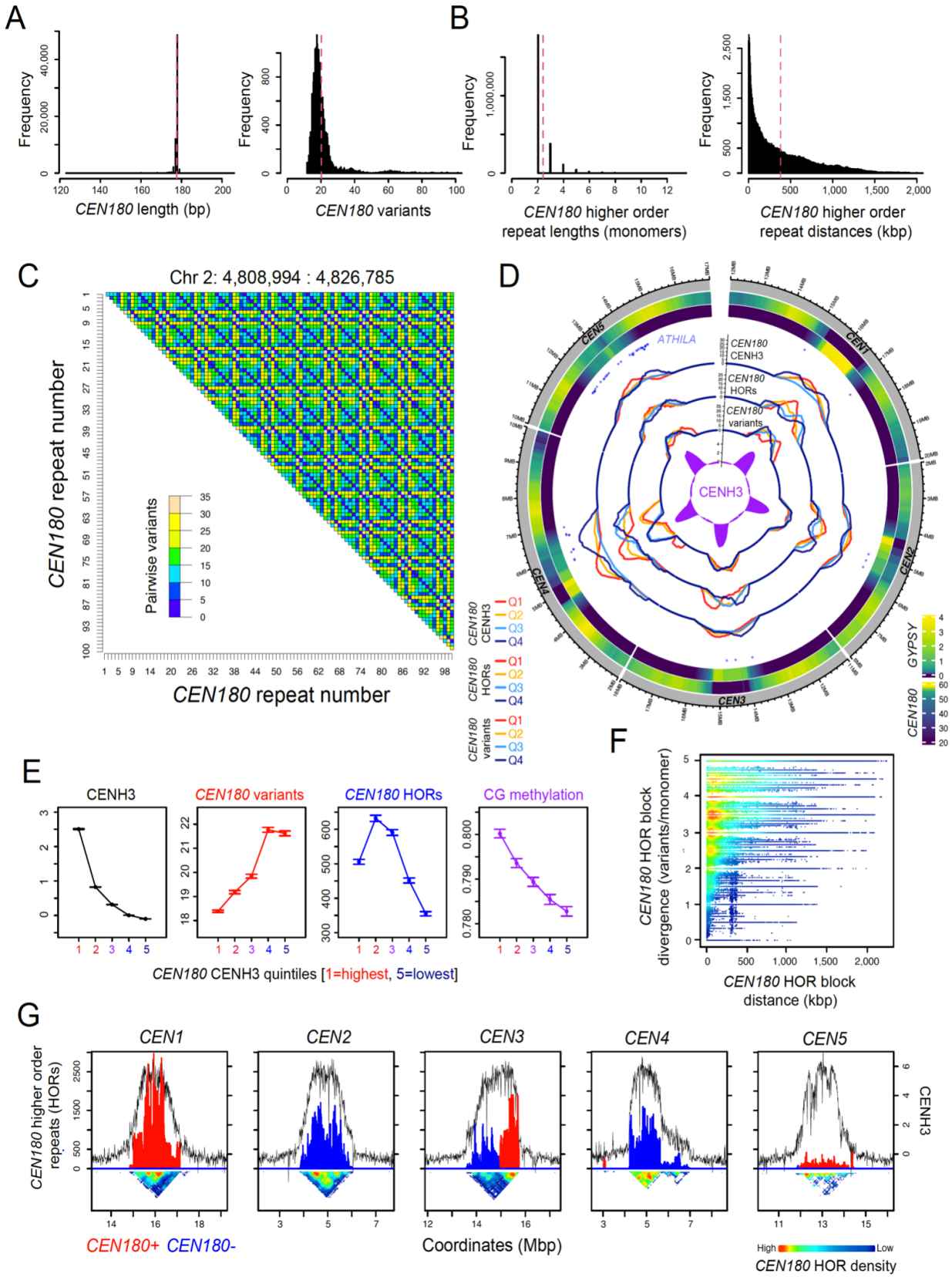

A
Frequency
CEN180 length (bp)
CEN180 variants
B
CEN180 higher order repeat lengths (monomers)
CEN180 higher order repeat distances (kbp)
C
Chr 2: 4,808,994 : 4,826,785
CEN180 repeat number
Pairwise variants
D
ATHILA
CEN180 CENH3
CEN180 HORs
CEN180 variants
CENH3
GYPSY
CEN180
E
CENH3
CEN180 variants
CEN180 HORs
CG methylation
CEN180 CENH3 quintiles [1=highest, 5=lowest]
F
CEN180 HOR block divergence (variants/monomer)
CEN180 HOR block distance (kbp)
G
CEN1
CEN2
CEN3
CEN4
CEN5
CEN180 higher order repeats (HORs)
CENH3
CEN180+ CEN180-
Coordinates (Mbp)
High
Low
CEN180 HOR density

Figure 2. The Arabidopsis *CEN180* satellite repeat library. A. Histograms of *CEN180* monomer lengths (bp), and variant distances relative to the genome-wide consensus (mean=red dotted lines). **B.** As for A, but showing widths of *CEN180* higher order repeat (HOR) blocks (monomers, 'mers'), and the distance between HORs (kbp). **C.** Heatmap of a representative region within centromere 2, shaded according to pairwise variants between *CEN180*. **D.** Circos plot showing: (i) *GYPSY* density; (ii) *CEN180* density; (iii) centromeric *ATHILA* 'rainfall'; (iv) *CEN180* density grouped by decreasing CENH3 $\log_2$(ChIP/input) (red=high; navy=low); (v) *CEN180* density grouped by decreasing higher-order repetition (red=high; navy=low); (vi) *CEN180* grouped by decreasing variant distance (red=high; navy=low); and (vii) CENH3 $\log_2$(ChIP/input) (purple), across the centromeres. **E.** *CEN180* were divided into quintiles according to CENH3 $\log_2$(ChIP/input) and mean values with 95% confidence intervals plotted. The same groups were analyzed for *CEN180* variant distance (red), higher-order repetition (blue) and CG-context DNA methylation (purple). **F.** Plot of the distance between pairs of HORs (kbp) and divergence (variants/monomers) between the HORs. **G.** Plots of CENH3 $\log_2$(ChIP/input) (black) across the centromeres, compared to *CEN180* higher-order repetition on forward (red) or reverse (blue) strands. The heatmap beneath is shaded according to HOR density.

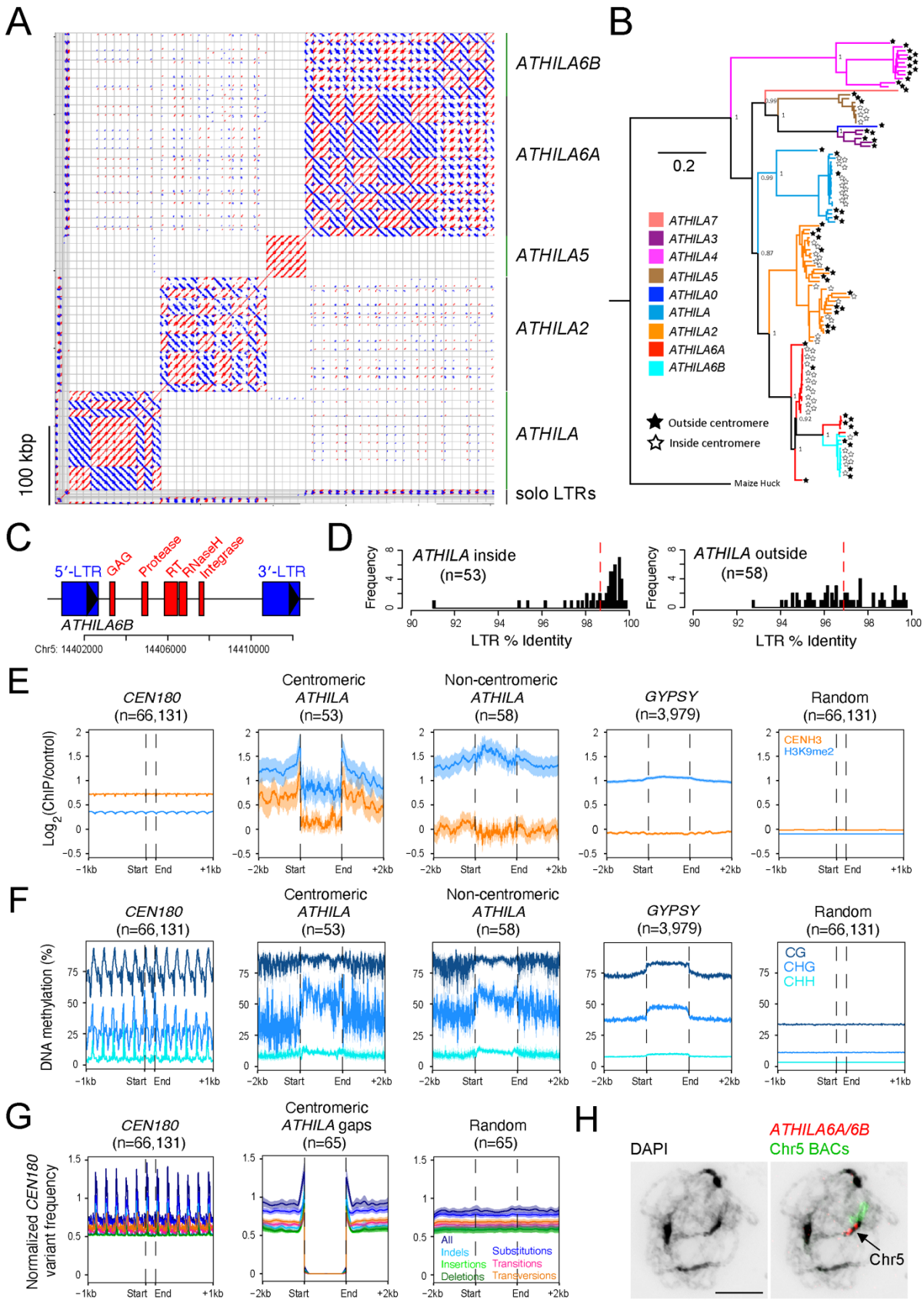

A
ATHILA6B
ATHILA6A
ATHILA5
ATHILA2
ATHILA
solo LTRs
100 kbp
B
0.2
ATHILA7
ATHILA3
ATHILA4
ATHILA5
ATHILA0
ATHILA
ATHILA2
ATHILA6A
ATHILA6B
Outside centromere
Inside centromere
Maize Huck
C
5′-LTR
GAG
Protease
RT
RNaseH
Integrase
3′-LTR
ATHILA6B
Chr5: 14402000
14406000
14410000
D
ATHILA inside
(n=53)
ATHILA outside
(n=58)
Frequency
LTR % Identity
E
CEN180
(n=66,131)
Centromeric
ATHILA
(n=53)
Non-centromeric
ATHILA
(n=58)
GYPSY
(n=3,979)
Random
(n=66,131)
Log2(ChIP/control)
CENH3
H3K9me2
F
DNA methylation (%)
CG
CHG
CHH
G
Normalized CEN180
variant frequency
Centromeric
ATHILA gaps
(n=65)
Random
(n=65)
All
Indels
Insertions
Deletions
Substitutions
Transitions
Transversions
H
DAPI
ATHILA6A/6B
Chr5 BACs
Chr5

Figure 3. Invasion of the Arabidopsis centromeres by *ATHILA* retrotransposons. A. Dotplot of centromeric *ATHILA* using a 50-bp search window. Red and blue indicate forward- and reverse-strand similarity. *ATHILA* subfamilies and solo LTRs are indicated. **B.** Maximum likelihood phylogenetic tree of 111 intact *ATHILA* elements, color-coded according to subfamily. Stars at the branch tips indicate *ATHILA* inside (white) or outside (black) the centromeres. **C.** An annotated map of an *ATHILA6B* with LTRs (blue) and core protein domains (red) highlighted. **D.** Histograms of LTR sequence identity for centromeric *ATHILA* elements (n=53), compared to *ATHILA* outside of the centromeres (n=58). (Mean values=red dashed lines). **E.** Meta-profiles of CENH3 (orange) and H3K9me2 (blue) ChIP-seq signals around *CEN180* (n=66,131), centromeric intact *ATHILA* (n=53), *ATHILA* located outside the centromeres (n=58), *GYPSY* retrotransposons (n=3,979), and random positions (n=66,131). Shaded ribbons represent 95% confidence intervals for windowed mean values. **F.** As for E, but analyzing ONT-derived percent DNA methylation in CG (dark blue), CHG (blue) and CHH (light blue) contexts. **G.** Meta-profiles of *CEN180* sequence edits (insertions, deletions and substitutions relative to the *CEN180* consensus), normalized by *CEN180* presence/absence, in positions surrounding *CEN180* gaps containing *ATHILA* (n=65), or random positions (n=65). All edits (dark blue), substitutions (blue), indels (light blue), insertions (light green), deletions (dark green), transitions (pink) and transversions (orange) are shown. Shaded ribbons represent 95% confidence intervals for windowed mean values. **H.** Pachytene-stage chromosome spread stained with DAPI (black), an *ATHILA6A/6B GAG* FISH probe (red) and chromosome 5 specific BACs (green). Scale bar represents 10 μM.

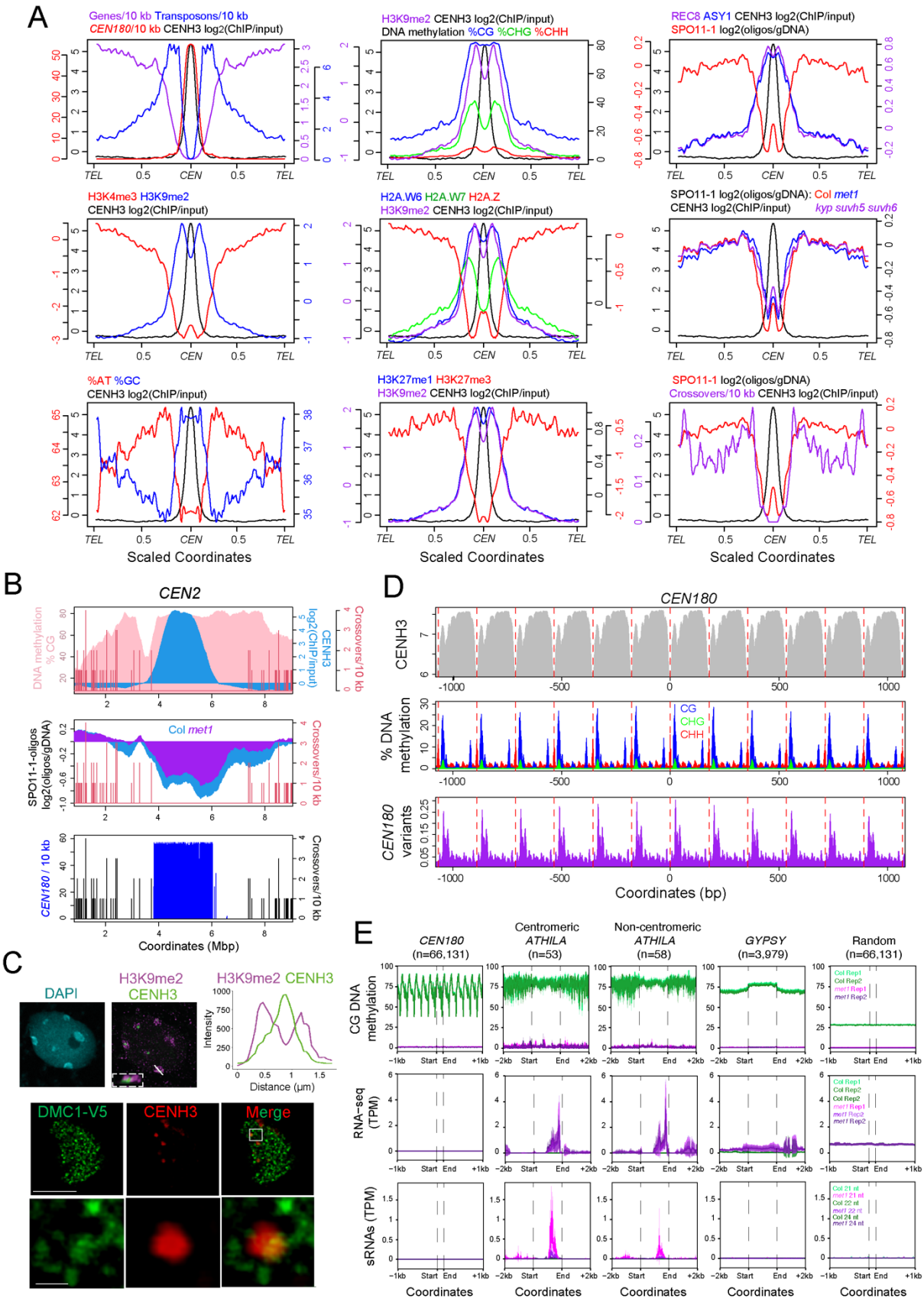

A
Genes/10 kb Transposons/10 kb
CEN180/10 kb CENH3 log2(ChIP/input)
H3K9me2 CENH3 log2(ChIP/input)
DNA methylation %CG %CHG %CHH
REC8 ASY1 CENH3 log2(ChIP/input)
SPO11-1 log2(oligos/gDNA)
H3K4me3 H3K9me2
CENH3 log2(ChIP/input)
H2A.W6 H2A.W7 H2A.Z
H3K9me2 CENH3 log2(ChIP/input)
SPO11-1 log2(oligos/gDNA): Col met1
CENH3 log2(ChIP/input) kyp suvh5 suvh6
%AT %GC
CENH3 log2(ChIP/input)
H3K27me1 H3K27me3
H3K9me2 CENH3 log2(ChIP/input)
SPO11-1 log2(oligos/gDNA)
Crossovers/10 kb CENH3 log2(ChIP/input)
TEL 0.5 CEN 0.5 TEL
Scaled Coordinates
B
CEN2
DNA methylation %CG
CENH3 log2(ChIP/input)
Crossovers/10 kb
SPO11-1-oligos log2(oligos/gDNA)
Col met1
CEN180 / 10 kb
Coordinates (Mbp)
C
DAPI
H3K9me2 CENH3
H3K9me2 CENH3
Intensity
Distance (µm)
DMC1-V5
CENH3
Merge
D
CEN180
CENH3
% DNA methylation
CG
CHG
CHH
CEN180 variants
Coordinates (bp)
E
CEN180 (n=66,131)
Centromeric ATHILA (n=53)
Non-centromeric ATHILA (n=58)
GYPSY (n=3,979)
Random (n=66,131)
CG DNA methylation
RNA-seq (TPM)
sRNAs (TPM)
Col Rep1
Col Rep2
met1 Rep1
met1 Rep2
Col 21 nt
met1 21 nt
Col 22 nt
met1 22 nt
Col 24 nt
met1 24 nt
Start End
Coordinates

Figure 4. Epigenetic organization and meiotic recombination within the centromeres. A. Quantification of genomic features plotted along chromosome arms that were proportionally scaled between telomeres (*TEL*) and centromere midpoints (*CEN*) (defined by maximum CENH3 ChIP-seq $\log_2$(ChIP/input) enrichment). Data analyzed were gene, transposon and *CEN180* density, CENH3, H3K4me3, H3K9me2, H2A.W6, H2A.W7, H2A.Z, H3K27me1, H3K27me3, REC8 and ASY1 $\log_2$(ChIP/input), % AT/GC base composition, DNA methylation, SPO11-1-oligos (in wild type and *met1*) and crossovers (**Table S7**). **B.** Plot quantifying crossovers (red), % CG DNA methylation (pink), CENH3 (blue), SPO11-1-oligos in wild type and *met1*, and *CEN180* density along centromere 2. **C.** An interphase nucleus immunostained for H3K9me2 (magenta) and CENH3-GFP (green). The white line indicates the confocal section used for the intensity plot shown on the right. Scale bar represents 5 μM. Beneath is a male meiocyte (early prophase I) immunostained for CENH3 (red) and V5-DMC1 (green). Scale bars are 10 μM (upper) and 1 μM (lower). **D.** Plots of CENH3 ChIP enrichment (grey), DNA methylation in CG (blue), CHG (green) and CHH (red) contexts and *CEN180* variants (purple), averaged over windows centered on *CEN180* starts. The red lines show 178-bp increments. **E.** Meta-profiles of CG-context DNA methylation, RNA-seq and siRNA-seq in wild type (green) or *met1* (pink/purple) (*29*), around *CEN180* (n=66,131), centromeric intact *ATHILA* (n=53), *ATHILA* located outside the centromeres (n=58), *GYPSY* (n=3,979) and random positions (n=66,131). Shaded ribbons represent 95% confidence intervals for windowed mean values.

Supplementary Materials

1. **Materials and Methods**
2. **Figures S1 – S22**
3. **Tables S1 – S8**
4. **Supplementary References (*42-94*)**

Materials and Methods

Genomic DNA extraction and ONT and PacBio HiFi sequencing

For genomic DNA extraction associated with ONT sequencing, 3 week-old Col-0 seedlings were grown on ½ MS media and 1% sucrose and kept in the dark for 48 hours prior to harvesting. Approximately 10 g of tissue was used per 200 ml of MPD-Based Extraction Buffer pH 6 (MEB). Tissue was flash frozen and ground tissue in liquid nitrogen, using a pestle and mortar, and resuspended in 200 ml MEB. Ground tissue was thawed in MEB with frequent stirring. The homogenate was forced through 4 layers of miracloth, and then filtering again through 4 layers of fresh miracloth by gravity. 20% Triton x-100 was added to a final concentration of 0.5% on ice, followed by incubation with agitation on ice for 30 minutes. The suspension was centrifuged at 800*g* for 20 minutes at 4ºC. The supernatant was removed and the pellet resuspended using a paintbrush in 10 ml 2-methyl-2,4 pentanediol buffer pH 7.0 (MPDB). The suspension was centrifuged at 650*g* for 20 minutes at 4ºC. The supernatant was removed and the pellet was washed with 10 ml of MPDB. Washing and centrifugation was repeated until the pellet appeared white and was finally resuspended in a minimal volume of MPDB. From this point onwards all transfers were performed using wide bore pipette tips. 5 ml CTAB buffer was added to the nuclei pellet and mixed via gentle inversion, followed by incubation at 60ºC until full lysis had occurred, taking between 30 minutes and 2 hours. An equal volume of chloroform was added and incubated on a rocking platform, with a speed of 18 cycles per minute, for 30 minutes, followed by centrifugation at 3000*g* for 10 minutes. An equal volume of phenol/chloroform/isoamyl alcohol (PCI, 25:24:1) was

added to the lysate, followed by incubation on a rocking platform (18 cycles per minute) for 30 minutes. The lysate was centrifuged at 3000*g* for 10 minutes and the upper aqueous phase was transferred into a fresh tube. The PCI extraction was then repeated. The extraction was then repeated using only chloroform. 1/10th volume of 3M Sodium Acetate was added to the lysate and mixed by gentle inversion. Two volumes of ice cold ethanol were added and mixed by inversion. DNA was precipitated at -20ºC for 48 hours. The precipitated DNA was removed using a glass hook and washed three times in fresh 70% ethanol. The DNA was dissolved in 120 µl of 10 mM Tris-Cl (pH 8.5).

Approximately 5 µg of DNA was size selected to be >30 kbp, using the BluePippin™ Size-Selection System (Sage Science) and the 0.75% DF Marker U1 cassette definition, with Range mode and BP start set at 30,000 bp. Library preparation followed the Nanopore SQK-LSK109 protocol and kit**.** Approximately 1.2-1.5 µg of size-selected DNA in a volume of 48 µl was used for library preparation. DNA was nic-repaired and end-prepped by the addition of 3.5 µl of NEBNext FFPE Buffer and NEBNext Ultra II End Prep Reaction Buffer, followed by 2 µl of NEBNext DNA Repair Mix and 3 µl NEBNext Ultra II End Prep Enzyme Mix (New England Biolab, E7180S), with incubation for 30 minutes at 20°C, followed by 30 minutes at 65°C. The sample was cleaned using 1×volume AMPure XP beads and eluted in 61 µl of nuclease-free water. Adapters were ligated at room temperature using 25 µl Ligation Buffer, 10 µl NEBNext T4 DNA Ligase and 5 µl Adapter Mix for 2 hours. The library was cleaned with 0.4×volume AMPure XP beads, washed using ONT Long Fragment buffer and eluted in 15 µl elution buffer.

For genomic DNA associated with PacBio HiFi sequencing, Col-0 plants were grown at the Max Planck Institute for Plant Breeding Research, Cologne, Germany. DNA extraction (from an individual plant), library preparation and DNA sequencing was performed at the Max Planck Genome Center, Cologne, Germany. High molecular weight DNA was isolated from 1.5 gram of vegetative material with a NucleoBond HMW DNA kit (Macherey Nagel). Quality was assessed with a FEMTOpulse device (Agilent) and quantity measured by fluorometry Quantus (Promega). A HiFi library was then prepared

according to the manual "Procedure & Checklist - Preparing HiFi SMRTbell® Libraries using SMRTbell Express Template Prep Kit 2.0" with initial DNA fragmentation by g-Tubes (Covaris) and final library size binning by SageELF (Sage Science). Size distribution was again controlled by FEMTOpulse (Agilent). The size-selected library was sequenced on one SMRTcell on a Sequel II device with Binding kit 2.0 and Sequel II Sequencing Kit 2.0 for 30 hours.

Col-CEN genome assembly

Libraries were sequenced on 6 ONT R9 flow cells and 1 ONT R10 flow cell, and the resulting .fast5 files were basecalled with Guppy (v4.0.15), using the dna_r9.4.1_450bps_hac.cfg and dna_r10.3_450bps_hac.cfg configurations, respectively. This yielded a total of 73.6 Gb of sequence (~613× total coverage). The fastq files of ONT reads used for genome assembly are available for download at ArrayExpress accession E-MTAB-10272 (http://www.ebi.ac.uk/arrayexpress/). We trimmed adapters using Porechop (v0.2.4) and filtered for read lengths greater than 30 kbp and mean read quality scores >90%, using Filtlong (v0.2.0) (https://github.com/rrwick/Filtlong), which yielded 436,146 reads with a mean length of 43.9 kbp (19.15 Gbp), equivalent to 161× coverage of the TAIR10 genome with ~55x coverage of ultra-long reads (>50 kbp). Flye (version 2.7) was used to assemble the reads, specifying a minimum read overlap of 10 kbp and a *k*-mer size of 17 (*42*).

Contig screen

We performed a comprehensive contig screen using methods inspired by the Vertebrate Genomes Project (VGP), though adapted for an inbred plant genome (*32*). We first aligned Flye contigs to the Columbia reference chloroplast (GenBank accession NC_000932.1) (*43*), and mitochondria (GenBank accession NC_037304.1) (*44*) genomes with Minimap2 (v2.17-r941, -x asm5) (*45*). Contigs with at least 50% of their bases covered by alignments were considered to be chloroplast or mitochondria genome sequences and were removed from the assembly.

We next used BLAST to screen for contigs representing bacterial contamination. We first masked the Flye assembly with windowmasker (v1.0.0, -mk_counts -genome_size 131405362) (*46*). We then

aligned the Flye contigs to all RefSeq bacterial genomes (downloaded on 2020/05/21) with megablast (v2.5.0, -outfmt "6 std score"), providing the windowmasker annotations with "-window_masker_db" (*47*). We removed BLAST alignments with an E value greater than or equal to 0.0001, a score less than 500, and a Percent Identity less than 98%, and any contigs (four in total) with remaining alignments were manually inspected. Two of the four contigs were already identified as being chloroplast or mitochondria sequence and the other two were clearly nuclear contigs, so we determined that no contigs were derived from bacterial contaminants.

After removing chloroplast and mitochondria contigs, we performed one final screen to remove contigs with low read support. We aligned ONT reads (>=40 kbp) to the contigs with Minimap2 (v2.17-r941, -x map-ont) and removed any contigs (one in total) with more than 50% of its bases covered by fewer than 15 reads. Though we did not use its standard pipeline, we made use of purge_dups scripts for this analysis (*48*). After screening, the assembly consisted of 10 contigs with an N50 of 22,078,741 bp.

Contig scaffolding

Though the five Columbia chromosomes were represented by only 10 contigs, we used homology-based scaffolding to order and orient contigs, assign chromosome labels, and orient pseudomolecules to match the orientation of TAIR10 chromosomes. We ran RagTag (v1.0.1, --debug --aligner=nucmer --nucmer-params='--maxmatch -l 100 -c 500') using TAIR10 as the reference genome, but excluding ChrC and ChrM (-e) (*49*, *50*). Three small contigs (3,200, 90,237 and 8,728 bp) consisting of low complexity sequence were not ordered and oriented and were removed from the assembly. After scaffolding, the 131,388,895 bp assembly was represented in five pseudomolecules corresponding to the five chromosomes of the Columbia genome. Chromosome 1 was gapless, while the other chromosomes contained one to four 100 bp gaps each (9 in total).

Initial pseudomolecule polishing and gap filling

We corrected mis-assemblies and filled gaps in the Columbia pseudomolecules with two rounds of Medaka (v1.2.1) ONT polishing (https://github.com/nanoporetech/medaka). For the first round of

polishing, we aligned R9 ONT reads (>=50 kbp) to the pseudomolecules with mini_align (minimap2 v2.17-r941, -m). To avoid overcorrection in the centromere satellite sequences, we performed "marker-assisted filtering" to remove alignments not anchored in putatively unique sequences (*6*, *15*) (https://github.com/malonge/T2T-Polish). We defined "marker" *k*-mers as 21-mers that occurred once in the assembly and between 14 and 46 times (inclusive) in the Illumina reads. The first round of polishing was completed using `medaka consensus` (--model r941_min_high_g360 --batch_size 200) and `medaka stitch`. The second round of polishing was performed as for the first round, except we aligned all R10 reads instead of R9 reads and the `medaka consensus` model was set to "r103_min_high_g360". As a result of ONT polishing, the assembly improved from a QV of 32.38 to 33.17 and 34.12 after the first and second rounds, respectively (*17*). After medaka polishing, the assembly contained only a single gap on chromosome 2.

Long-read ONT polishing was followed by short-read polishing of non-centromeres with DeepVariant (*51*). We first aligned Col-0 genomic DNA Illumina reads to the pseudomolecules with bwa mem (v0.7.17-r1198-dirty) and we compressed and sorted alignments with samtools (v1.10) (*52*, *53*). We then created a VCF file of potential polishing edits with DeepVariant (v1.1.0, --model_type=WGS),"bcftools view" (v1.11, -e 'type="ref"' -i 'QUAL>1 && (GT="AA" || GT="Aa")') and "bcftools norm". To avoid error-prone short-read polishing in the centromeres, we used Bedtools to remove polishing edits within the centromeres and we used BCFtools to derive a final consensus FASTA file (*54*, *55*). Though short-read polishing did not alter the centromeres, it improved the overall assembly QV to 41.4616.

Telomere patching

We locally re-assembled and patched telomeric sequences for the 8 Columbia telomeres not adjacent to NORs (all but the beginning of chromosomes 2 and 4). We aligned all R9 reads to the TAIR10 reference with Winnowmap (v1.11, k=15, --MD -ax map-ont) and for each telomere, we collected all reads that aligned once to within 50 bp of the chromosome terminus (*56*). Using Bowtie (*57*) (v1.3.0, -S --all -v 0), we counted the occurrences of the telomeric repeat motif ('CCCTAAA') in each read, and the read

with the most occurrences was designated as the "reference" and all other reads were designated as the "query". Local re-assembly was completed by aligning the query reads to the reference read and computing a consensus with `medaka_consensus` (v1.2.1, -m r941_min_high_g360). To patch these telomere consensus sequences into the Columbia pseudomolecules, we identified the terminal BAC sequences for each of the 8 chromosome arms. For each chromosome arm, we aligned the terminal BAC sequence to the Columbia pseudomolecules and the telomere consensus sequence with Nucmer (v3.1, --maxmatch). Using these alignment coordinates, the consensus sequences were manually patched such that everything after the terminal BAC sequence was replaced with telomere consensus sequence. Telomeres were then manually confirmed to be structurally valid.

Assembly curation and preparation

After polishing and telomere patching, we performed final curation steps to correct lingering misassemblies and screen for contamination. First, while it was not straightforward to fill the remaining chromosome 2 gap *de novo*, we were able to replace the gap locus with the corresponding region in TAIR10. We found two BAC sequences flanking the gap locus that aligned concordantly to both the Col-0 pseudomolecules and TAIR10. These BAC contigs were aligned to the pseudomolecules and TAIR10 with Nucmer (v3.1, --maxmatch -l 250 -c 500) and the gap locus between the BAC contigs in the Columbia pseudomolecules was replaced with the corresponding TAIR10 locus between the BAC contigs.

To identify and correct structural mis-assemblies, we aligned Columbia long-reads to the Columbia pseudomolecules and called structural variants (SVs). First, we used Bedtools `random` (v2.29.2, -l 100000 -n 50000 -seed 23) to simulate 50,000 100 kbp exact reads from TAIR10. These reads, along with R9 (>=50 kbp) and R10 Columbia reads were aligned to the Columbia pseudomolecules with Winnowmap (v1.11, k=15, "--MD -ax map-pb" for TAIR10 reads and "--MD -ax map-ont" for ONT reads). After compressing and sorted alignments with samtools (v1.10), Sniffles (v1.0.12, -d 100 -n -1 -s 3) was used to infer SVs from each of the alignments (*58*). SVs with fewer than 30% of reads supporting the ALT allele were removed and the three resulting VCF files were merged with Jasmine

(v1.0.10, max_dist=500 spec_reads=3 --output_genotypes) (*59*). There were a total of three variants called by all three read sets, including two deletions and one insertion that we corrected. REF and ALT alleles for these SVs were manually refined and validated, and ALT alleles were incorporated into the pseudomolecules using `bcftools consensus`.

Next, we manually inspected all gaps filled by Medaka and found that a 181 bp region containing a 100 bp gap on chromosome 5 was incorrectly replaced with 103 bp of sequence and we manually replaced the filled sequence with the original gap locus. This ultimately produced the Col-CEN v1.1 assembly. We used VecScreen to do a final contamination screen. We first aligned the Columbia pseudomolecules to the VecScreen database with blastn (v2.5.0, -task blastn -reward 1 -penalty -5 -gapopen 3 -gapextend 3 -dust yes -soft_masking true -evalue 700 -searchsp 1750000000000 -outfmt "6 std score"). The BLAST alignments did not yield any “moderate” or “strong” matches to the database, so we determined that there was no contamination.

Additional polishing and generation of the Col-CEN v1.2 assembly

To further polish the Col-CEN v1.1 assembly, we aligned all HiFi reads that were at least 16 kbp long to the Col-CEN v1.1 assembly with Winnowmap2 (v2.0, k=15 greater-than distinct=0.9998 --MD -ax map-pb) and we filtered alignments with Samtools “view” (v1.10, -F 256) (*53*, *56*). We then used “falconc bam-filter-clipped”, a part of the IPA package, to remove chimeric read alignments (-t -F 0x104) (https://github.com/PacificBiosciences/pbipa). Using these filtered alignments, we polished the Col-CEN v1.1 assembly with a special branch of Racon that outputs polishing edits in VCF format (v1.6.0, -L -u) (https://github.com/isovic/racon/tree/liftover) (*60*). Polishing edits were then filtered with Merfin, using 21-mers derived from the Col-0 Illumina reads (-peak 30) (*61*) and incorporated into the assembly with BCFtools “consensus” (*54*).

To identify and correct putative larger mis-assemblies with a second, independent method, we assembled all HiFi reads at least 16 kbp long with Hifiasm (v0.15-r327, -l0), and aligned the resulting primary contigs to the Racon polished assembly with minimap2 (v2.20-r1061, --cs -cx asm5). We called

variants with paftools "call" and manually inspected all variants larger than 1 kbp in IGV (https://github.com/lh3/minimap2/tree/master/misc) (*45*). Ultimately, two sequences were inserted into the Racon assembly, ultimately producing the Col-CEN v1.2 assembly. The Col-CEN v1.2 assembly contained five pseudomolecules, two missing telomeres, and partially resolved NOR sequence at the beginning of chromosomes 2 and 4. Chromosomes 1, 3 and 5 were completely sequence resolved from telomere-to-telomere. The final Col-CEN v1.2 assembly FASTA file includes these 5 pseudomolecules and the Columbia chloroplast and mitochondria reference genomes.

To catalog variation between Col-0 lab strains, heterozygous loci, or potential lingering misassemblies, we aligned Col-0 reads to Col-CEN v1.2 and called variants. To call small variants, we aligned all HiFi reads at least 16 kbp long to the Col-CEN v1.2 assembly with Winnowmap2 (v2.0, k=15 greater-than distinct=0.9998 --MD -ax map-pb) and called variants with DeepVariant (v1.1.0, --model_type=PACBIO). The same HiFi alignments were used to call SVs with Sniffles (v1.0.12, -d 50 -n -1 -s 10) and variants with less than 30% of reads supporting the ALT allele were removed. The same process was used to call SVs with ONT data (Winnowmap v2.0) (k=15 greater-than distinct=0.9998 --MD -ax map-ont). The resulting VCF files are available on GitHub (https://github.com/schatzlab/Col-CEN). During analysis, we uncovered two potentially misassembled loci, though plausible corrections were not apparent. We have listed these loci in an "issues" file on GitHub (https://github.com/schatzlab/Col-CEN). These, and potential future issues identified by ourselves or the community, will be considered in future assembly updates.

For assembly validation, we aligned Hi-C reads to Col-CEN with bwa mem (v0.7.17-r1198-dirty) and processed the alignments with the Arima mapping pipeline (https://github.com/ArimaGenomics/mapping_pipeline) (http://broadinstitute.github.io/picard/) (*62*). Hi-C heatmaps were made with Cooler and HiGlass (Cooler v0.8.10, 50 kbp resolution) (*63*) (https://higlass.io/).

Genome annotation

Genes were lifted-over from TAIR10 with Liftoff (v1.5.1, -copies -a 1 -s 1) (*64*). Since ChrC and ChrM were directly copied from TAIR10, their lift-over genes were replaced with their original TAIR10 annotations. We inspected every TAIR10 gene that did not lift over to provide an explanation for the discrepancy. All presence/absence variable genes are listed in **Table S2** and all missing genes (including for reasons other than genuine biological variation) are documented on GitHub (https://github.com/schatzlab/Col-CEN). We also inspected every gene that lifted over in multiple copies. All copy-number-variable genes are listed in **Table S3** and all genes that lifted over in multiple copies (including for reasons other than genuine biological variation) are listed on GitHub (https://github.com/schatzlab/Col-CEN). We used EDTA (v1.9.6, --sensitive 1 --anno 1 --evaluate 1) to perform *de novo* transposable element (TE) annotation, providing transcripts with "--cds" and the TAIR10 TE library with "--curatedlib" (*65*, *66*). The TE annotation was supplemented with a manual annotation of centromere gaps using dotplot analysis and further manual annotation of the centromeric *ATHILA* elements (see section below). We used LASTZ to identify regions with similarity to *5S*, *45S* rDNA and the mitochondrial genome. To generate similarity heatmaps, the centromere region was divided into adjacent 5 kbp regions, which were compared using the pairwiseAlignment (type='global') and pid functions in R, using the Biostrings library. Sequences were compared in forward and reverse directions, and the highest percent sequence identity value kept. These values were then plotted in the heatmap.

***CEN180* repeat annotation**

To identify repetitive regions, we divided the genome assembly into adjacent 1 kbp windows. In each window, for each position, we defined 12-mers and exactly matched these sequences to the rest of the window. We identified windows where the proportion of non-unique 12-mers was greater than 10%, and merged contiguous windows that were above this threshold. For each region, we generated a histogram of the distances between 12-mers to test for periodic repeats. For example, if a region contains an arrayed tandem repeat of monomer size N, then a histogram of the 12-mer distances will show peaks at values N, N×2, N×3 … . The N value was obtained for each region, using the most frequent 12-mer

distance. Next, 5 sequences of length N were randomly chosen from within the region and matched back to the sequence using the R function matchPattern (max.mismatch=N/3 with.indels=T). For each set of matches we identified overlapping repeats. If the overlap was less than 10 nucleotides, the overlap was divided at the midpoint between the repeats. If the overlap was 10 nucleotides or greater, the larger repeat was kept. The set of non-overlapping matches with the highest number was kept for further analysis. These sequence matches were aligned using mafft (--retree 2 --inputorder) (*67*), and a consensus repeat monomer was derived from the multiple sequence alignment. This consensus sequence was matched back to the region using matchPattern (max.mismatch=N/3 with.indels=T), and overlaps were treated in the same way.

Our approach identified 66,131 *CEN180* repeats with a mean length of 178 bp. The set of unique *CEN180* sequences (n=22,440) were aligned using mafft (--sparsescore 1000 --inputorder) (*67*). A consensus sequence was generated from the multiple sequence alignment, which was:

5′-

AGTATAAGAACTTAAACCGCAACCCGATCTTAAAAGCCTAAGTAGTGTTTCCTTGTTAGA

AGACACAAAGCCAAAGACTCATATGGACTTTGGCTACACCATGAAAGCTTTGAGAAGCA

AGAAGAAGGTTGGTTAGTGTTTTGGAGTCGAATATGACTTGATGTCATGTGTATGATTG-

3′. In order to analyze *CEN180* diversity, for each position of the multiple sequence alignment (809 positions), we calculated the proportion of A, T, G, C and gaps. The alignment for each monomer at each position was then compared to these proportions and used to calculate a variant distance for the monomer. For example, if a monomer had an A in the alignment at a given position, and the overall proportion of A at that position was 0.7, the variant distance for that monomer would increase by 1-0.7. This was repeated for each position of the alignment, for each monomer. This ‘weighted’ variant distance was used to assess how similar a given *CEN180* monomer is to the genome-wide consensus. Alternatively, to compare pairwise differences between two specific monomers, the two sequences were compared along the length of the multiple sequence alignment and each instance of disagreement counted to give a ‘pairwise’ variant score.

To identify higher order repeats (HORs) in a head-to-tail (tandem) orientation, each monomer was taken in turn and compared to all others using a matrix of pairwise variant scores. If a pair of monomers had a variant score of 5 or less, and were on the same strand, they were considered a match. For each match, monomers were extended by +1 unit in the same direction on the chromosome, and these were again compared for pairwise variants. This process was repeated until the next monomers had a pairwise variant score higher than threshold, or the repeats were on opposite strands, or the end of the array was reached, with these conditions defining the end of the HOR. We also searched for repeats in head-to-head (inverted) orientation, which was identical apart from that repeats must be on opposite strands, and when monomers are extended to search for HORs, one is extended +1 position along the chromosome, whereas the other decreases -1. HORs were defined for each instance of 2 or more consecutive monomer matches. We define each HOR as consisting of block1 and block2 of *CEN180* monomers. The size of each block was recorded, in terms of monomers and base pairs, in addition to the distance between the block start coordinates. Cumulative pairwise variants per *CEN180* monomer were also calculated between each pair of blocks to provide a 'block' variant score. To measure higher order repetition of each monomer, we summed the HOR block sizes in mers, such that if a monomer was represented in three 5-mer blocks, it would score 15.

ATHILA annotation

To resolve the sequence of the centromeric *ATHILA* elements, we used LTRharvest (*68*) to complement the EDTA run that was used for the annotation of all Arabidopsis TEs (see above). We ran LTRharvest three times using 'normal', 'strict' and 'very strict' parameters. The parameters were gradually adjusted to allow us to capture the full-length sequence of the *ATHILA* subfamilies, based on older studies that reported the total and LTR lengths of intact *ATHILA* elements (*18*). These parameters were -maxlenltr 2500 -minltrlen 400 -mindistltr 2000 -maxdistltr 20000 -similar 75 -mintsd 0 -motif TGCA -motifmis 1 for the 'normal' run; -maxlenltr 2000 -minlenltr 1000 -mindistltr 4000 -maxdistltr 16000 -similar 80 -mintsd 3 -motif TGCA -motifmis 1 for the 'strict' run; and -maxlenltr 2100 -minlenltr 1100 -mindistltr 5000 -maxdistltr 14000 -similar 85 -mintsd 4 -motif TGCA -motifmis 1 -vic 20 for the 'very strict' run. Coordinates of predicted intact elements from EDTA, LTRharvest and the manual dotplot annotation

of centromeric TEs were merged and sequences aligned using mafft (*69*). Through these steps, we were able to pinpoint with base-pair resolution the external junctions of every *ATHILA* element, and the internal junctions of the LTRs with the internal domain (5′-LTR with PBS; PPT with 3′-LTR). Overall, we identified 111 intact elements, 53 inside and 58 outside of the centromeres, of which 43 (81%) and 40 (69%) respectively have a detectable target site duplication (TSD), 20 fragmented *ATHILA* and 12 solo LTRs (10 with a TSD, 83%) (**Table S6**). We further identified open reading frames (minimum 300 bp) in the internal domain of the intact elements using getorf in EMBOSS (*70*), and the core domains of the *gag* and *pol* genes by running HMMER v3.3.2 (http://hmmer.org/) (-E 0.001 --domE 0.001) and using a collection of Hidden Markov Models (HMMs) downloaded from Pfam (http://pfam.xfam.org/) that describe the genes of *GYPSY* LTR retrotransposons: PF03732 for gag; PF13650, PF08284, PF13975 and PF09668 for protease; PF00078 for reverse transcriptase; PF17917, PF17919 and PF13456 for RNase-H; PF00665, PF13683, PF17921, PF02022, PF09337 and PF00552 for integrase; PF03078 for an *ATHILA*-specific domain. Given that many *ATHILA* subfamilies do not appear to contain the core domains of reverse transcriptase, RNase-H and integrase (Table S4), as these are described by the Pfam models, we used the full-length sequence of the intact elements to examine their phylogenetic relationships. The multiple alignment file was produced using mafft with the G-INS-i parameter (*69*), and FastTree (-nt) to generate the maximum likelihood tree (*71*). The tree was visualized and annotated with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

ONT DNA methylation analysis

To identify CG, CHG and CHH methylation contexts we used DeepSignal-plant (v. 0.1) (*20*), which uses a deep-learning method based on bidirectional recurrent neural network (BRNN) with long short-term memory (LSTM) units to detect DNA 5mC methylation. R9 reads were filtered for length and accuracy using Filtlong (v0.2.0) (--min_mean_q 90, --min_length 30000. https://github.com/rrwick/Filtlong). Basecalled read sequence was annotated onto corresponding .fast5 files, and re-squiggled using Tombo (v 1.5.1). Methylation prediction for the CG, CHG, and CHH contexts were called using DeepSignal-plant using the respective models:

model.dp2.CG.arabnrice2-1_R9.4plus_tem.bn13_sn16.balance.both_bilstm.b13_s16_epoch6.ckpt, model.dp2.CHG.arabnrice2-1_R9.4plus_tem.bn13_sn16.denoise_sig1nal_bilstm.both_bilstm.b13_s16_epoch4.ckpt model.dp2.CHH.arabnrice2-1_R9.4plus_tem.bn13_sn16.denoise_signal_bilstm.both_bilstm.b13_s16_epoch7.ckpt. The script call_modification_frequency.py provided in the DeepSignal-plant package was then used to generate the methylation frequency at each CG, CHG and CHH site.

To identify CG methylation in Nanopore reads we also used Nanopolish (v 0.13.2), which uses a Hidden Markov model on the nanopore current signal to distinguish 5-methylcytosine from unmethylated cytosine. Reads were first filtered for length and accuracy using Filtlong (v0.2.0) (--min_mean_q 95, --min_length 15000. https://github.com/rrwick/Filtlong). The subset was then indexed to the fast5 files, and aligned to the genome using Winnowmap (v1.11, -ax map-ont). The read fastq, alignment bam, and fast5 files were used as an input to the Nanopolish call-methylation function. The script calculate_methylation_frequency.py provided in the Nanopolish package was then used to generate the methylation frequency at each CG containing *k*-mer.

Bionano optical mapping

DNA was extracted following Bionano's Plant DNA Isolation Kit (#80003) and protocol. Isolated DNA was labeled with Bionano's Direct Label and Stain Kit (DLS #80005) and samples were run on a Saphyr chip and analyzed with BionanoAccess software v1.6, Bionano Tools v1.6 and Bionano Solve v3.6_09252020. Data generation reached 2,290 Gb equating to roughly 1,523× coverage after quality filtering for molecules containing at least 10 labels per molecule (read). *De novo* assembly of the Bionano data was performed with default assembly settings resulting in 19 contigs for a total assembly length of 132.961 Mbp. Further comparison of the Bionano contig maps was made with the Col-CEN v1.2 genome assembly. Bionano maps and molecules support the Col-CEN genome assembly where Bionano maps are capable of alignment. However, due to a lack of labelling sites, the centromere sequences generally result in breakage of the Bionano maps.

Chromatin immunoprecipitation and sequencing (ChIP-seq)

Approximately 12 grams of 2 week old Col-0 seedlings were ground in liquid nitrogen. Nuclei were isolated in nuclei isolation buffer (1 M sucrose, 60 mM HEPES pH 8.0, 0.6% Triton X-100, 5 mM KCl, 5 mM $MgCl_2$, 5 mM EDTA, 0.4 mM PMSF, 1 mM pepstatin-A, 1×protease inhibitor cocktail), and crosslinked in 1% formaldehyde at room temperature for 25 minutes. The crosslinking reaction was quenched with 125 mM glycine and incubated at room temperature for a further 25 minutes. The nuclei were purified from cellular debris via two rounds of filtration through one layer of Miracloth and centrifuged at 2,500*g* for 25 minutes at 4 °C. The nuclei pellet was resuspended in EB2 buffer (0.25 M sucrose, 1% Triton X-100, 10 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 1 mM EDTA, 5 mM DTT, 0.1 mM PMSF, 1 mM pepstatin-A, 1×protease inhibitor cocktail) and centrifuged at 14,000*g* for 10 minutes at 4 °C.

The nuclei pellet was resuspended in lysis buffer (50 mM Tris-HCl pH 8.0, 1% SDS, 10 mM EDTA, 0.1 mM PMSF, 1 mM pepstatin-A) and chromatin was sonicated using a Covaris E220 evolution with the following settings: power=150V, bursts per cycle=200, duty factor=20%, time=60 seconds. Sonicated chromatin was centrifuged at 14,000g and the supernatant was extracted and diluted with 1×volume of ChIP dilution buffer (1.1% Triton X-100, 20 mM Tris-HCl pH 8.0, 167 mM NaCl, 1.1 mM EDTA, 1mM pepstatin-A, 1×protease inhibitor cocktail). The chromatin was incubated overnight at 4 °C with 50µl Protein A magnetic beads (Dynabeads, Thermo Fisher) pre-bound with either 5µl α-CENH3 (*12*), or α-H3K9me2 antibody (mAbcam 1220). The beads were collected on a magnetic rack and washed twice with low-salt wash buffer (150 mM NaCl, 0.1% SDS, 1% Triton X-100, 20 mM Tris-HCl pH 8.0, 2 mM EDTA, 0.4 mM PMSF, 1 mM pepstatin-A, 1×protease inhibitor cocktail) and twice with high-salt wash buffer (500 mM NaCl, 0.1% SDS, 1% Triton X-100, 20 mM Tris-HCl pH 8.0, 2 mM EDTA, 0.4 mM PMSF, 1 mM pepstatin-A, 1×protease inhibitor cocktail). Immunoprecipitated DNA–protein complexes were eluted from the beads (1% SDS, 0.1 M $NaHCO_3$) at 65°C for 15 minutes. Samples were reverse crosslinked by incubating with 0.24 M NaCl at 65°C overnight. Proteins and RNA were digested with Proteinase K treatment, and RNase A, and DNA was purified with phenol:chloroform:isoamyl alcohol (25:24:1) extraction and ethanol precipitation. Library preparation

followed the Nanopore SQK-LSK109 protocol and kit (as above) and sequenced on separate flongle flowcells.

Per-read DNA methylation analysis following CENH3 and H3K9me2 ChIP and ONT sequencing

The resulting .fast5 files were basecalled with Guppy (v5.0.11+2b6dbffa5), using the dna_r9.4.1_450bps_sup.cfg and aligned to the Col-CEN reference with Winnowmap (v1.11, k=15, --MD -ax map-ont). Reads overlapping centromeric positions (Chr1: 14840000-17560000, Chr2: 3823000-6046000, Chr3: 13597000-15734000, Chr4: 4204000-6978000, Chr5: 11784000-1456000) were extracted, providing a set of 5,130 and 11,150 CENH3- or H3K9me2-associated centromeric reads, respectively. The methylation predictions for CG, CHG and CHH methylation contexts were extracted using DeepSignal-plant (v0.1) (*20*) within these read sets. The resulting .tsv files were filtered to remove ambiguous calls (prob_cf=0.5) and used to calculate the mean methylation state of each context, across individual reads within both data sets. These values were then plotted in R version 4.0.0.

ChIP-seq and MNase-seq data alignment and processing

Deduplicated paired-end ChIP-seq and MNase-seq Illumina reads (**Table S7**) were processed with Cutadapt v1.18 to remove adapter sequences and low-quality bases (Phred+33-scaled quality <20) (*72*). Trimmed reads were aligned to the Col-CEN genome assembly using Bowtie2 v2.3.4.3 with the following settings: --very-sensitive --no-mixed --no-discordant -k 10 (*73*). Up to 10 valid alignments were reported for each read pair. Read pairs with Bowtie2-assigned MAPQ <10 were discarded using Samtools v1.9 (*53*). For retained read pairs that aligned to multiple locations, with varying alignment scores, the best alignment was selected. Alignments with more than 2 mismatches or consisting of only one read in a pair were discarded. Single-end SPO11-1-oligo reads were processed and aligned to the Col-CEN assembly using an equivalent pipeline without paired-end options, as described (*28*). For each data set, bins per million mapped reads (BPM; equivalent to transcripts per million, TPM, for RNA-seq data) coverage values were generated in bigWig and bedGraph formats with the bamCoverage tool from deepTools v3.1.3 (*74*). Reads that aligned to chloroplast or mitochondrial DNA were excluded from this coverage normalization procedure.

RNA-seq data alignment and processing

Paired-end RNA-seq Illumina reads (2×100 bp) (**Table S7**) (*29*) were processed with Trimmomatic v0.38 to remove adapter sequences and low-quality bases (Phred+33-scaled quality <3 at the beginning and end of each read, and average quality <15 in 4-base sliding windows) (*28*, *75*). Trimmed reads were aligned to the Col-CEN genome assembly using STAR v2.7.0d with the following settings: --outFilterMultimapNmax 100 --winAnchorMultimapNmax 100 --outMultimapperOrder Random --outFilterMismatchNmax 2 --outSAMattributes All --twopassMode Basic --twopass1readsN -1 (*76*). Read pairs with STAR-assigned MAPQ <3 were discarded using Samtools v1.9 (*53*). For retained read pairs that aligned to multiple locations, with varying alignment scores, the best alignment was selected. Alignments with more than 2 mismatches, or consisting of only one read in a pair, were discarded.

Small RNA-seq data alignment and processing

Small RNA-seq Illumina reads (**Table S7**) (*29*) were processed with BBDuk from BBMap v38.22 (*77*) to remove ribosomal sequences, and with Cutadapt v1.18 (*72*) to remove adapter sequences and low-quality bases (Phred+33-scaled quality <20). Trimmed reads were aligned to the Col-CEN genome assembly using Bowtie v1.2.2, allowing no mismatches (*57*). For reads that aligned to multiple locations, with varying alignment scores, the best alignment was selected. For each small RNA size class (18–26 nucleotides), TPM values in adjacent genomic windows were calculated based on the total retained alignments (across all size classes) in the library.

Bisulfite sequencing data alignment and processing

Paired-end bisulfite sequencing Illumina reads (2×90 bp) (**Table S7**) (*29*) were processed with Trim Galore v0.6.4 to remove sequencing adapters, low-quality bases (Phred+33-scaled quality <20) and 3 bases from the 5′ end of each read (*78*). Trimmed reads were aligned to the Col-CEN assembly using Bismark v0.20.0 (*79*). Read pairs that aligned equally well to more than one location and duplicate alignments were discarded. Methylated cytosine calls in CG, CHG and CHH sequence contexts were extracted and context-specific DNA methylation proportions were generated in bedGraph and bigWig formats using the bismark2bedGraph and UCSC bedGraphToBigWig tools. DNA methylation

proportions for cytosines covered by <6 reads were excluded. Single-end bisulfite sequencing reads (50 bp) (**Table S7**) (*21*, *22*) were processed and aligned to the Col-CEN assembly using an equivalent pipeline without paired-end options.

Fine-scale profiling around feature sets

Fine-scale profiles around *CEN180* (n=66,131), randomly positioned loci of the same number and width distribution (n=66,131), centromeric intact *ATHILA* elements (n=53), *ATHILA* elements located outside the centromeres (n=58), and *GYPSY* retrotransposons (n=3,979) were calculated for ChIP-seq, RNA-seq, small RNA-seq and bisulfite-seq data sets by providing the above-described bigWig files to the computeMatrix tool from deepTools v3.1.3 in 'scale-regions' mode (*74*). Each feature was divided into non-overlapping, proportionally scaled windows between start and end coordinates, and flanking regions were divided into 10-bp windows. Mean values for each data set were calculated within each window, generating a matrix of profiles in which each row represents a feature with flanking regions and each column a window. Coverage profiles for a ChIP input sequencing library and a gDNA library (**Table S7**) were used in conjunction with those for ChIP-seq and SPO11-1-oligo libraries, respectively, to calculate windowed $\log_2$([ChIP+1]/[control+1]) coverage ratios for each feature. Meta-profiles (windowed means and 95% confidence intervals) for each group of features were calculated and plotted using the feature profiles in R version 4.0.0.

Crossover mapping

Total data from 96 Col×Ler genomic DNA F_2 sequencing libraries (2×150 bp) were aligned to the Col-CEN assembly using bowtie2 (default settings). Polymorphisms were identified using the alignment files with samtools mpileup (-vu -f) and bcftools call (-mv -Oz). The resulting polymorphisms were filtered for SNPs (n=522,112), which was used as the 'complete' polymorphism set in TIGER. These SNPs were additionally filtered by, (i) removing SNPs with a quality score less than 200, (ii) removing SNPs where total coverage was greater than 300, or less than 50, (iii) removing SNPs that had reference allele coverage less than 20 or greater than 150, (iv) removing SNPs that had variant allele coverage greater than 130, (v) masking SNPs that overlapped transposon and repeat annotations and (vi) masking

SNPs within the main *CEN180* arrays. This resulted in a 'filtered' set of 248,695 SNPs for use in TIGER. DNA sequencing data from 260 wild type Col×Ler F_2 genomic DNA (192 from ArrayExpress E-MTAB-4657 and 68 from E-MTAB-6577) was aligned to the Col-CEN assembly using bowtie2 (default settings) and the alignment analyzed at the previously defined 'complete' SNPs using samtools mpileup (-vu -f) and bcftools call (-m -T). These sites were used as an input to TIGER, which identifies crossover positions by genotype transitions (*80*). A total of 2,080 crossovers were identified with a mean resolution of 1,047 bp.

Epitope tagging of *V5-DMC1*

The *DMC1* promoter region was PCR amplified from Col-0 genomic DNA using the Dmc1-PstI-fw and Dmc1-SphI-rev oligonucleotides. The remainder of the *DMC1* promoter, gene and terminator were amplified with oligonucleotides Dmc1-SphI-fw and Dmc1-NotI-rev. The resulting PCR fragments were digested with *Pst*I and *Sph*I, or *Sph*I and *Not*I, respectively, and cloned into *Pst*I-*Not*I-digested pGreen0029 vector to yield a pGreen-DMC1 construct. To insert 3 N-terminal V5 epitope tags, first two fragments were amplified with DMC1-Nco-F and 3N-V5-R and 3N-V5-F and Dmc1-Spe-rev and then used in an overlap PCR reaction using the DMC1-Nco-F and Dmc1-Spe-rev oligonucleotides. The PCR product resulting from the overlap PCR was digested with *Nco*I and *Spe*I and cloned into *Nco*I- and *Spe*I-digested pGreen-DMC1. The resulting binary vector was used to transform *dmc1-3/+* heterozygotes (SAIL_126_F07). We used dmc1-seq11 and Dmc1-Spe-rev oligonucleotides to amplify wild type *DMC1* allele and Dmc1-Spe-rev and LA27 to amplify the *dmc1-3* T-DNA mutant allele. The presence of the *V5-DMC1* transgene was detected with N-screen-F and N-screen-R oligonucleotides. This oligonucleotide pair amplifies a 74 bp product in Col and a 203 bp product in *V5-DMC1*. To identify *dmc1-3* homozygotes in the presence of *V5-DMC1* transgenes, we used DMC1-genot-compl-F and DMC1-genot-compl-R oligonucleotides, which allowed us to distinguish between the wild type *DMC1* gene and *V5-DMC1* transgene. All oligonucleotide sequences are provided in **Table S8**.

Cytogenetic and immunocytological analyses

For fluorescence *in situ* hybridization (FISH), spreads of meiotic chromosomes at pachytene stage of meiosis were prepared from young flower buds fixed in ethanol:acetic acid (3:1) and stored in 70% ethanol until use. Chromosome spreads were prepared as described (*81*). To identify individual chromosome arms, chromosome-specific *A. thaliana* BAC clones were arranged into contigs. More specifically, the following BAC contigs were used: five (F10C21/AC051630 – F12K21/AC023279; **Fig. 1D**, **1F**, **S5A** and **S5D**), 15 (F13M18/AL087094 – F12K21/AC023279; **Fig. S5**C and **S5E**) or 29 (F6F9/AC007797 – F12K21/AC023279; **Fig. S5B**) chromosome 1 upper-arm-specific BACs; five (F2J6/AC009526 – T2P3/B21868; **Fig. 1D** and **S5A**) or 36 (F2J6/AC009526 – T6H22/AC009894; **Fig. S5B**) chromosome 1 bottom-arm-specific BACs; five (T21B4/AF007271 – T8M17/AF296835; **Fig. S5A**) or 29 (T20O7/AB026660 – T8M17/AF296835; **Fig. 3H** and **S5E**) chromosome 5 upper-arm-specific BACs; five (F5M8/AL082902 – T31G3/AB026662; **Fig. S5A**) chromosome 5 bottom-arm-specific BACs. The Arabidopsis $(TTTAGGG)_n$ telomere repeat probe was prepared by PCR, as described (*82*). All DNA probes were labeled with biotin-dUTP, digoxigenin-dUTP, or Cy3-dUTP by nick translation, then pooled, ethanol-precipitated and pippeted on pepsin-treated and ethanol-dehydrated slides containing suitable chromosome spreads. The slides were heated to 80°C for 2 minutes and incubated at 37°C for 12 hours. The hapten-labeled probes were immuno-detected as described (*81*). BAC contigs and other DNA probes were visualised using fluorescently labeled antibodies against biotin-dUTP (avidin-Texas red, Vector Laboratories, cat. no. A-2006-5) and digoxigenin-dUTP (mouse anti-digoxigenin, Jackson ImmunoResearch, 200-002-156, goat anti-mouse Alexa Fluor 488, Invitrogen, A11001, and goat anti-mouse Alexa Fluor 647, Invitrogen, A21235). Chromosomes were counterstained with DAPI (2 μg/mL) in Vectashield (Vector Laboratories). Fluorescence signals were analyzed and imaged using a Zeiss AxioImager epifluorescence microscope (Carl Zeiss) with a CoolCube camera (MetaSystems). Images were acquired separately using the Isis software (MetaSystems) for all four fluorochromes using appropriate excitation and emission filters (AHF Analysentechnik). The four monochromatic images were pseudocoloured, merged, and cropped using Photoshop CS (Adobe Systems), and chromosome length was measured using ImageJ (National Institutes of Health).

The *CEN180* pAL FISH probe, which labels all centromeres, was amplified using primers ATH_cen180F and ATH_cen180R (**Table S8**) (*83*). PCR amplification was performed as follows: initial denaturation at 95°C for 5 minutes; 35 cycles of denaturation at 95°C for 20 seconds, annealing at 46°C for 20 seconds and extension at 72°C for 20 seconds; and a final extension at 72°C for 5 minutes. PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) and labelled by nick translation. To design *CEN180* probes specific to individual chromosomes or sets of chromosomes, *CEN180* sequences identified in the Col-CEN assembly were aligned using MAFFT (v7.450) and used to identify repeats with high copy number and distributions biased to specific chromosomes. Oligonucleotide FISH probes homologous to specific *CEN180* sequences were designed that were 60 nucleotides in length, with a GC content between 30-50% and selected to minimize self-annealing and formation of hairpin structures, using Geneious (v11.1.5) (**Table S8**). Double-stranded DNA probes were prepared and labelled, as described (*81*).

To design FISH probes against *ATHILA* transposons the sequences encoding the highly variable GAG domains for each sub-family were aligned using MAFFT (v7.450) and consensus sequences were generated. PCR primers were then designed to amplify subfamily GAG domain genes, using Primer3 (v2.3.7) implemented in Geneious (**Table S8**). PCR amplification was performed as follows: initial denaturation at 95°C for 5 minutes; 35 cycles of denaturation at 95°C for 20 seconds, annealing at 58°C for 20 seconds and extension at 72°C for 20 seconds; and a final extension at 72°C for 5 minutes. PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit and subsequently cloned into the pGEM-T Easy Vector System (Promega), using TOP10 competent cells. Positive colonies were screened using SP6/T7 primers and five clones of each *ATHILA-GAG* gene were Sanger sequenced. Subsequently, clones with the highest pairwise sequence similarity to specific *ATHILA* sub-family consensus sequences were used as templates for PCR amplification. Purified PCR products were labelled by nick translation, as described (*81*).

For analysis of chromatin during mitotic interphase, nuclei were isolated from 1 week old seedlings (wild type Col-0 and CENH3-GFP (*84*)) and treated as described (*24*). Primary antibodies were diluted 1:200 while the secondary antibodies Alexa488 and Alexa555 goat anti rabbit or goat anti mouse

conjugates (Molecular Probes) were diluted 1:500. The primary antibodies used were anti-GFP (mouse, Roche 11814460001), anti-H3K4me1 (rabbit, Abcam Ab8895), anti-H3K4me3 (rabbit, Abcam Ab8580), anti-H3K9me1 (rabbit, Abcam Ab8896), anti-H3K9me2 (mouse, Abcam Ab1220), anti-H3K27me1 (rabbit, Abcam Ab194688), anti-H3K27me3 (rabbit, Sigma Aldrich 07-449) and anti-K36me3 (rabbit, Abcam Ab9050). To visualize DNA, nuclei were mounted in Vectashield containing DAPI. Images were acquired with the LSM980 Axio Observer with the Airyscan2 detector from Zeiss. Images were Airscan processed using the Zen Black software. Images were further analyzed using Fiji software. To correct for 3D shifts between channels in the Z plane, differences between the channels were estimated by imaging fluorescent beads. The channels were then aligned to correct for this shift. Areas of interest were resliced in Image J to obtain line plots. Intensity plots were then made using the ggplot2 package in R 3.5.1.

To immunocytologically analyse meiosis, fresh buds at floral stage 8 and 9 were dissected to release the anthers that contain male meiocytes (*85*). Chromosome spreads of meiotic and mitotic cells from anthers were performed, followed by immunofluorescent staining of proteins as described (*26*). The antibodies used in this study were: α-ZYP1 (rabbit, 1/500 dilution) (*86*), α-H3K9me2 (mouse, 1/200 dilution) (Abcam, ab1220), α-CENH3 (rabbit, 1/100 dilution) (Abcam, ab72001) and α-V5 (chicken, 1/200 dilution) (Abcam, ab9113). Chromosomes stained with ZYP1, CENH3 and H3K9me2 were visualized with a DeltaVision Personal DV microscope (Applied Precision/GE Healthcare). Chromosomes stained with DMC1-V5 and CENH3 were visualized with a Leica SP8 confocal microscope. Chromosomes stained with H3K9me2 were visualized with a Stimulated emission depletion nanoscopy mounted on an inverted IX71 Olympus microscope.

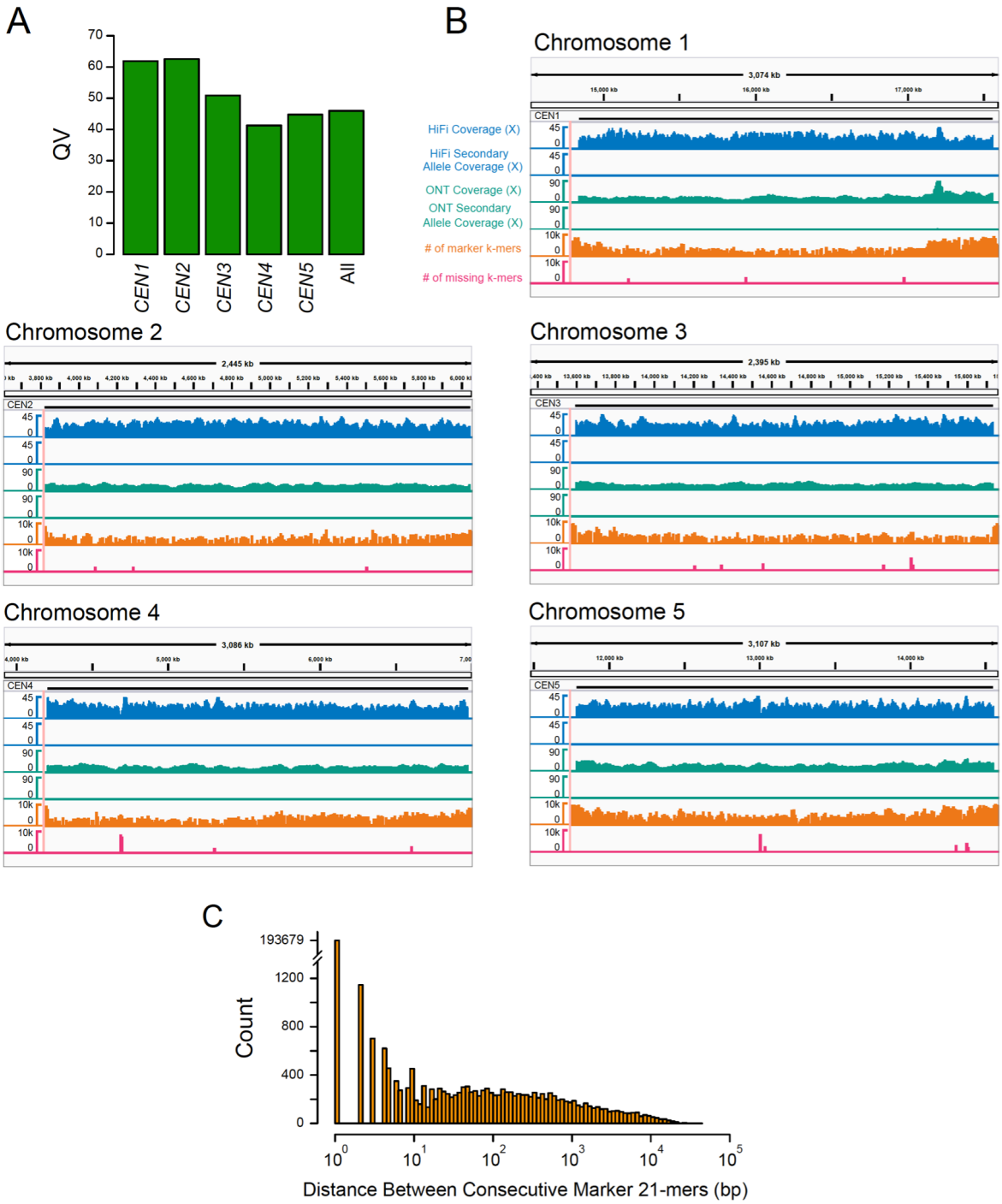


Figure S1. Validation of the Col-CEN centromere assembly. A. Assembly consensus quality (QV) scores of the individual and collective (All) centromeres. **B.** IGV screenshots depicting quantitative tracks across the five centromeres. All coverage tracks are binned via averaging, whereas the marker and missing k-mer tracks are aggregated in 10 kbp windows with no IGV binning. Secondary Allele Coverage tracks depict the coverage of the most covered alternate sequence (if any) indicated by the

alignments at every position. The “marker” and “missing” k-mer tracks are plotted with a y-axis log scale. **C**. Distribution of distances (bp) between consecutive marker 21-mers.

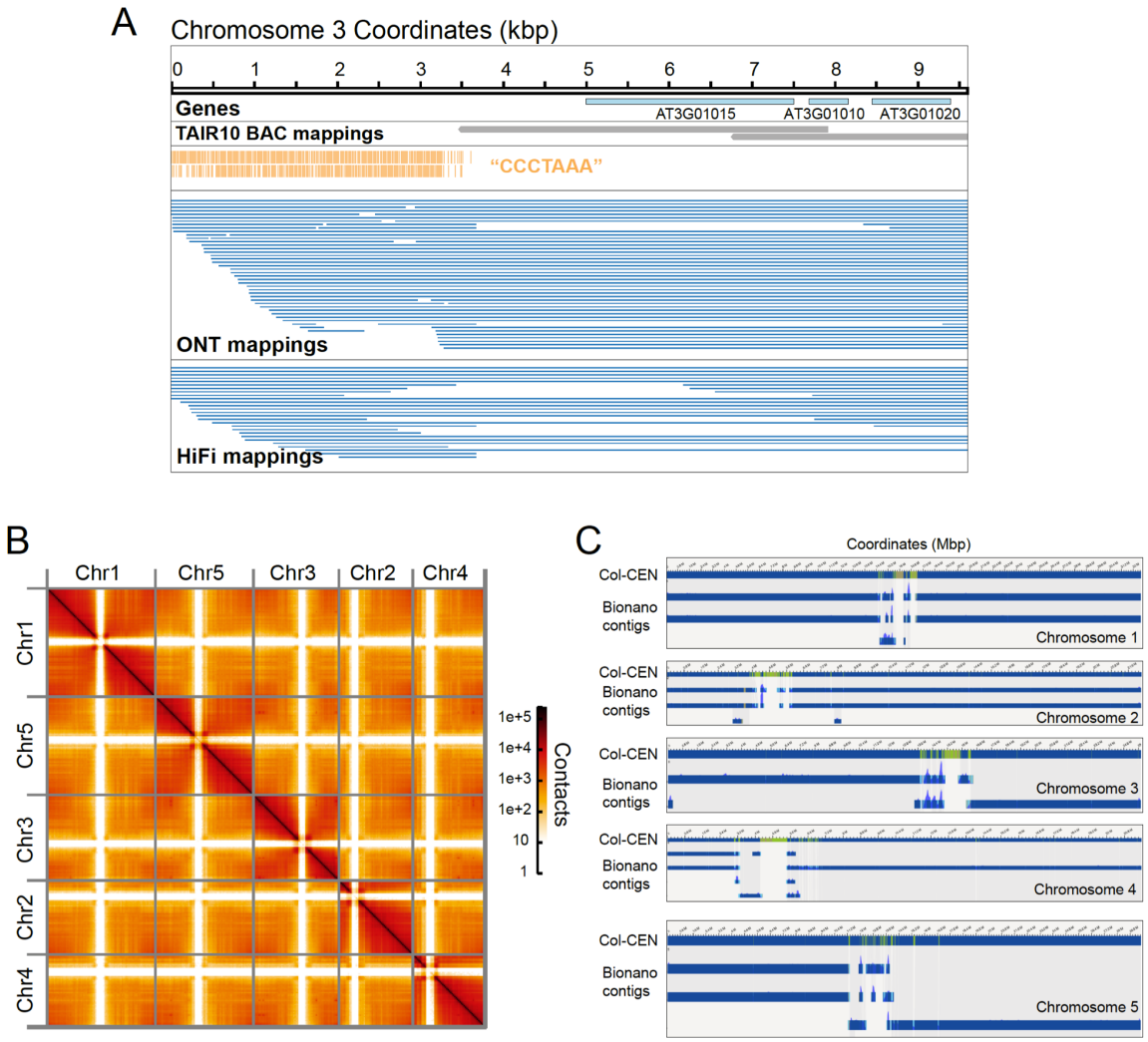


Figure S2. Telomere assembly and validation of the Col-CEN assembly using Bionano and Hi-C data. A. IGV screenshot showing the start of Col-CEN chromosome 3, including the assembled telomere. Gene models and mapped TAIR10 BACs are indicated, in addition to matches to the telomeric repeat (orange, CCCTAAA). Also shown in blue are ONT and HiFi read mappings to the Col-CEN assembly. **B.** A Hi-C heatmap generated by aligning Col-0 Hi-C reads to the Col-CEN assembly (*90*). **C.** Bionano *de novo* assembly contigs were mapped to the Col-CEN reference assembly. The green and blue bars represent the expected labeling positions in the ONT reference assembly, where blue bars are expected labeling positions, green regions lack Bionano labels and light brown bars represent predicted labeling positions not linked to a Bionano optical contig. Centromere regions generally lack predicted labeling sequences and therefore Bionano *de novo* assembled contigs are broken.

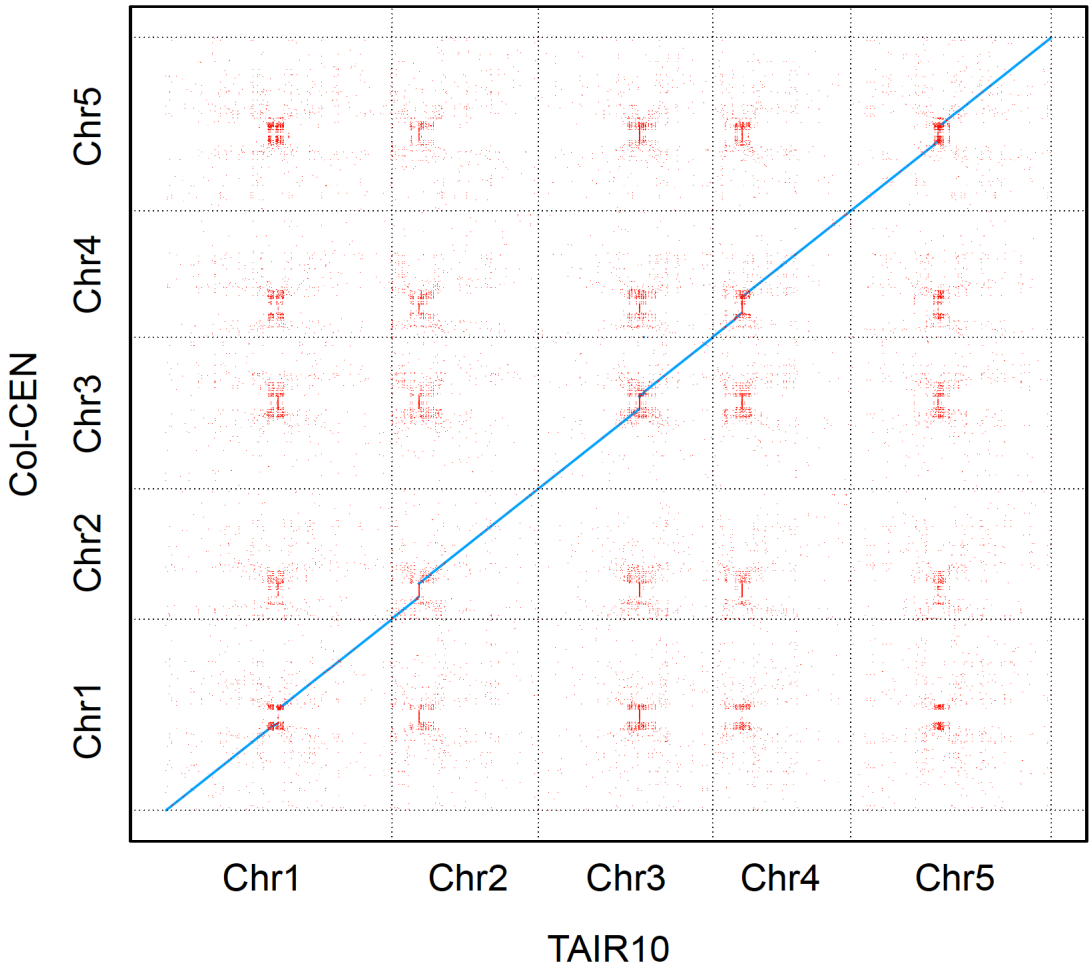


Figure S3. Dotplot sequence similarity comparison of TAIR10 and the Col-CEN genome assembly. A dotplot depicting unique (blue) and repetitive (red) Nucmer alignments (--maxmatch -l 50 -c 250) between TAIR10 and Col-CEN.

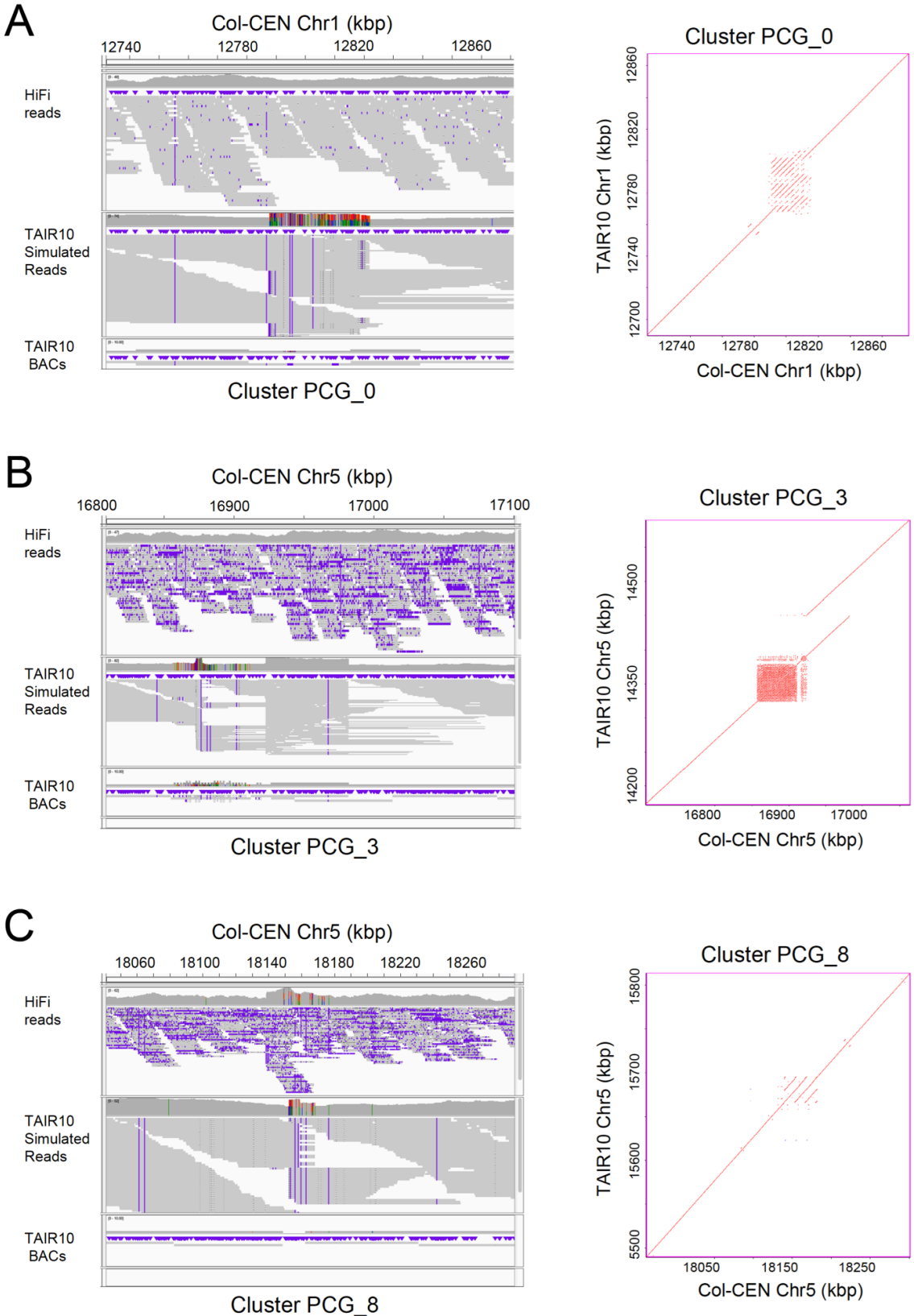

A
Col-CEN Chr1 (kbp)
12740
12780
12820
12860
HiFi reads
TAIR10 Simulated Reads
TAIR10 BACs
Cluster PCG_0
Cluster PCG_0
TAIR10 Chr1 (kbp)
12700
12740
12780
12820
12860
Col-CEN Chr1 (kbp)
B
Col-CEN Chr5 (kbp)
16800
16900
17000
17100
HiFi reads
TAIR10 Simulated Reads
TAIR10 BACs
Cluster PCG_3
Cluster PCG_3
TAIR10 Chr5 (kbp)
14200
14350
14500
16800
16900
17000
Col-CEN Chr5 (kbp)
C
Col-CEN Chr5 (kbp)
18060
18100
18140
18180
18220
18260
HiFi reads
TAIR10 Simulated Reads
TAIR10 BACs
Cluster PCG_8
Cluster PCG_8
TAIR10 Chr5 (kbp)
5500
15600
15700
15800
18050
18150
18250
Col-CEN Chr5 (kbp)

Figure S4. Genic copy number variation loci between the TAIR10 and Col-CEN assemblies. A. On the left is an IGV screenshot showing a region of chromosome 1 from the Col-CEN assembly that contains a thionin gene cluster that shows a deletion relative to TAIR10 with 4 genes that did not map to Col-CEN (Cluster PCG_0, see **Table S2**). The screenshot shows alignment of PacBio Hifi reads (upper track). Below, 100 kbp exact WGS reads were simulated from TAIR10 and their alignments are shown (middle track). Finally, TAIR10 BAC contig alignments are shown (lower). Purple marks indicate insertions and additional colors in the coverage tracks indicate substitutions. Uneven TAIR10 simulated read and BAC contig coverage indicates a structural difference between TAIR10 and Col-CEN at this locus, yet uniform HiFi coverage supports Col-CEN assembly accuracy, suggesting that this discrepancy is due to genuine biological variation, rather than misassembly. To the right a dotplot of the PCG_0 cluster in Col-CEN versus TAIR10 is shown. **B.** As for A., but showing Cluster PCG_3 on chromosome 5, where 8 TAIR10 genes did not map to Col-CEN (see **Table 2**). **C.** As for A., but showing Cluster PCG_8 on chromosome 5, where 3 TAIR10 genes mapped with an extra copy to Col-CEN (see **Table 3**).

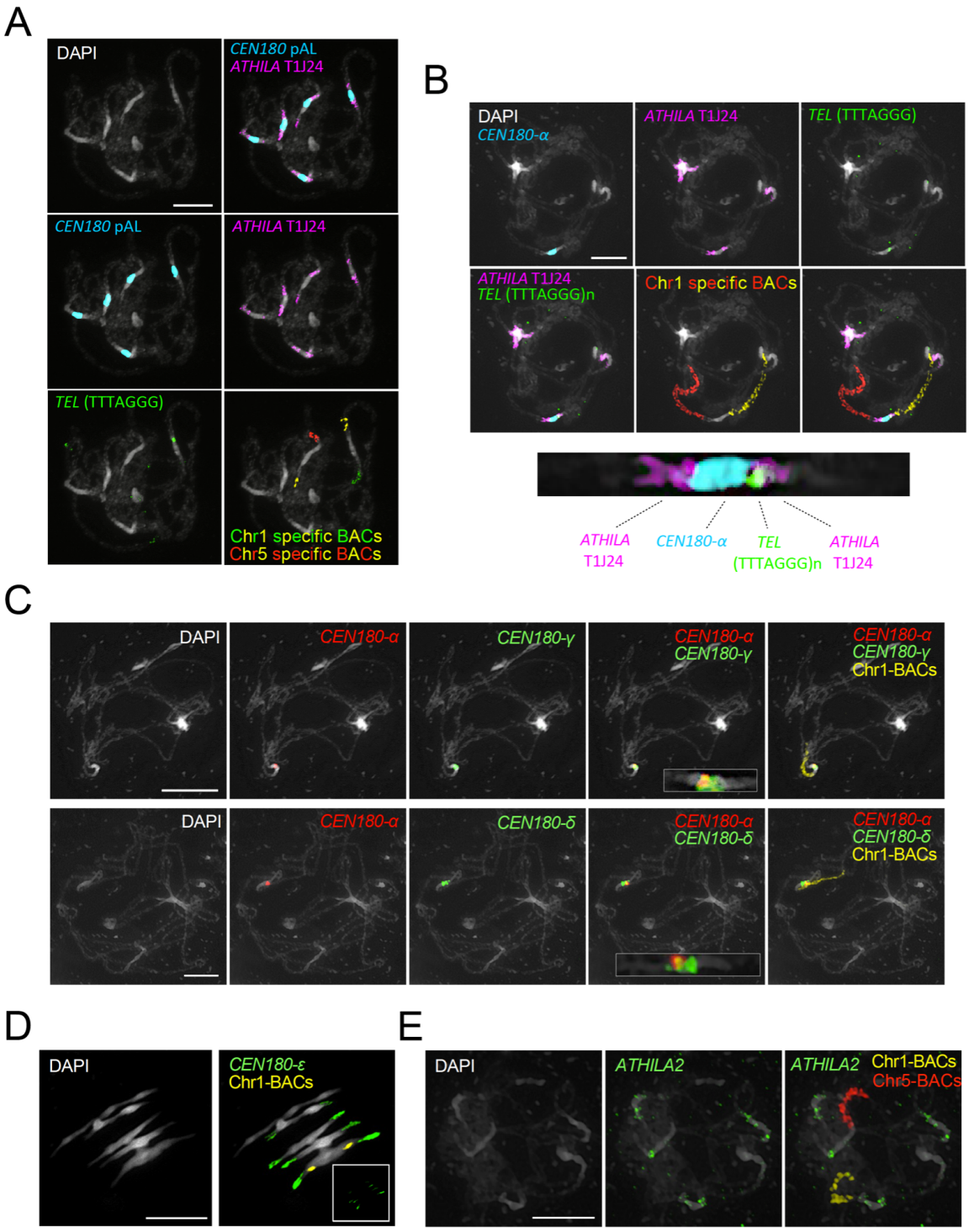


Figure S5. Fluorescent *in situ* hybridization (FISH) analysis of the Arabidopsis centromeres. A. Pachytene-stage meiotic chromosomes were spread and stained with DAPI (white), and FISH performed using probes designed to label all *CEN180* (blue, pAL), pericentromeric *ATHILA* (purple, BAC T1J24), the telomeric repeat (green, *TEL* $(TTTAGGG)_n$), chromosome 1 specific BACs (yellow and green) and chromosome 5 specific BACs (red and yellow). The scale bar represents 10 μM. **B.** As

for A., apart from the *CEN180-α* probe (blue) was used for FISH, together with chromosome 1 specific BACs labelled in red and yellow. A blow-up of centromere 1 is shown beneath. **C.** As for A., but labelling with the *CEN180-α* (red), *CEN180-γ* (green) and *CEN180-δ* (green) FISH probes, together with chromosome 1 specific BACs (yellow). Blow-ups of the centromere 1 region are shown inset. **D.** A cell dividing at metaphase I of meiosis is shown that was stained by DAPI (white), and the *CEN180-ε* FISH probe (green). **E.** As for A, but labelling with an *ATHILA2* subfamily specific *GAG* probe (green) and chromosome 1 (yellow) and 5 (red) specific BACs.

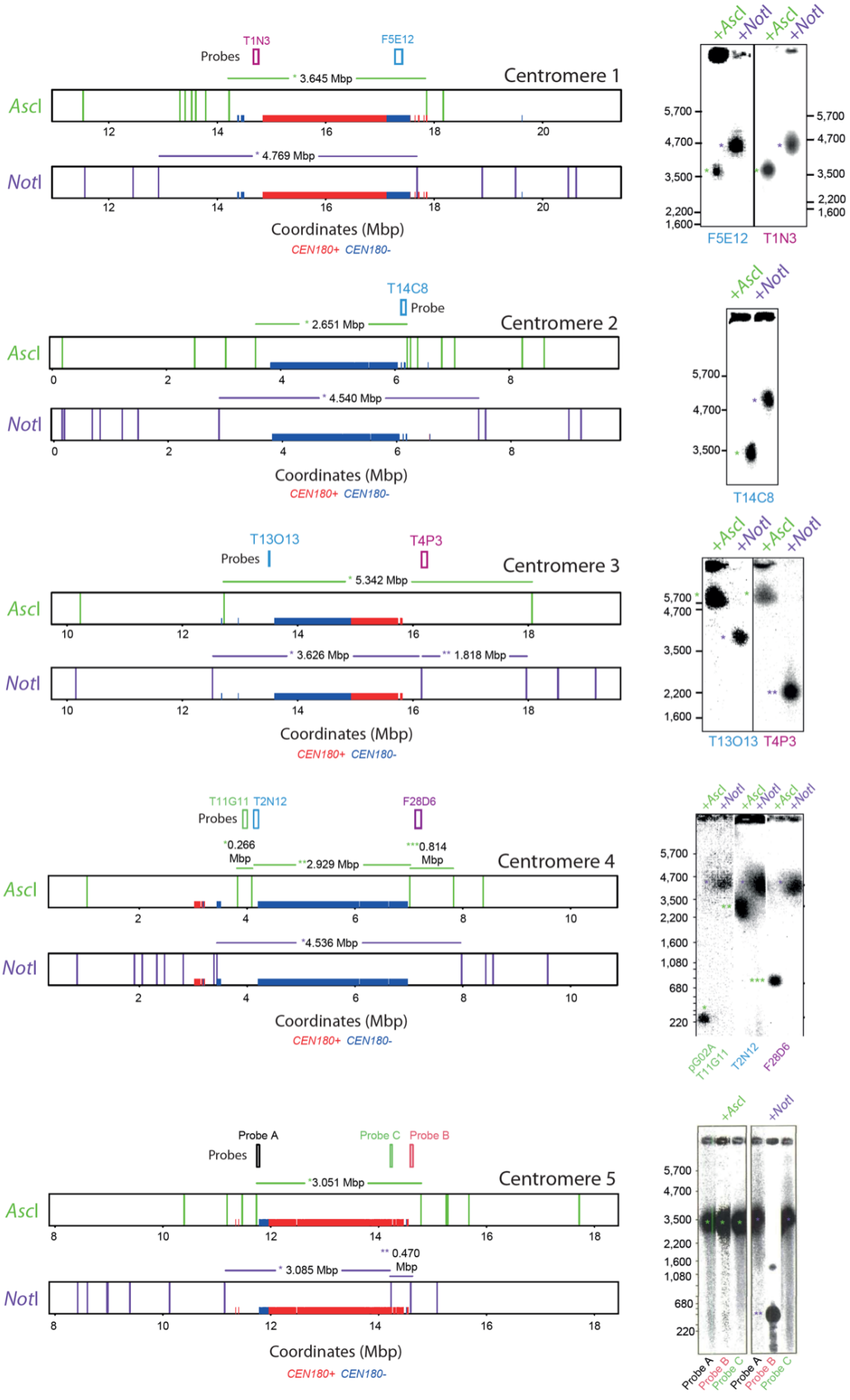

T1N3
Probes
F5E12
* 3.645 Mbp
Centromere 1
AscI
NotI
* 4.769 Mbp
Coordinates (Mbp)
CEN180+ CEN180-
+AscI +NotI +AscI +NotI
5,700
4,700
3,500
2,200
1,600
F5E12 T1N3
T14C8
Probe
* 2.651 Mbp
Centromere 2
* 4.540 Mbp
T14C8
T13O13
T4P3
* 5.342 Mbp
Centromere 3
* 3.626 Mbp
** 1.818 Mbp
T13O13 T4P3
T11G11 T2N12
F28D6
*0.266 Mbp
**2.929 Mbp
***0.814 Mbp
Centromere 4
*4.536 Mbp
1,080
680
220
pG02A
T11G11
T2N12
F28D6
Probe A
Probe C
Probe B
*3.051 Mbp
Centromere 5
** 0.470 Mbp
* 3.085 Mbp
Probe A
Probe B
Probe C

Figure S6. Comparison of the Col-CEN assembly with physical maps derived from pulsed-field gel electrophoresis and Southern blotting. On the right hand side of the figure published pulsed-field gel electrophoresis and Southern blotting data are shown, where genomic DNA was digested using either *Asc*I or *Not*I (*16*, *91*, *92*). The probe used for hybridization is labelled underneath the blots. To the left are physical maps of the Col-CEN assembly that have been virtually digested for *Asc*I (green) or *Not*I (purple) and site locations indicated relative to chromosome coordinates. The position of plus strand (red) and minus strand (blue) *CEN180* are indicated on the x axis. Above each physical map the location of the probes used for Southern blot hybridization are indicated. We further annotate the predicted size of cross-hybridizing fragments following restriction digestion, for comparison with the reproduced data. We note that for *CEN1* the authors interpret probe hybridization as indicating binding to two separate ~4.7 Mbp arrays. However, an incorrect BAC sequence used when designing the restriction maps (specifically, BAC F8L2 sequence: https://www.ncbi.nlm.nih.gov/nuccore/AC087569) predicted an incorrect *Not*I site, which was inside of the *Asc*I cutting site. However, based on analysis of our assembly the *Not*I site is in fact outside of the *Asc*I site and thus the probes are binding to the same fragment (*16*). This region has now also been resolved correctly in the TAIR10 reference assembly.

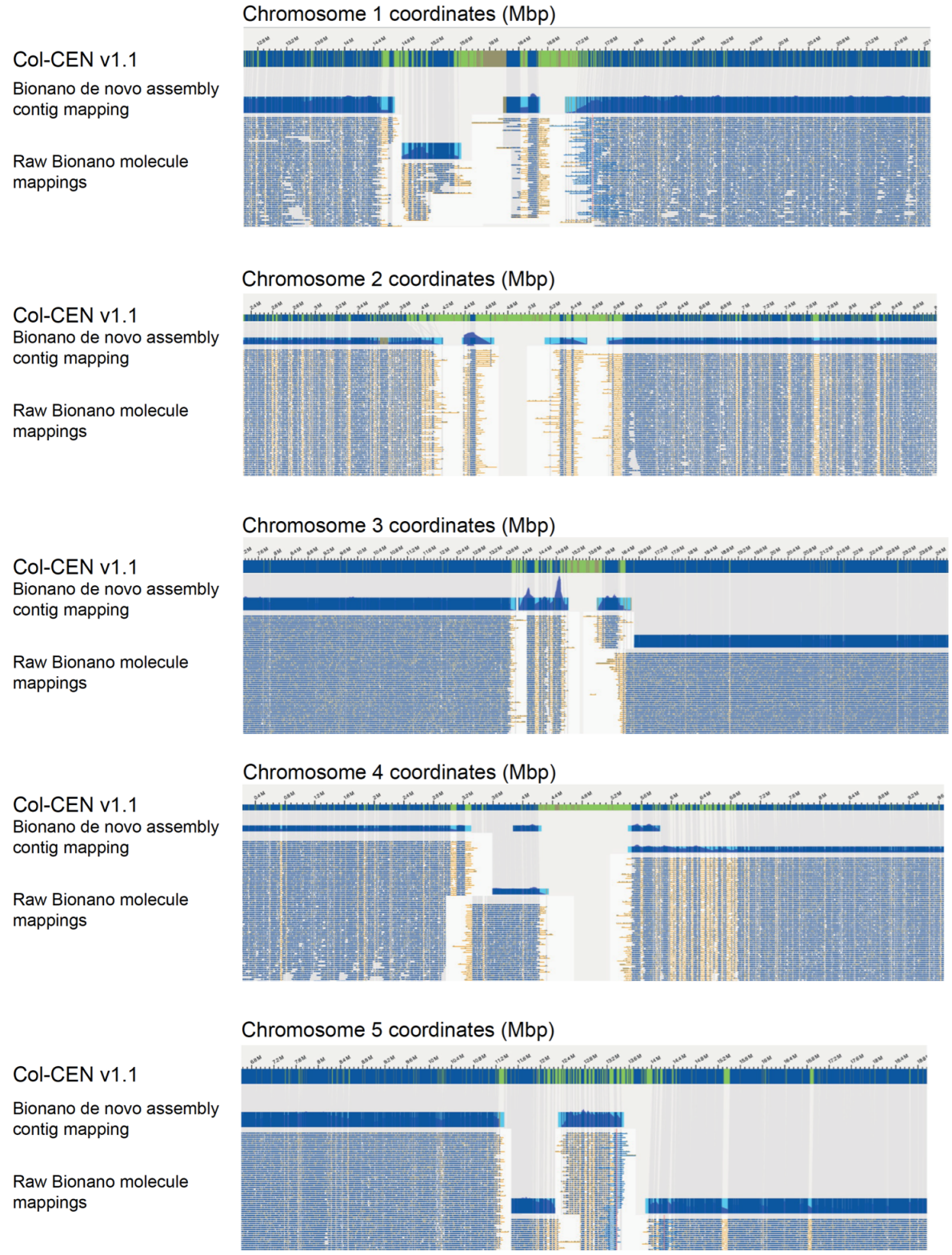


Figure S7. Bionano optical mapping across the Col-0 centromeres. Bionano *de novo* assembly contigs mapped to the Col-CEN reference assembly. The green and blue bars represent the expected labeling positions in the ONT reference assembly, where blue bars are expected labeling positions,

green regions lack Bionano labels and light brown bars represent predicted labeling positions not linked to a Bionano optical contig. Centromere regions generally lack predicted labeling sequences and therefore Bionano *de novo* assembled contigs are broken. Below the Bionano contigs (blue background with blue bars) are raw molecule mappings to the Bionano contigs at ~1,000× coverage (yellow background with blue dots indicating labelled sites).

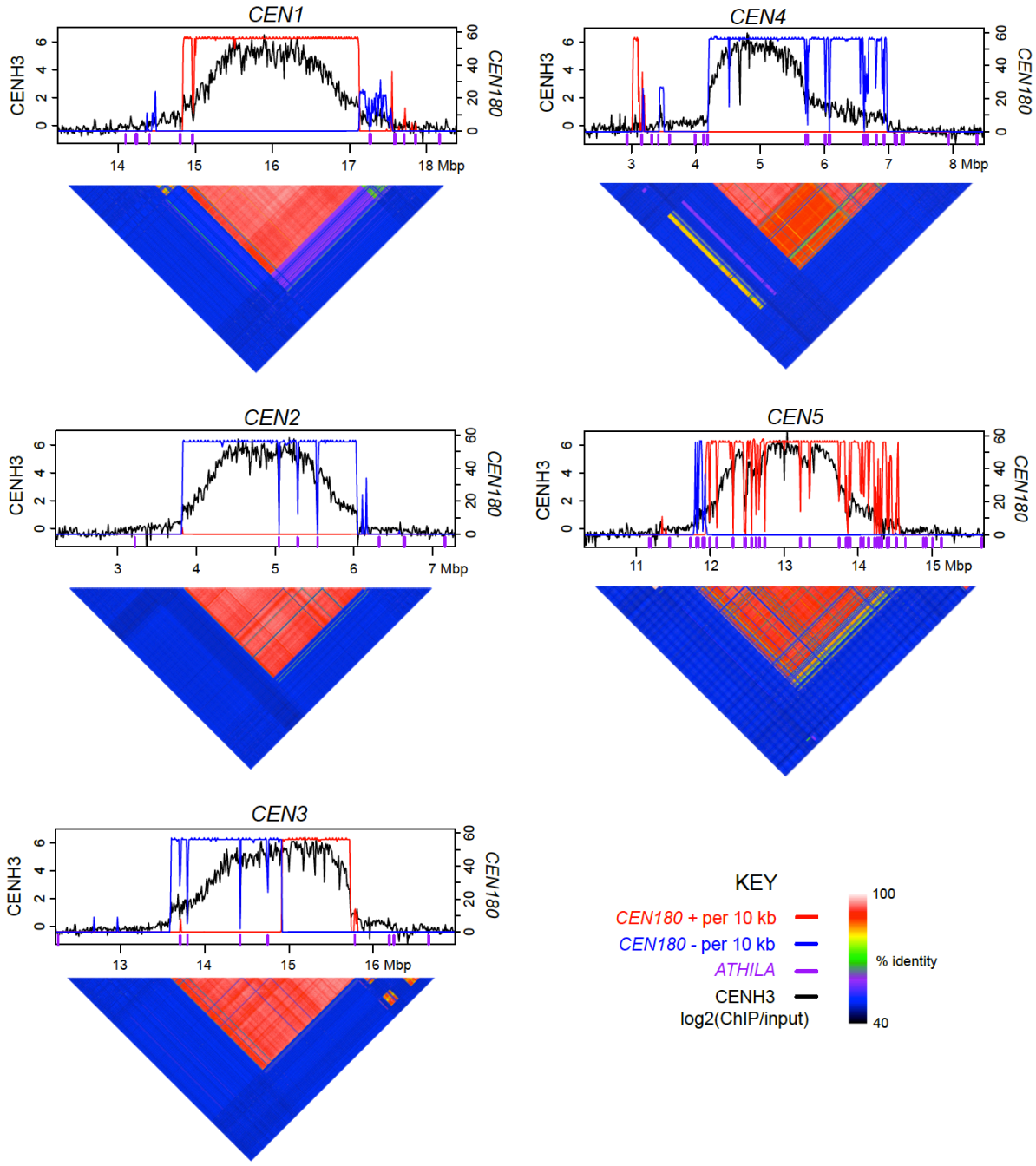


Figure S8. CENH3, *CEN180* and sequence identity across the Arabidopsis centromeres. CENH3 log2(ChIP/input) (black) (*10*), plotted over each centromere. *CEN180* density per 10 kbp is plotted showing forward (red) or reverse (blue) strand orientation. The location of *ATHILA* retrotransposons is indicated by purple ticks on the x axis. Beneath the plot are heatmaps indicating pairwise % identity values of all non-overlapping 5 kbp regions.

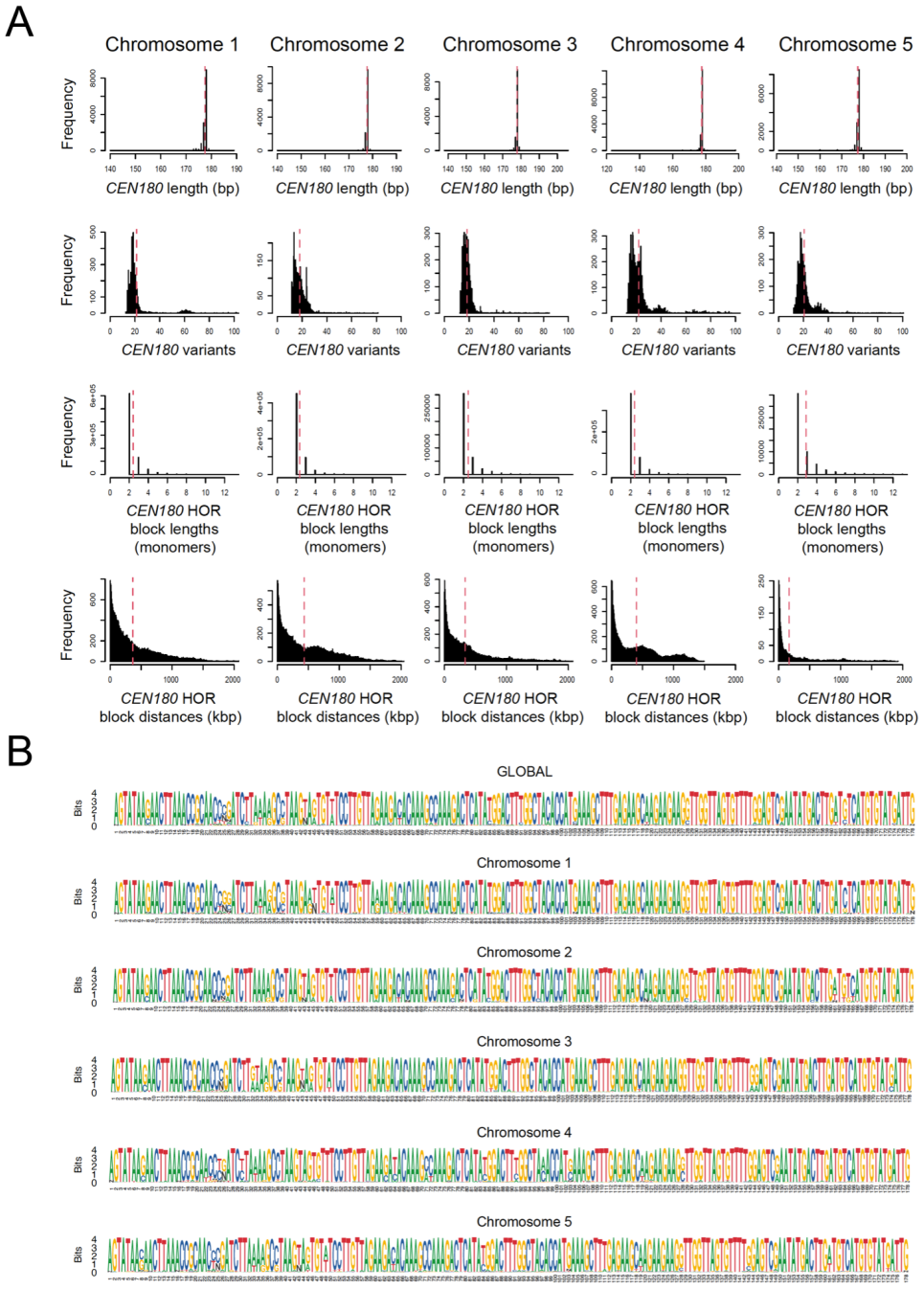


Figure S9. The Arabidopsis *CEN180* satellite repeat library analysed by chromosome. A. Histograms of *CEN180* monomer lengths (bp), and variants relative to the genome-wide consensus, shown for each chromosome. Mean values are shown by the red dotted line. **B.** *CEN180* sequence conservation represented by sequence logo plots. The global genome-wide sequence logo is shown first, followed by each individual chromosome. Positions with less than 50% coverage are not shown.

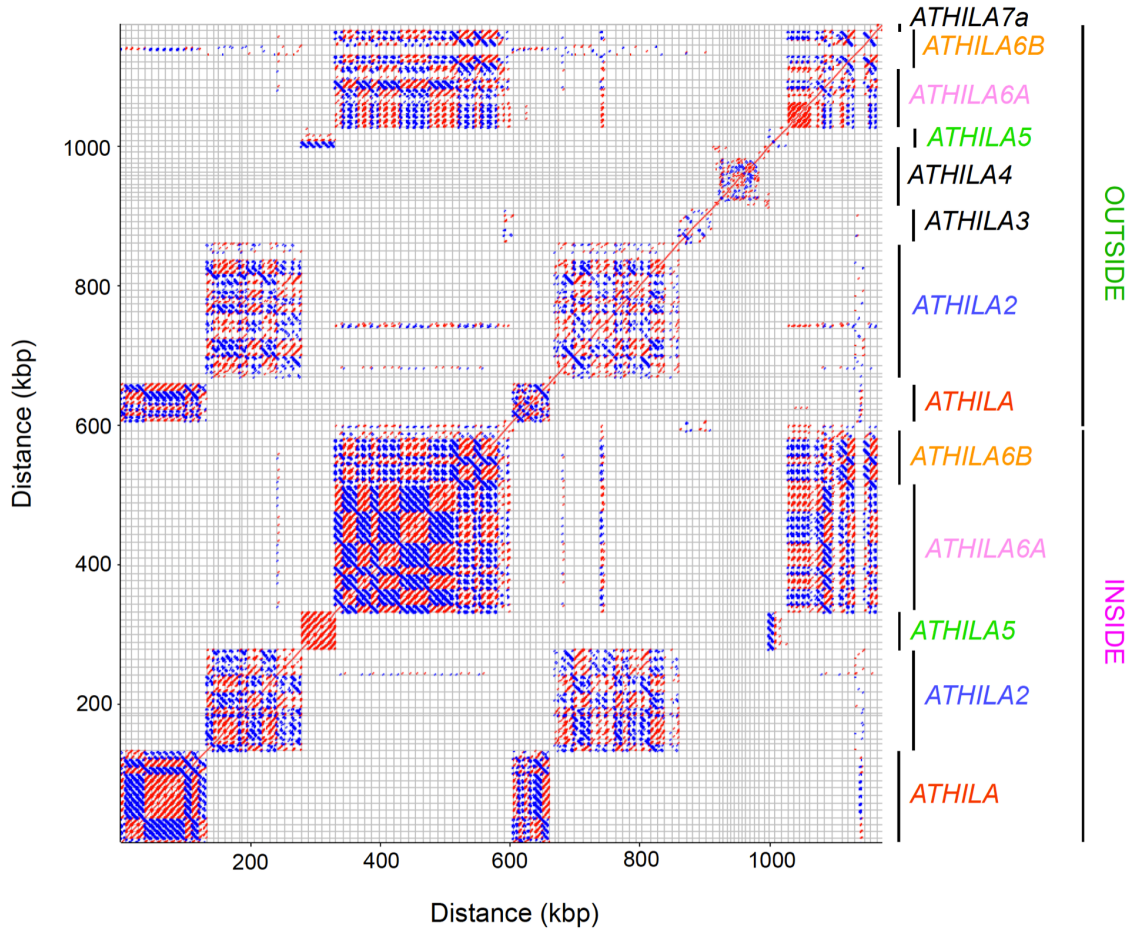
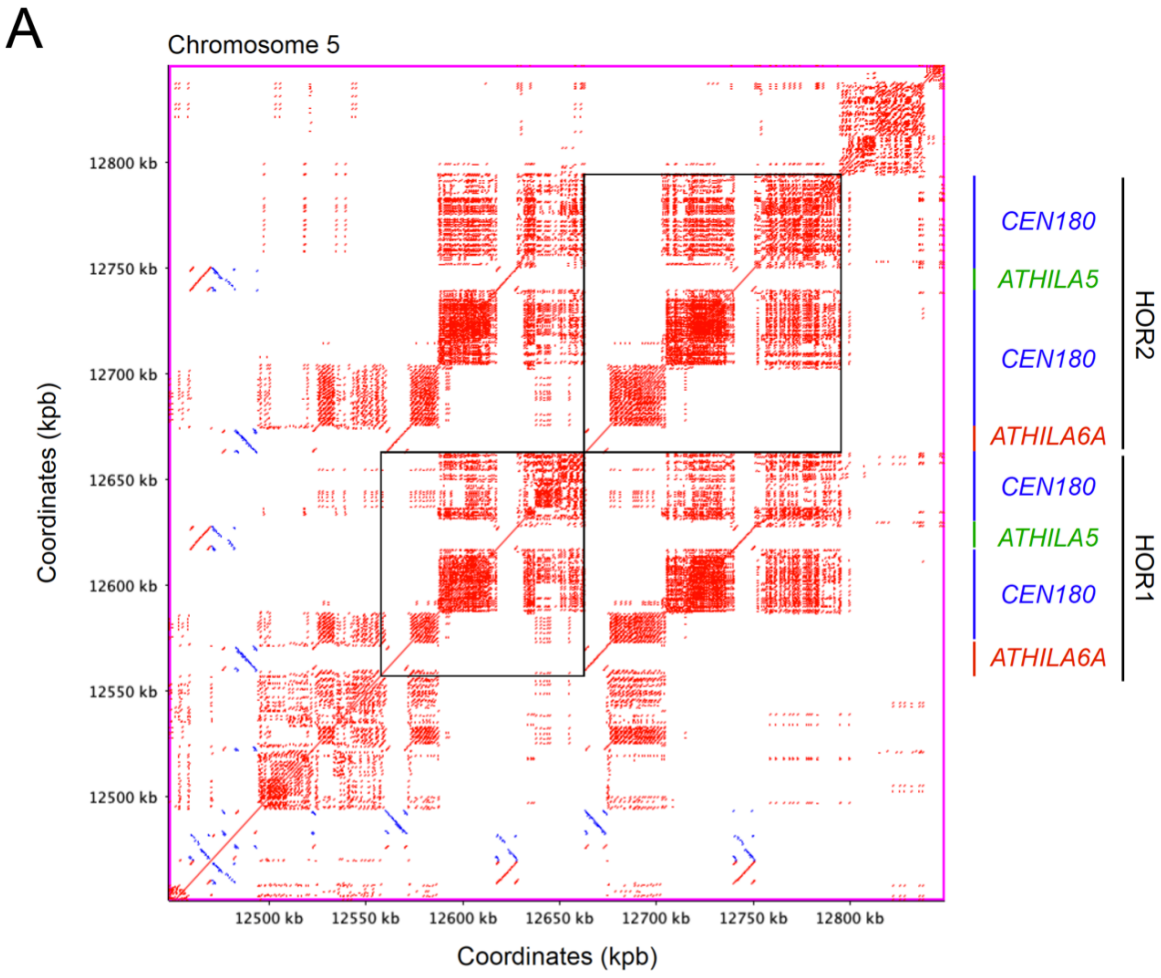


Figure S10. Dotplot comparison of *ATHILA* retrotransposons located inside or outside the main centromeric *CEN180* arrays. Dotplot of centromeric *ATHILA* retrotransposons using a search window of 75 bp. Red and blue indicate forward and reverse strand similarity. The elements assigned to different *ATHILA* subfamilies are indicated, in addition to whether they are located inside or outside the main centromeric *CEN180* arrays.


A
Chromosome 5
12800 kb
12750 kb
12700 kb
12650 kb
12600 kb
12550 kb
12500 kb
Coordinates (kpb)
12500 kb
12550 kb
12600 kb
12650 kb
12700 kb
12750 kb
12800 kb
Coordinates (kpb)
CEN180
ATHILA5
CEN180
ATHILA6A
HOR2
CEN180
ATHILA5
CEN180
ATHILA6A
HOR1

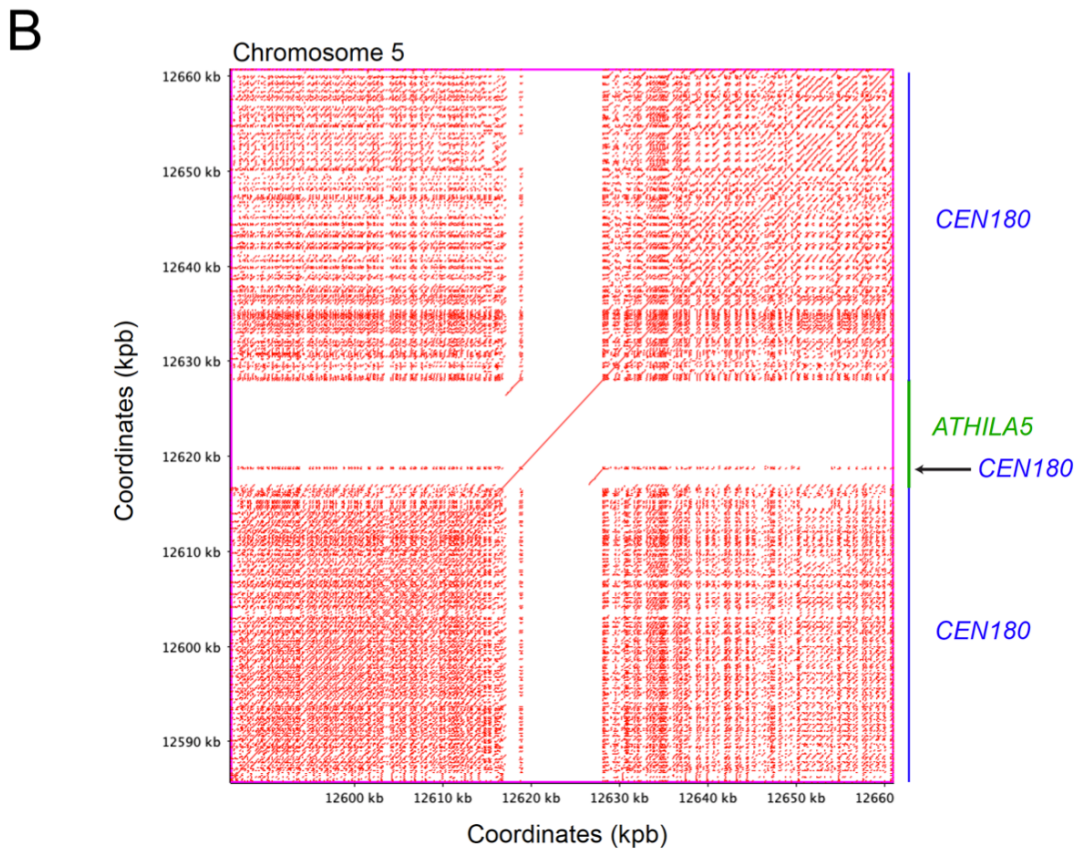

B
Chromosome 5
12660 kb
12650 kb
12640 kb
12630 kb
12620 kb
12610 kb
12600 kb
12590 kb
Coordinates (kpb)
12600 kb
12610 kb
12620 kb
12630 kb
12640 kb
12650 kb
12660
Coordinates (kpb)
CEN180
ATHILA5
CEN180
CEN180

Figure S11. Higher order duplication of *ATHILA* elements post-integration. A. Dotplot analysis of a large region that has duplicated within the centromere of chromosome 5, forming higher order repeats (HOR1 and HOR2). The boundaries of each HOR are indicated by the black boxes within the dotplot. Each higher order repeat contains one *ATHILA5* and one *ATHILA6A* element that show high identity (99.5 and 99.6%) between copies. In contrast, the surrounding blocks of *CEN180* repeats within each HOR are more variable in size and show lower sequence identity (94.3-97.3%). Additional evidence that this region was duplicated after the insertion of the *ATHILA5* and *ATHILA6A* copies includes, i) their nearly identical lengths (11,345 vs. 11,346 bp for *ATHILA6A*, and 10,968 vs. 10,961 bp for *ATHILA5*), ii) the identical target site duplication (TSD) for the *ATHILA5* copies (GTAGT), iii) the identical flanking sequences (CCTAAGTAGT for the upstream and GTAGTGTTTC for the downstream region of *ATHILA5*, and AGACACAAAG for the downstream region of *ATHILA6A*), and iv) the fact that both *ATHILA5* contain internal *CEN180* copies in identical positions within their 5'-LTRs (see B). **B.** Dotplot analysis of one of the duplicated *ATHILA5* elements from A, which contains one complete and one partial copy of *CEN180*, located internally and downstream of the 5'-LTR. We postulate that the *CEN180* repeats inserted within the original *ATHILA5* copy prior to this region being duplicated.

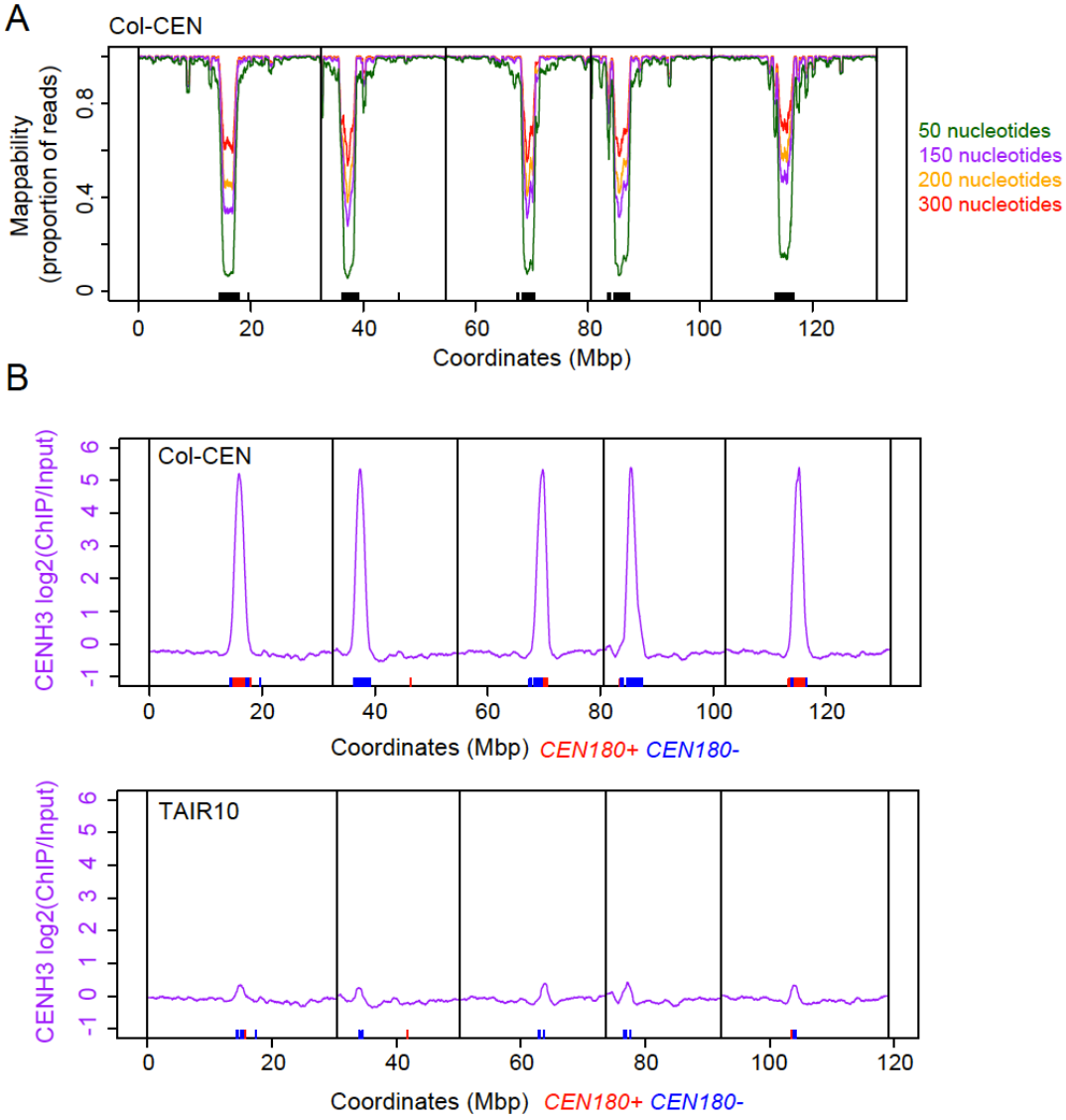


Figure S12. Mappability within the centromeres and CENH3 ChIP-enrichment compared between the Col-CEN and TAIR10 assemblies. A. Genome mappability was computed based on the uniqueness of *k*-mers for each genomic position, with up to *e* mismatches permitted (zero mismatches were permitted) using GenMap v1.3.0 (*93*, *94*). The uniqueness of *k*-mers, or (*k*,*e*)-mappability, was calculated for each position using 50-, 150-, 200- and 300-mers. (*k*,*e*)-mappability for a given position represents the reciprocal value of the frequency with which the *k*-mer occurs in the genome. Chromosome-scale profiles were generated by calculating mean (*k*,*e*)-mappability values within adjacent 10-kb genomic windows. **B.** CENH3 $\log_2$(ChIP/Input (purple) plotted along the Col-CEN (upper) or TAIR10 (lower) chromosomes. *CEN180* are indicated as ticks on the x-axis for forward (red) and reverse (blue) strand.

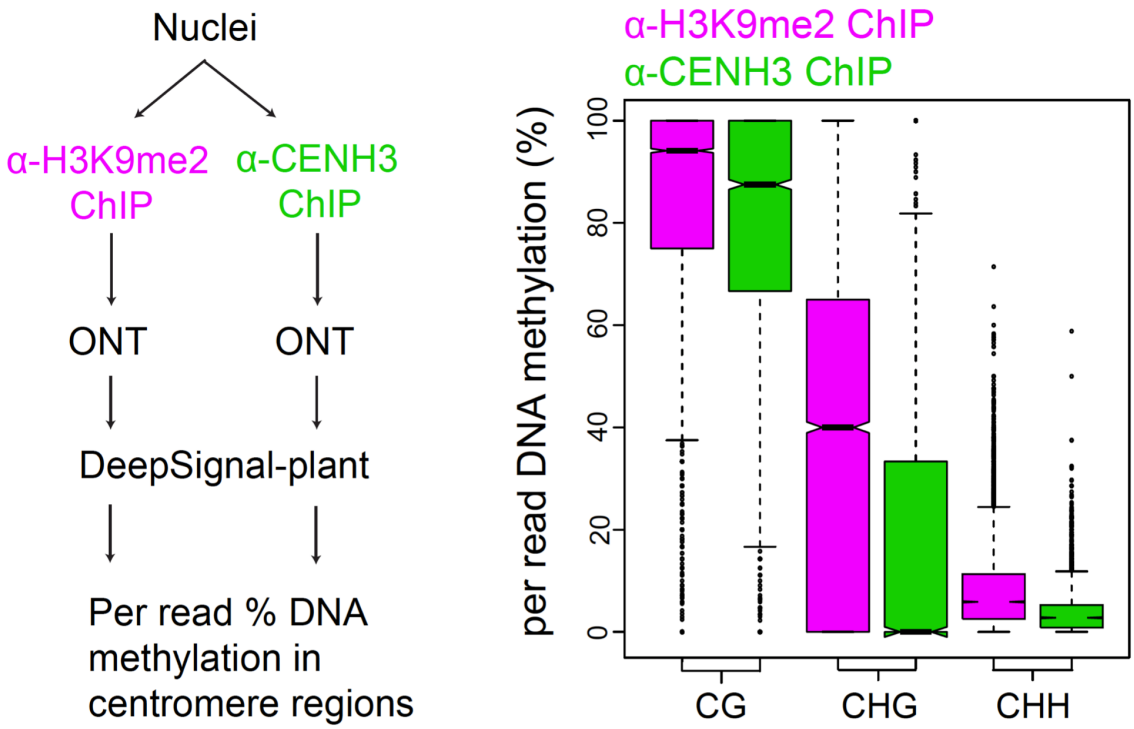


Figure S13. Profiling DNA methylation of H3K9me2 and CENH3 ChIP DNA using ONT. We performed ChIP-seq on Col-0 nuclei using H3K9me2 or CENH3 antibodies. The resulting DNA was then sequenced using a ONT Flongle flow cell. Reads were mapped to the Col-CEN assembly and filtered for those aligning within the centromeres. Read IDs were extracted, duplicates removed, and then used to extract fast5 files. The fast5 files were then analysed using DeepSignal-plant in order to calculate the mean methylation value for each context across each read. The boxplot shows mean DNA methylation levels across single reads for the CG, CHG and CHH sequence contexts. We observe that methylation is significantly lower in the CENH3 ChIP reads compared to H3K9me2, and that the difference is strongest for the CHG and CHH sequence contexts. CG context methylation is high in both H3K9me2 or CENH3 ChIP-seq read sets.

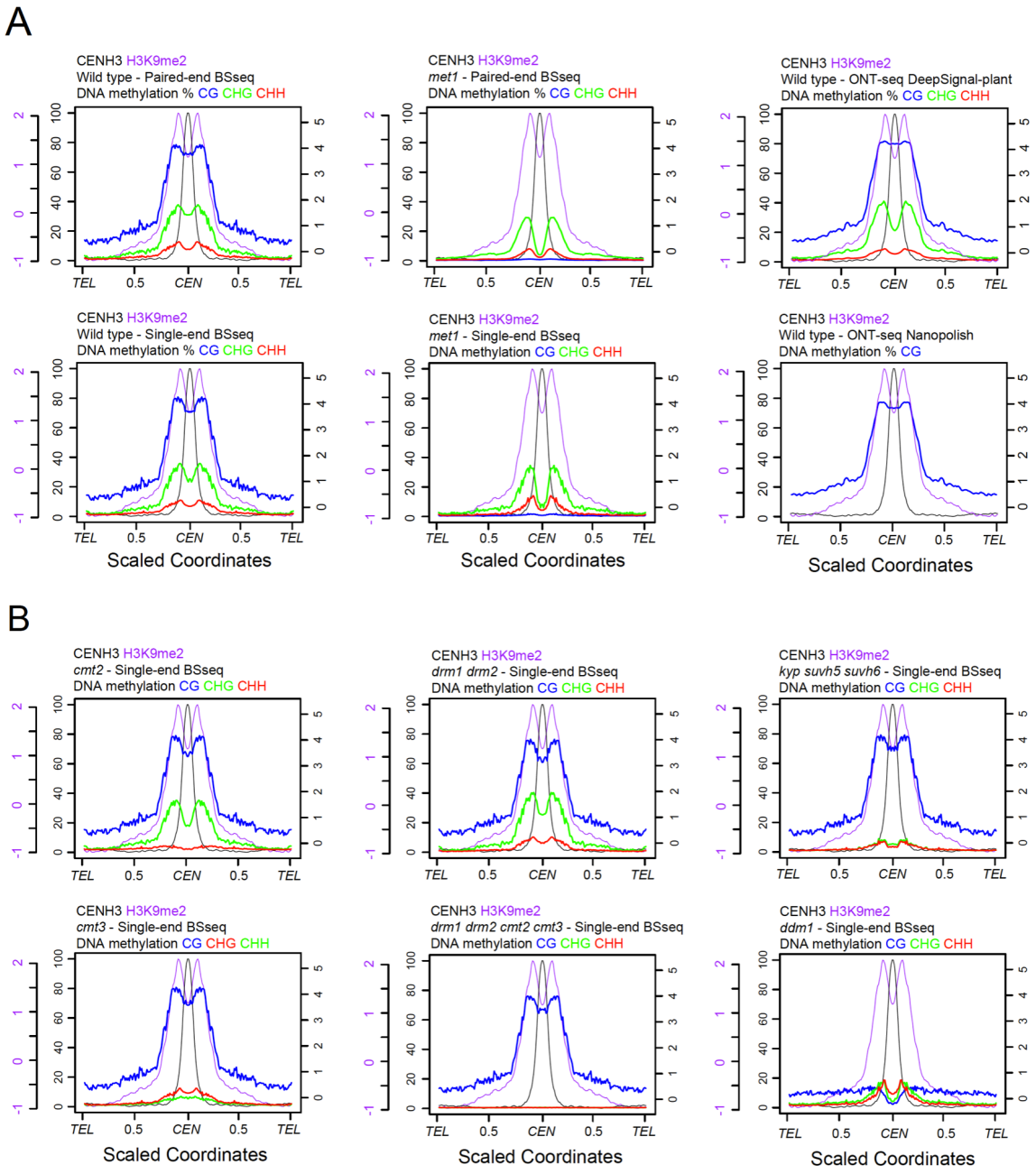


Figure S14. Centromeric DNA methylation in wild type and CG and non-CG context pathway mutants. A. Plots of CENH3 (black) and H3K9me2 (purple) ChIP-seq enrichment along chromosomes scaled proportionally along the telomere-centromere axes (*10*, *26*). DNA methylation profiles calculated from BS-seq data are plotted for CG (blue), CHG (red) and CHH (green) sequence contexts in the indicated genotypes (*22*, *29*). Comparison of Col-0 and *met1* is shown using independent data sets that were sequenced with either paired-end or single-end reads (*22*, *29*). As a comparison, DNA methylation profiles generated from ONT reads using the DeepSignal-plant and Nanopolish algorithms

are shown to the right. **B.** As for A., but comparing data from *cmt2*, *cmt3*, *drm1 drm2*, *drm1 drm2 cmt2 cmt3*, *kyp suvh5 suvh6* and *ddm1* (*21*, *22*).

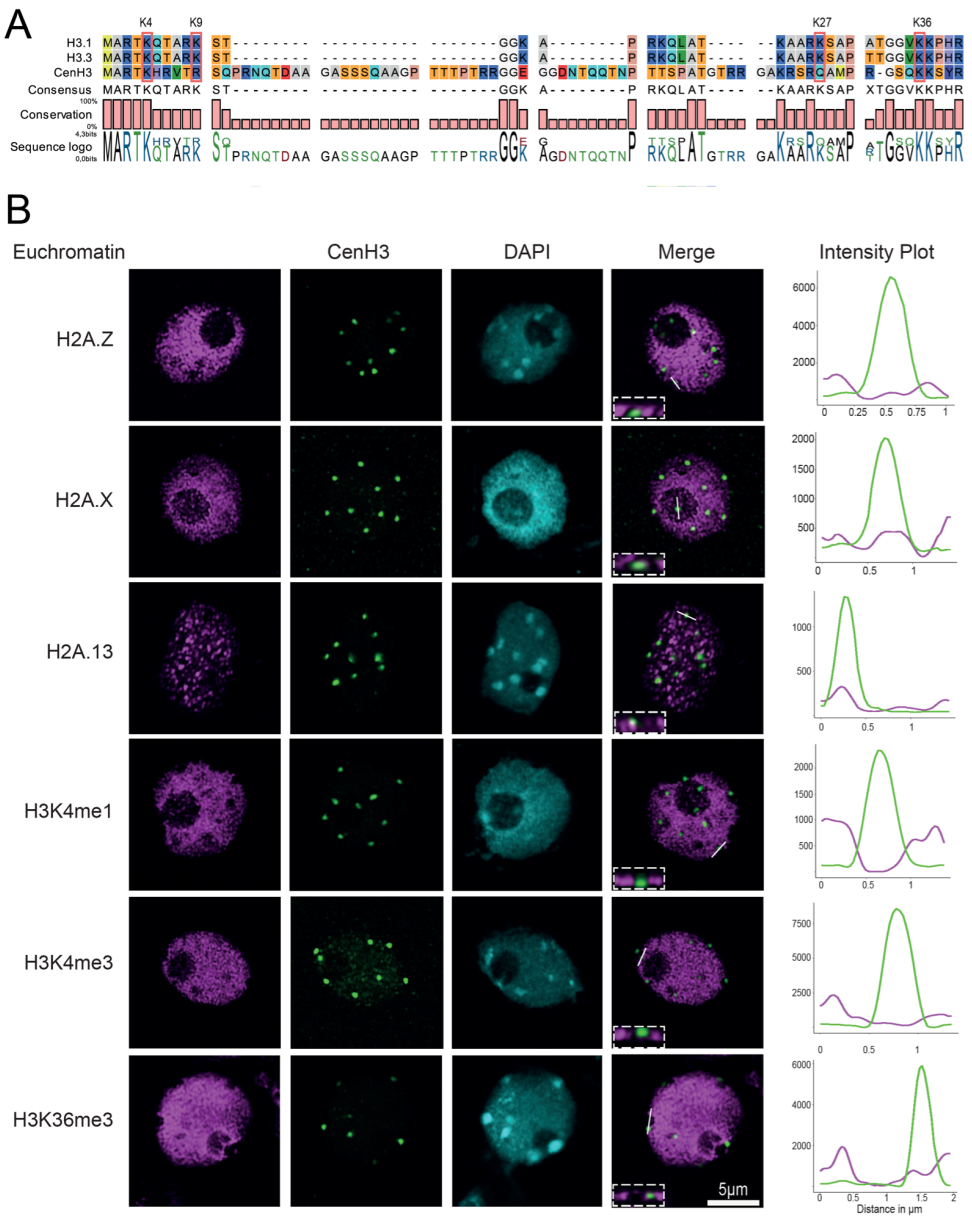


Figure S15. Immunofluorescence analysis of euchromatic marks in isolated nuclei relative to CENH3. A. Protein sequences of Arabidopsis H3.1, H3.3 and CENH3 were aligned using CLC Main Workbench. H3 N-terminal lysine residues known to be modified and investigated here are highlighted in red. **B.** Arabidopsis nuclei were stained for euchromatic marks (Magenta) and CENH3-GFP (green) and DNA (cyan=DAPI). The white line indicates the area of the confocal section. The confocal section is also depicted at the left bottom of each merged image. The intensity plot for the confocal section is shown on the right. Scale bars are 5µm.

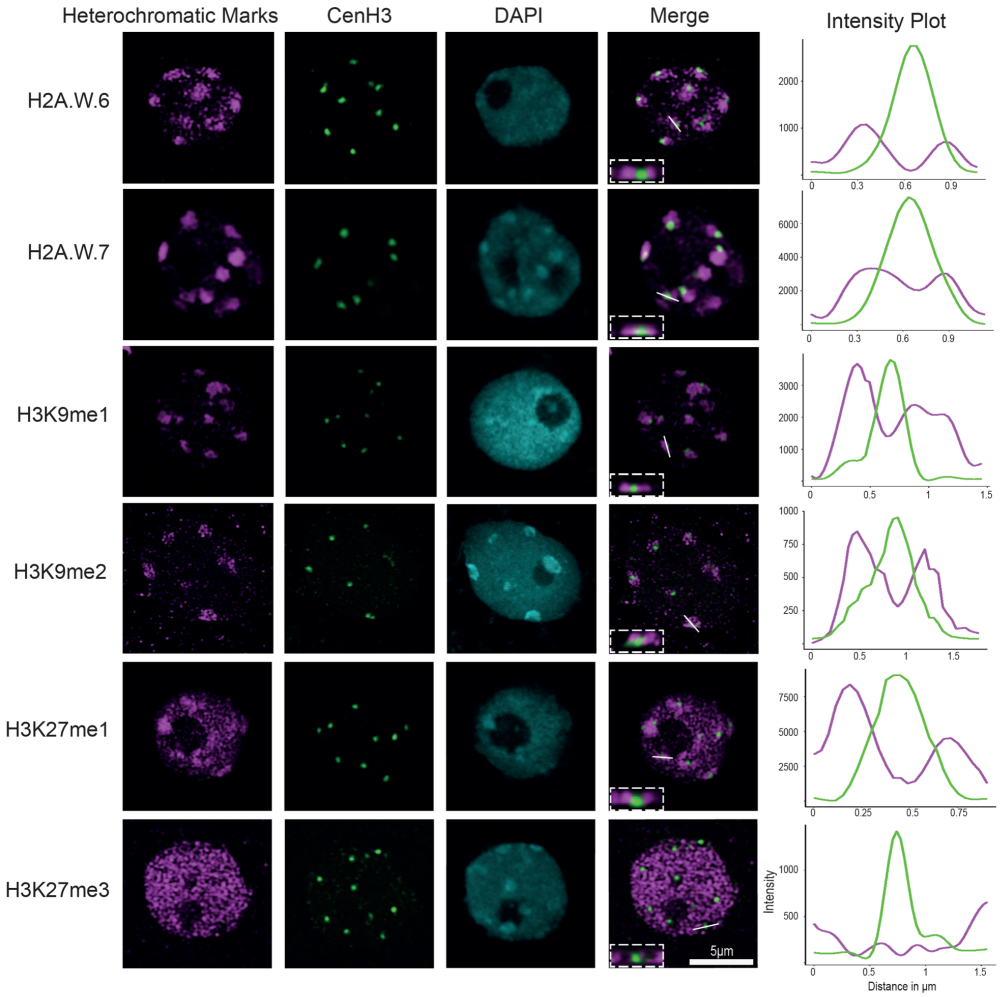


Figure S16. Immunofluorescence analysis of heterochromatic marks in isolated nuclei relative to CENH3. Arabidopsis nuclei were stained for heterochromatic marks (Magenta) and CENH3-GFP (green) and DNA (cyan=DAPI). The white line indicates the area of the confocal section. The confocal section is also depicted at the left bottom of each merged image. The intensity plot for the confocal section is shown on the right. Scale bars are 5 µm.

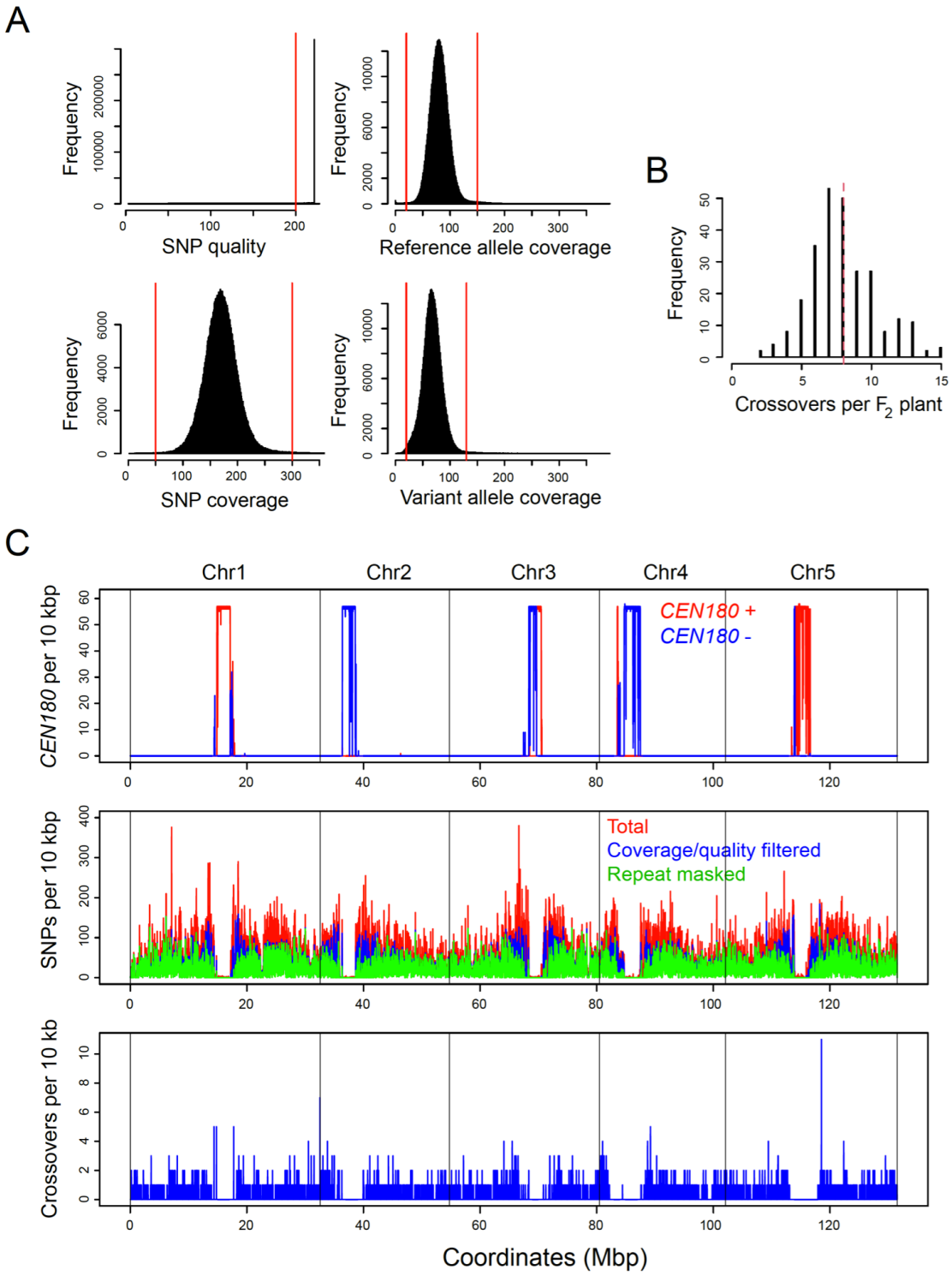
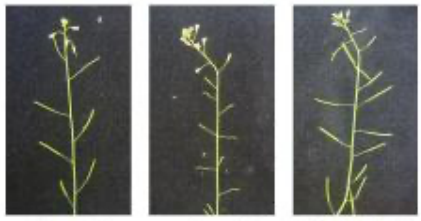


Figure S17. Mapping Col×Ler single nucleotide polymorphisms (SNPs) and crossovers against the Col-0 centromere assembly. A. Histograms showing the frequency of qualities, coverage, reference and variant allele coverages for single nucleotide polymorphisms (SNPs) called against the assembly using data from 260 Col×Ler genomic DNA F_2 sequencing libraries. The red lines indicate thresholds where sites were filtered out of analysis. **B.** Histogram of crossovers mapped against the

assembly per Col×Ler F_2 plant. The red dotted line indicates the mean value. **C.** Plot of the assembly showing *CEN180* satellite density per 10 kbp for forward (red) and reverse (blue) strands (upper). Beneath, the frequency per 10 kbp of total Col×Ler SNPs (red) are plotted, in addition to SNP frequency filtered for quality and coverage values, as in A (blue), and SNPs following repeat-masking (green). The lower plot shows crossovers per 10 kbp (blue) mapped against the assembly.

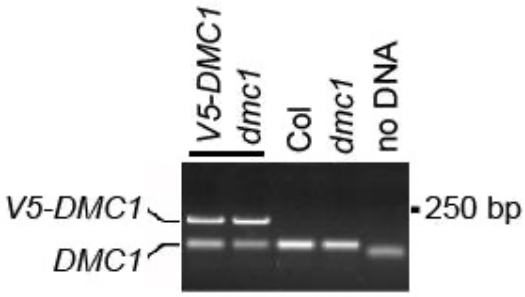
A



B

Genotype	Average Seed per Silique	Standard Deviation
Col	50.9	4.7
dmc1	2.3	0.6
V5-DMC1 dmc1	51.2	3.9

C



D

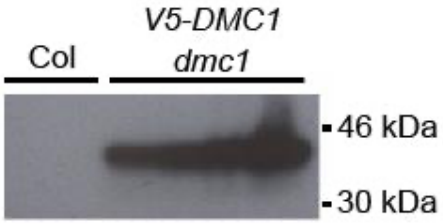


Figure S18. Epitope-tagging and functional complementation of *V5-DMC1*. A. Inflorescences of wild type (Col-0), *dmc1-3* and *V5-DMC1 dmc1-3*. Fertility is evident from silique length. **B.** Quantification of seed set per silique in wild type (Col-0), *dmc1-3* and *V5-DMC1 dmc1-3*. **C.** PCR based detection of the N-terminally epitope-tagged *V5-DMC1* transgene, alongside Col-0 and *dmc1-3* null controls. PCR primers flank the *DMC1* ATG translation start site. The expected PCR product sizes are 203 and 74 bp for epitope-tagged and wildtype *DMC1*, respectively. Unincorporated oligonucleotides are seen in 'no DNA' control. **D.** α-V5 western blot from Col-0 and *V5-DMC1 dmc1-3* protein extracts from closed flower buds. The expected size of *V5-DMC1* is 41.7 kDa.

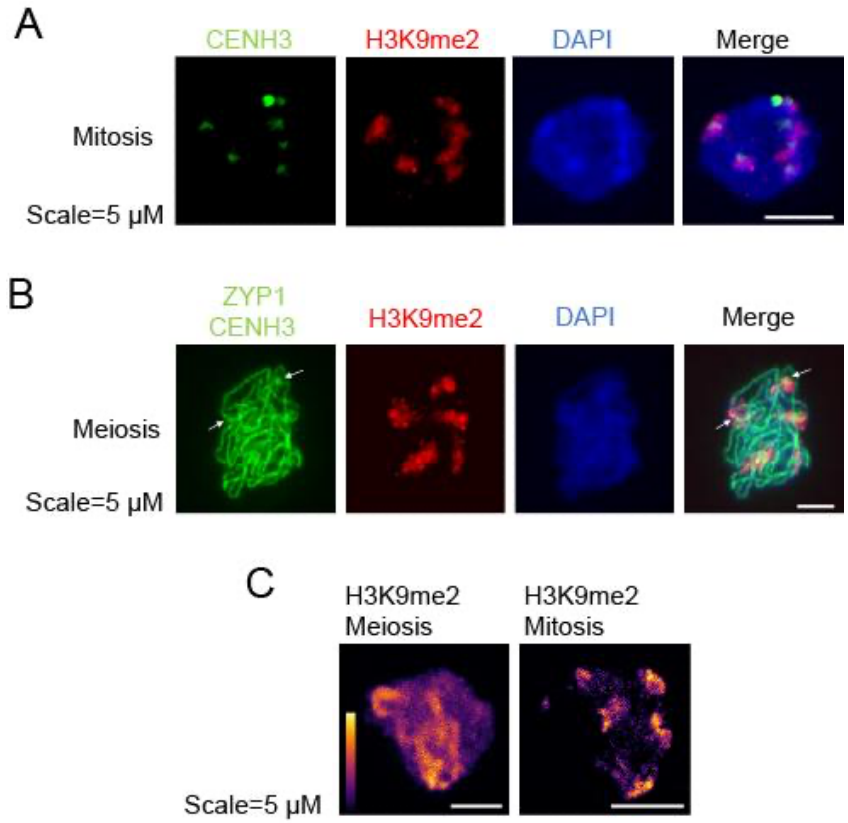


Figure S19. Immunocytological staining of the Arabidopsis centromeres. A. Somatic interphase nucleus immunostained for CENH3 (green), H3K9me2 (red) and stained for DAPI. Scale bar = 5 µM. **B.** As for A, but showing an Arabidopsis male meiocyte in pachytene immunostained for CENH3 (green), ZYP1 (green) and H3K9me2 (red), and stained for DAPI (blue). Scale bar=5 µM. **C.** Mitotic and meiotic cells immunostained for H3K9me2 and imaged using STED super resolution microscopy**.** The colour-scale indicates the intensity of staining, with yellow representing the maximum intensity. Scale bars = 5 µM.

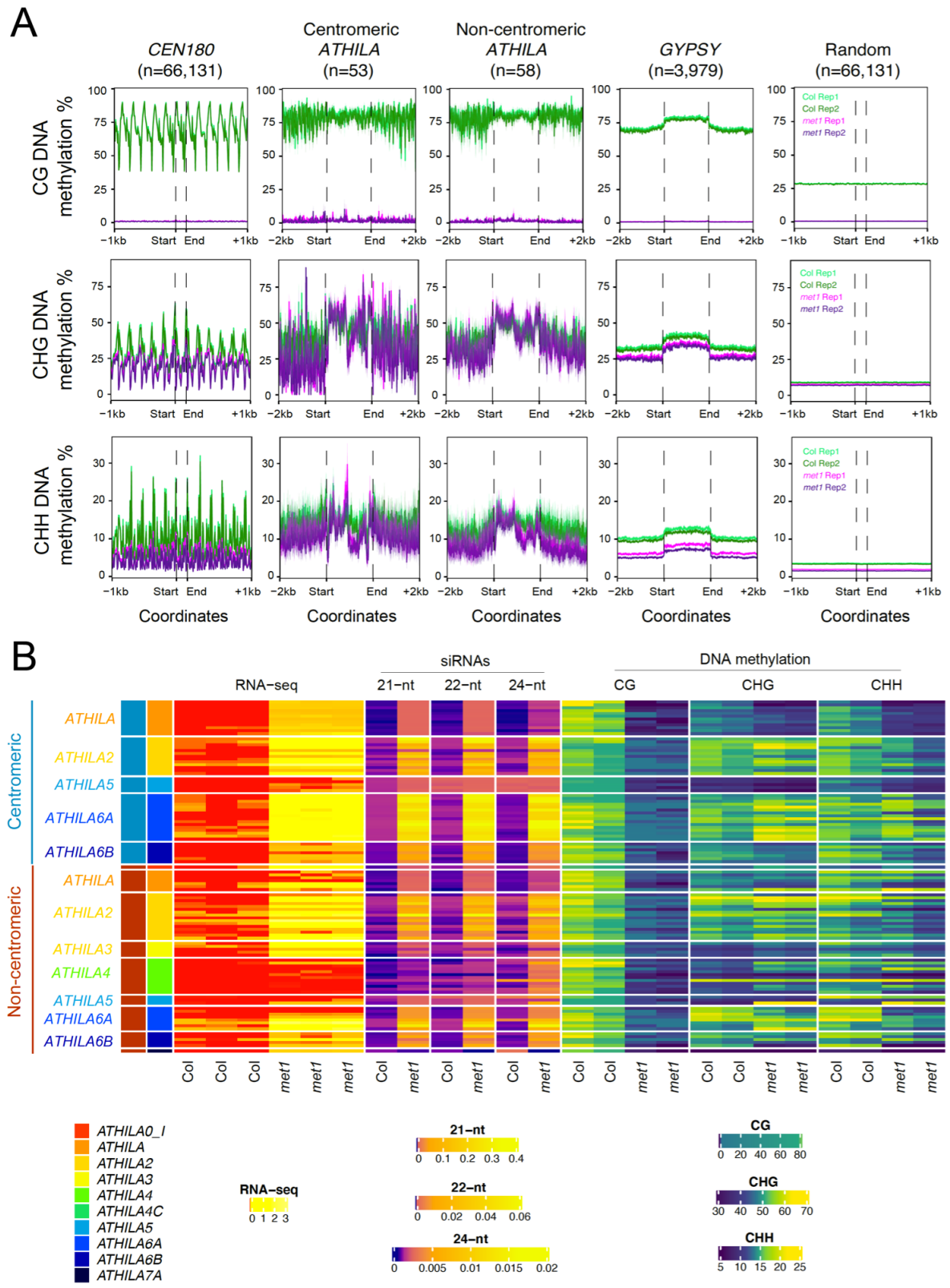


Figure S20. DNA methylation, RNA and siRNA expression associated with *ATHILA* elements in wild type and *met1*. A. CG, CHG and CHH context DNA methylation in wild type (Col-0, green) or *met1* (pink/purple) measured using BS-seq (*29*), over *CEN180* (n=66,131), centromeric *ATHILA*

(n=53), non-centromeric *ATHILA* (n=58), all *GYPSY* retrotransposons in the genome (n=3,979) and random positions (n=66,131). Shaded ribbons represent 95% confidence intervals for windowed mean values. **B.** Heatmap analysis of RNA-seq (*29*), siRNA-seq (*29*) and DNA methylation (*29*) data from wild type (Col-0) or *met1*. Each row represents an individual *ATHILA*, ordered according to their location within or outside the main centromeric *CEN180* arrays, and then by subfamily.

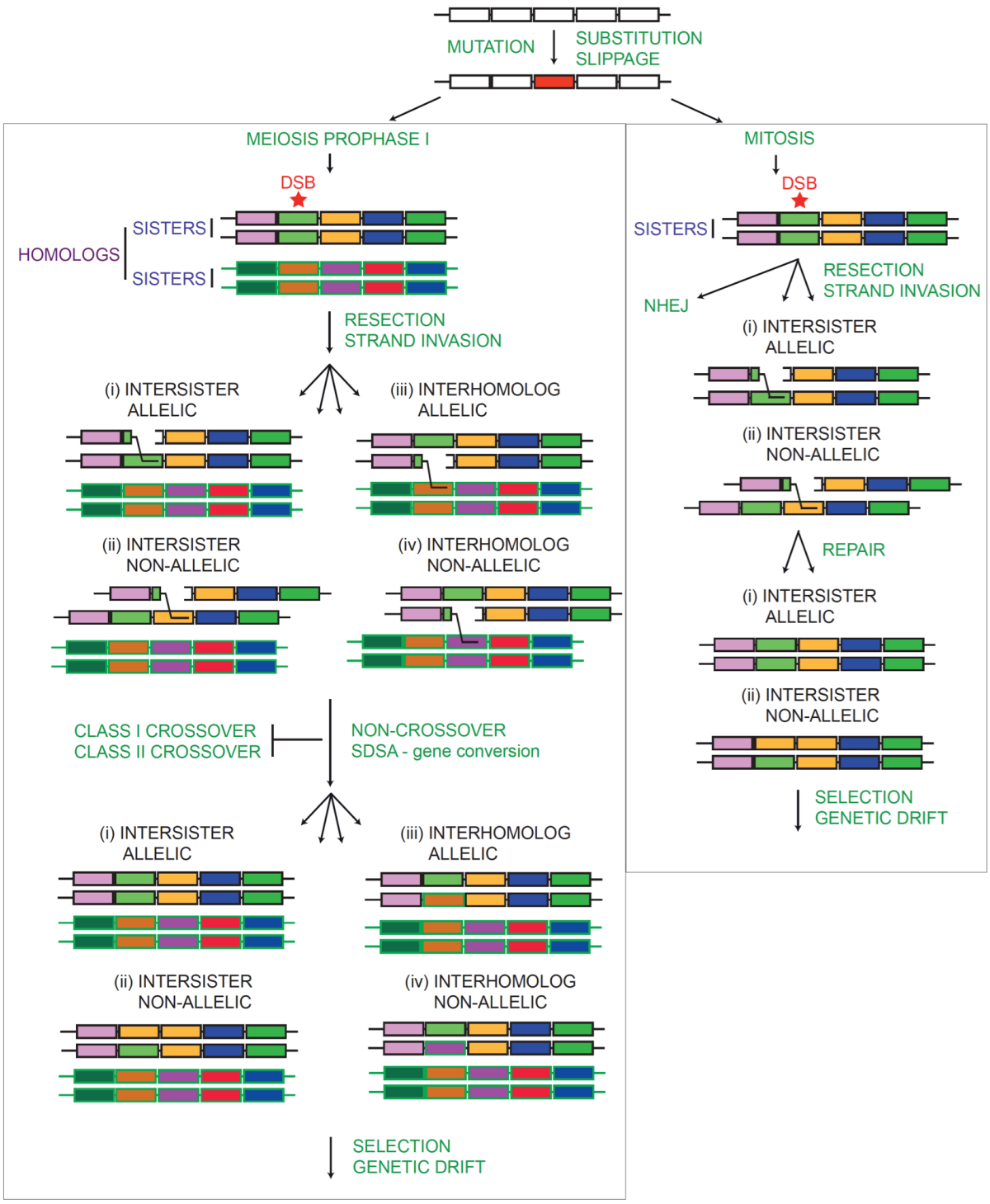


Figure S21. Model for *CEN180* sequence evolution in Arabidopsis. At the top of the diagram a representative array of five *CEN180* monomers (rectangles) is shown. Mutations, including base substitutions and replication slippage, generate monomer sequence variants (red). On the left hand side of the diagram we consider a similar representative region of five *CEN180* passing through meiosis, each of which has a distinct sequence, indicated by color. The 4 chromosomes are shown as two sisters

of each homolog. During meiotic prophase I, one chromosome experiences a DNA double strand break (DSB, red star). The DSB is processed via resection to form single stranded DNA that is bound by RAD51/DMC1, which promote invasion of another chromosome. We show four possible scenarios where the invading strand enters, (i) an allelic location on the sister chromatid, (ii) a non-allelic location on the sister chromatid, (iii) an allelic location on a homolog, or (iv) a non-allelic location on a homolog. Crossover repair, via either the Class I or Class II pathways, are suppressed within the centromere. Therefore, we propose that centromeric strand invasion events are instead repaired via meiotic non-crossover pathways, including synthesis-dependent strand annealing (SDSA), which can result in gene conversion. For simplicity conversion of single *CEN180* repeats is indicated, although based on patterns of higher order repetition we propose resection and conversion may involve multiple monomer repeats (up to 60). Recombinant *CEN180* arrays generated by these pathways are then subject to selection and genetic drift in populations. On the right hand side of the diagram, we indicate that DSB formation and repair within the *CEN180* arrays may also occur outside of meiosis. In this case, repair may proceed via non-homologous end joining (NHEJ), or using intersister homologous recombination in either allelic or non-allelic locations. These pathways may also generate variation in *CEN180* arrays that will be subject to selection and genetic drift.

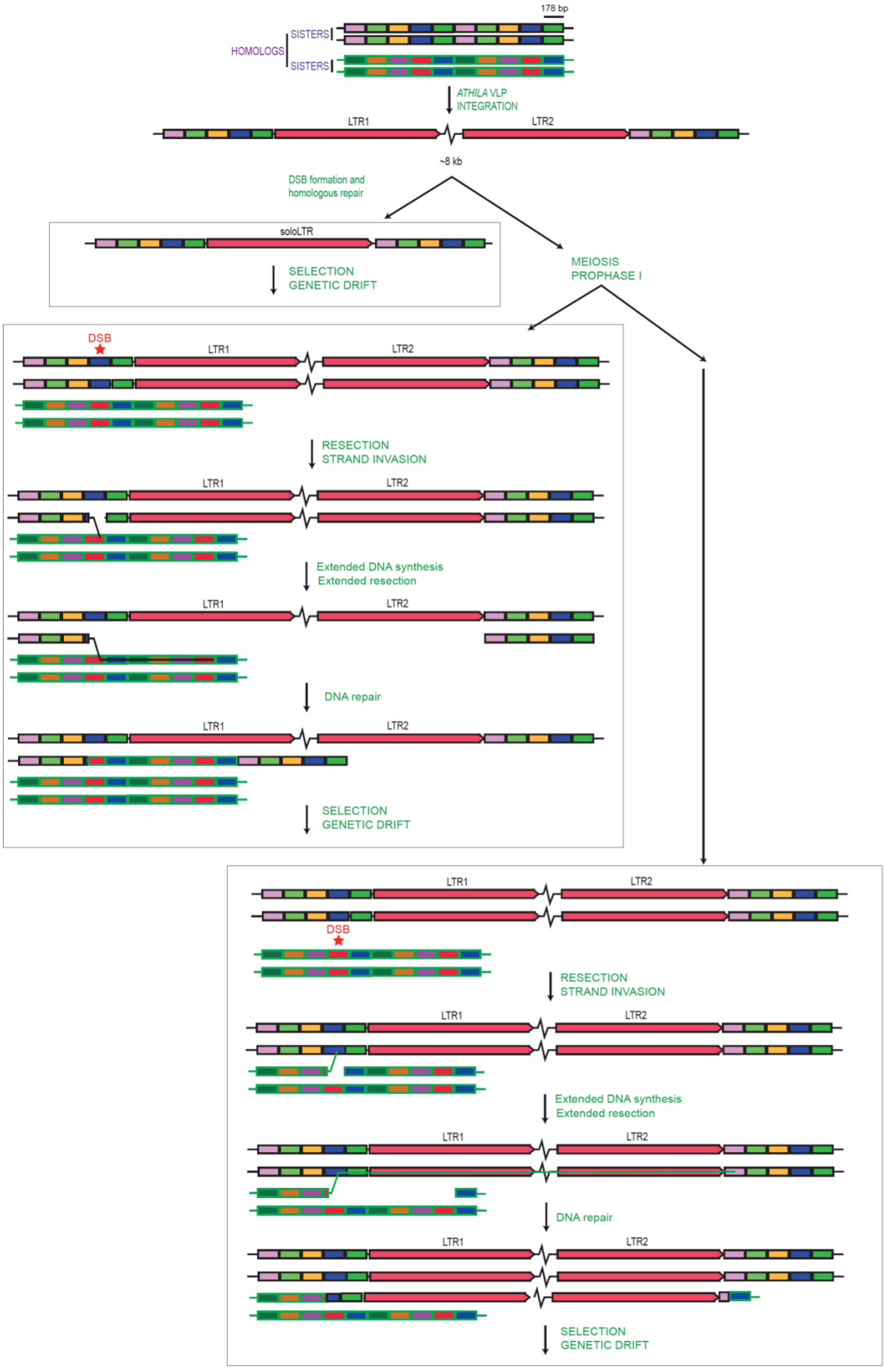

178 bp
HOMOLOGS
SISTERS
SISTERS
ATHILA VLP
INTEGRATION
LTR1
LTR2
~8 kb
DSB formation and
homologous repair
soloLTR
SELECTION
GENETIC DRIFT
MEIOSIS
PROPHASE I
DSB
LTR1
LTR2
RESECTION
STRAND INVASION
Extended DNA synthesis
Extended resection
DNA repair
SELECTION
GENETIC DRIFT

Figure S22. Model for *ATHILA* integration and sequence evolution within the Arabidopsis centromeres. We consider a representative region of ten *CEN180* monomers, with distinct monomers color-coded. The sister and homologous chromosomes are shown. A *de novo ATHILA* integration event is shown within one of the chromosomes. The paired long terminal repeats (LTRs, red) are shown approximately to scale, but the internal region of the transposon is not represented, but would typically consist of ~8 kbp of sequence. Following integration we consider three potential further changes to the *ATHILA* insertion. As we observe multiple centromeric *ATHILA* solo LTRs, we propose that DNA double strand break (DSB) formation and repair may occur within the *ATHILA* that results in formation of a solo LTR. This pathway may occur during mitosis or meiosis, and the resulting solo LTR would then be subject to selection and/or genetic drift. On the right hand side of the diagram we consider an alternative pathway during meiotic prophase I, showing two potential outcomes. In the left hand branch, a meiotic DSB (red star) forms in a *CEN180* linked to the *ATHILA* insertion (which is hemizygous). The DSB undergoes resection to form single stranded DNA (ssDNA) which is able to invade a homologous chromosome that lacks the *ATHILA* insertion. Based on the large size (10-100s kbp) of *CEN180* higher order repeats that we observe, we propose that an extended form of resection may occur that causes deletion of the *ATHILA* from the donor chromosome. The invading strand then undergoes template driven DNA synthesis that copies *CEN180* sequence from a different chromosome. Following dissolution of strand invasion and non-crossover repair with the parental chromosome, the *ATHILA* has effectively been eliminated. The resulting chromosomes are then subject to selection and genetic drift. An alternative outcome of this pathway is shown on the right hand side. In this case, a meiotic DSB forms on the homolog that lacks the *ATHILA*, followed by resection, ssDNA formation and strand invasion of the homolog that carries an *ATHILA* insertion. In this case, template driven DNA synthesis and non-crossover repair copies and duplicates the *ATHILA*. We propose that this recombination process represents a mechanism to eliminate the *ATHILA*, as although in some situations new copies of *ATHILA* are generated, due to the greater abundance of *CEN180* satellites in the centromeres there is a higher chance overall of this pathway eliminating the transposons.

Table S1. Consensus quality (QV) score of the Col-CEN Arabidopsis genome assembly. Consensus quality scores (QV) were calculated from "missing" 21-mers (k_asm) present in the Col-CEN assembly, but not present in the short read Illumina library. k_total shows the total number of 21-mers. QV scores were calculated for Col-CEN individual chromosomes (green), centromeres (blue), chromosome arms (orange), or the whole genome (yellow).

feature_label	chrom	start	end	k_asm	region_len	k_total	error_rate	QV
Chr1	Chr1	0	32540122	491	32540122	32540102	0.00000071 85324858	61.44
Chr2	Chr2	0	22217084	12938	22217084	22217064	0.00002773 841542	45.57
Chr3	Chr3	0	25743512	956	25743512	25743492	0.00000176 8393137	57.52
Chr4	Chr4	0	21578073	4606	21578073	21578053	0.00001016 568227	49.93
Chr5	Chr5	0	29480885	4525	29480885	29480865	0.00000730 9552866	51.36
CEN1	Chr1	14841109	17559778	37	2718669	2718649	0.00000064 80851998	61.88
CEN2	Chr2	3823791	6045243	26	2221452	2221432	0.00000055 73441539	62.54
CEN3	Chr3	13597187	15733925	368	2136738	2136718	0.00000820 1946565	50.86
CEN4	Chr4	4203901	6977949	4308	2774048	2774028	0.00007400 600007	41.31
CEN5	Chr5	11784130	14551809	1934	2767679	2767659	0.00003328 6578	44.78
Chr1-CEN1	NA	NA	NA	454	29821453	29821433	0.00000072 49552477	61.40
Chr2-CEN2	NA	NA	NA	12912	19995632	19995612	0.00003075 906296	45.12
Chr3-CEN3	NA	NA	NA	588	23606774	23606754	0.00000118 61153	59.26
Chr4-CEN4	NA	NA	NA	298	18804025	18804005	0.00000075 46574935	61.22
Chr5-CEN5	NA	NA	NA	2591	26713206	26713186	0.00000461 8941798	53.35
Whole Genome	NA	NA	NA	23516	NA	131559576	0.00000851 2530102	50.70
All CEN	NA	NA	NA	6673	NA	12618486	0.00002518 859647	45.99
Whole Genome - All CEN	NA	NA	NA	16843	NA	118941090	0.00000674 3688897	51.71

Table S2. TAIR10 gene models that show presence-absence variation (PAV) in Col-CEN. TAIR10 gene models were mapped onto Col-CEN using Liftoff (*64*). Genes that occurred as presence-absence variants (PAVs), as they did not map to Col-CEN, are listed and classified as loci in the CLUSTER_ID column.

ID	CHROM	START (1-based)	END	LEN	PLUS_STRAND	NOTE	CLUSTER_ID
AT1G34800	Chr1	12773164	12773442	279	0	protein_coding_gene	PCG_0 (THIONIN)
AT1G34805	Chr1	12776578	12776856	279	0	protein_coding_gene	PCG_0 (THIONIN)
AT1G34830	Chr1	12793536	12794023	488	0	protein_coding_gene	PCG_0 (THIONIN)
AT1G34840	Chr1	12796984	12797247	264	0	protein_coding_gene	PCG_0 (THIONIN)
AT1G38065	Chr1	14289578	14292060	2483	0	protein_coding_gene	PCG_1
AT1G56820	Chr1	21273314	21273395	82	1	tRNA	TRNA_6
AT1G56910	Chr1	21277861	21277942	82	1	tRNA	TRNA_6
AT1G57030	Chr1	21283986	21284067	82	1	tRNA	TRNA_6
AT1G57210	Chr1	21292992	21293073	82	1	tRNA	TRNA_6
AT1G57240	Chr1	21294341	21294422	82	1	tRNA	TRNA_6
AT1G57300	Chr1	21297221	21297302	82	1	tRNA	TRNA_6
AT1G57330	Chr1	21298753	21298834	82	1	tRNA	TRNA_6
AT1G58808	Chr1	21784645	21786869	2225	1	other_RNA	PCG_2
AT1G58848	Chr1	21791783	21797050	5268	1	protein_coding_gene	PCG_2
AT1G58983	Chr1	21806020	21807487	1468	0	protein_coding_gene	PCG_2
AT1G59030	Chr1	21808193	21809509	1317	0	protein_coding_gene	PCG_2
AT1G59077	Chr1	21810644	21813023	2380	0	protein_coding_gene	PCG_2
AT1G59124	Chr1	21816443	21820572	4130	1	protein_coding_gene	PCG_2
AT1G59312	Chr1	21839858	21841972	2115	1	protein_coding_gene	PCG_2
AT5G36670	Chr5	14401491	14406427	4937	1	protein_coding_gene	PCG_3

AT5G36680	Chr5	14406802	14409137	2336	0	protein_coding_gene	PCG_3
AT5G36690	Chr5	14415185	14417288	2104	0	protein_coding_gene	PCG_3
AT5G36700	Chr5	14421576	14424511	2936	0	protein_coding_gene	PCG_3
AT5G36720	Chr5	14429661	14429924	264	0	protein_coding_gene	PCG_3
AT5G36722	Chr5	14431599	14432216	618	0	protein_coding_gene	PCG_3
AT5G36800	Chr5	14484565	14485409	845	0	protein_coding_gene	PCG_3
AT5G36820	Chr5	14495617	14496849	1233	1	protein_coding_gene	PCG_3

Table S3. TAIR10 gene models that mapped as additional copies to Col-CEN. TAIR10 gene models are listed that mapped via Liftoff to more than one location in Col-CEN (*64*). The CLUSTER_ID column indicates close linkage of the duplicated genes.

TAIR10 ID	TAIR10 CHR	TAIR10 START	TAIR10 LEN	TAIR10 NOTE	COLCEN ID	ColCEN CHR	ColCEN START	Col-CEN LEN	CLUSTER_ID
AT2G16145	Chr2	7008520	80	miRNA	AT2G16145	Chr2	9512451	80	miRNA_1
AT2G16145	Chr2	7008520	80	miRNA	AT2G16145_1	Chr2	9517494	80	miRNA_1
AT1G24822	Chr1	8774997	2886	protein_coding	AT1G24822	Chr1	8780249	2886	PCG_5
AT1G24822	Chr1	8774997	2886	protein_coding	AT1G24822_1	Chr1	8848453	2884	PCG_5
AT1G24909	Chr1	8785785	2130	protein_coding	AT1G24909	Chr1	8830487	2130	PCG_5
AT1G24909	Chr1	8785785	2130	protein_coding	AT1G24909_1	Chr1	8844863	2130	PCG_5
AT1G25141	Chr1	8817678	705	protein_coding	AT1G25141_2	Chr1	8802028	705	PCG_5
AT1G25141	Chr1	8817678	705	protein_coding	AT1G25141_1	Chr1	8839849	705	PCG_5
AT1G25141	Chr1	8817678	705	protein_coding	AT1G25141	Chr1	8854226	705	PCG_5
AT1G25210	Chr1	8833018	2095	protein_coding	AT1G25210_1	Chr1	8772607	2094	PCG_5
AT1G25210	Chr1	8833018	2095	protein_coding	AT1G25210	Chr1	8840811	2093	PCG_5
AT1G59930	Chr1	22061083	399	protein_coding	AT1G59930	Chr1	24161766	399	PCG_6
AT1G59930	Chr1	22061083	399	protein_coding	AT1G59930_1	Chr1	24163425	399	PCG_6
AT1G77932	Chr1	29302725	795	protein_coding	AT1G77932	Chr1	31405241	795	PCG_7
AT1G77932	Chr1	29302725	795	protein_coding	AT1G77932_1	Chr1	31411502	795	PCG_7
AT1G77940	Chr1	29303897	1486	protein_coding	AT1G77940	Chr1	31406413	1486	PCG_7
AT1G77940	Chr1	29303897	1486	protein_coding	AT1G77940_1	Chr1	31412674	1486	PCG_7
AT5G39150	Chr5	15669898	911	protein_coding	AT5G39150	Chr5	18143268	911	PCG_8
AT5G39150	Chr5	15669898	911	protein_coding	AT5G39150_1	Chr5	18156831	911	PCG_8
AT5G39170	Chr5	15680731	595	protein_coding	AT5G39170_1	Chr5	18154101	595	PCG_8
AT5G39170	Chr5	15680731	595	protein_coding	AT5G39170	Chr5	18167662	595	PCG_8

AT5G39190	Chr5	15692591	991	protein_coding	AT5G39190_1	Chr5	18152385	991	PCG_8
AT5G39190	Chr5	15692591	991	protein_coding	AT5G39190	Chr5	18179541	991	PCG_8
AT5G40910	Chr5	16395507	3623	protein_coding	AT5G40910	Chr5	18882480	3623	PCG_9
AT5G40910	Chr5	16395507	3623	protein_coding	AT5G40910_1	Chr5	18887652	3623	PCG_9
ATCG00910	ChrC	100709	72	tRNA	ATCG00910_1	Chr4	8541426	72	tRNA_1
ATCG00910	ChrC	100709	72	tRNA	ATCG00910	ChrC	100709	72	tRNA_1

Table S4. Unique and repeated *CEN180* monomer sequences within and between chromosomes.

CEN180 monomers were compared across the genome to identify unique versus repeated sequences. For repeated sequences we show which chromosomes they occurred on.

Chr	Total	Unique	Repeated	Chr1	Chr2	Chr3	Chr4	Chr5	Chr2, Chr4, Chr5
Chr1	13,578	4,174	Chr1	9,372	0	265	0	2	25
Chr2	12,293	3,887	Chr2		8,363	20	20	7	
Chr3	11,848	3,944	Chr3			7,662	0	7	
Chr4	15,613	4,951	Chr4				10,660	0	
Chr5	12,799	5,484	Chr5					7,287	
All	66,131	22,440	Total	9,372	8,363	7,947	10,680	7,303	43,691

Table S5. *CEN180* higher order repeats. *CEN180* monomers were classified as being the same if they shared 5 or fewer pairwise variants, and consecutive blocks identified as higher order repeats (HORs). HORs are all in a tandem orientation and are classified as being intra- or inter-chromosome. The mean HOR block size, in monomers and bp, and the mean distance between intra-chromosome HORs (bp) are listed.

Chr	Monomers	Intra-chromosome HORs	Inter-chromosome HORs	Mean HOR monomers	Mean HOR block (bp)	Mean HOR distance (bp)
1	13,578	814,715	24,110	2.41	429	365,291
2	12,293	584,684	13,757	2.35	418	434,776
3	11,848	413,642	2,743	2.50	446	334,277
4	15,613	498,876	611	2.40	427	402,170
5	12,799	55,515	0	2.86	509	167,045
All	66,131	2,367,432	41,221	2.41	429	365,291

Table S6. Structural and sequence characteristics of centromeric *ATHILA* retrotransposons. Analysis of 111 gaps greater than 1 kbp in the main *CEN180* arrays identified 53 intact and 20 fragmented *ATHILA* retrotransposons, as well as 12 solo LTRs. For each sequence we report the *ATHILA* subfamily class based on the TAIR10 classification and our phylogenetic analysis, and information on element length, strand, target site duplications (TSDs), long terminal repeat (LTR) position and length, and hits with Hidden Markov Models (HMMs) that describe *GYPSY* LTR retrotransposon open reading frames (see Methods). The 'quality' column indicates whether the *ATHILA* is an 'intact' full-length element, i.e. it contains clearly identified LTRs and, possibly, a TSD; a fragment - note that we also included as fragments and not as intact elements, i) *ATHILA* copies with large internal deletions (e.g. the 4872 bp *ATHILA2* element in centromere 4 has complete and highly similar LTRs but also a ~6 kbp internal deletion), and ii) *ATHILA* copies with a deletion that included the whole LTR plus additional sequence in the internal domain; or a solo LTR. The 'comment' column' includes notes on interesting characteristics for some elements. For example, it highlights the *ATHILA5* duplicates in centromere 5 that contain the internal *CEN180* repeats, and some cases where two intact *ATHILA* of the same subfamily share one LTR (LTR-internal.region-LTR-internal.region-LTR), possibly as a result of post-integration interelement homologous recombination. Given that the LTRs of the *ATHILA6A* and *ATHILA6B* subfamilies appear identical, it was not possible to further allocate solo LTRs of the *ATHILA6* clade into their respective subfamilies. In addition to the *ATHILA* elements, a small number of other TEs were identified but not further analyzed due to their fragmented organization. The majority of these elements occur in centromere 1 and are shown at the end of the Table. Note that for these elements the coordinates refer to the position of the gaps and not the TEs within the gaps. Due to size, Table S6 is attached as a separate file 'Table_S6.xlsx'.

Table S7. Summary of short-read Illumina sequencing libraries aligned to the Col-CEN assembly. All data sets were generated from plants in a Col-0 background, with the exception of the Col×Ler F_2 genomic DNA sequencing libraries that were used to identify meiotic crossovers.

Library	Study accession	Run accession	Read length	Tissue	References
CENH3 ChIP-seq	PRJNA349052	SRR4430537	2×100 bp	Seedling	(*10*)
H3K9me2 ChIP-seq	PRJEB36221	ERR3813867	2×75 bp	Floral bud	(*26*)
H3K27me1 ChIP-seq	PRJEB36221	ERR3813864	2×75 bp	Floral bud	(*26*)
H3K4me1 ChIP-seq	PRJEB36221	ERR3813865	2×75 bp	Floral bud	(*26*)
H3K4me2 ChIP-seq	PRJEB36221	ERR3813866	2×75 bp	Floral bud	(*26*)
H3K4me3 ChIP-seq	PRJEB15183	ERR1590146	2×150 bp	Floral bud	(*28*)
H3K27me3 ChIP-seq	PRJNA252965	SRR1509478	2×100 bp	Floral bud	(*87*)
H2A.W6 ChIP-seq	N/A	N/A	50 bp	Seedling	This study
H2A.W7 ChIP-seq	N/A	N/A	50 bp	Seedling	This study

H2A.Z ChIP-seq	PRJNA219442	SRR988546	50 bp	Leaf	(*24*)
REC8 ChIP-seq	PRJEB36221	ERR3813871	2×75 bp	Floral bud	(*26*)
ASY1 ChIP-seq	PRJEB36320	ERR3829803	2×75 bp	Floral bud	(*27*)
SPO11-1-oligos	PRJEB15185	ERR1590157	50 bp	Floral bud	(*28*)
MNase-seq	PRJEB15184	ERR1590154	2×100 bp	Floral bud	(*28*)
gDNA	PRJEB23842	ERR2215865	2×100 bp	Floral bud	(*28*)
RNA-seq (Col-0 and *met1-3*)	PRJEB9919	ERR966157–ERR966162	2×100 bp	Leaf	(*29*)
Bisulfite-seq (Col-0 and *met1-3*)	PRJEB9919	ERR965674–ERR965677	2×90 bp	Leaf	(*29*)
Bisulfite-seq (Col-0 and mutants)	PRJNA172021, PRJNA222364	SRR534177–SRR869314, SRR1005412–SRR1005415	50 bp	Leaf	(*21*, *22*)
Small RNA-seq (Col-0 and *met1-3*)	PRJEB9919	ERR966148–ERR966149	50 bp	Leaf	(*29*)

Col-0×Ler-0 genomic DNA F_2	E-MTAB-4657 E-MTAB-6577	E-MTAB-4657 E-MTAB-6577	2×150 bp	Leaf	(*88*, *89*)
Hi-C (Col-0)	PRJNA253621	SRR1504819	2×50 bp	Leaf	(*90*)

Table S8. Oligonucleotides. The sequence of oligonucleotides used for *V5-DMC1* construction and genotyping, and FISH, are listed.

Oligo name	Sequence 5´ to 3´	Purpose
Dmc1-PstI-fw	ATATATACTGCAGGATATCAAACATTTACC TGAAAAGA	Cloning *3V5-DMC1*
Dmc1-SphI-rev	ATATATGCATGCTTCTTTTAACTCTTCTCAT	Cloning *3V5-DMC1*
Dmc1-SphI-fw	AAAGAAGCATGCTTAAGCCAACAGAG	Cloning *3V5-DMC1*
Dmc1-NotI-rev	ATATATATATATGCGGCCGCGAGTTTTGCA GCAATTATGAAA	Cloning *3V5-DMC1*
Dmc1-Spe-rev	TATCAAACTAGTGTAAAGTAAACCTTGGTT	Cloning *3V5-DMC1*, genotyping *dmc1-3*
DMC1-Nco-F	TTTCTTTCCATGGATTAAAAAAATTTG	Cloning *3V5-DMC1*
3N-V5-F	GGTAAACCAATCCCAAACCCACTCCTCGGT CTCGACTCAACAGGAAAGCCTATTCCTAAT CCTCTTCTTGGACTTGATTCTACTATGATG GCTTCTCTTAAGTAAGTGA	Cloning *3V5-DMC1*
3N-V5-R	GGGTTTGGGATTGGTTTACCAGTAGAATCA AGTCCAAGAAGAGGATTAGGAATAGGCTT TCCCATTTTCTCGCTCTAAGAGTCTCTA	Cloning *3V5-DMC1*
Dmc1-screen-N-fw	CTCTCACTCTTCCAAGCTTA	Genotyping *3V5-DMC1*

Dmc1-screen-N-rev	AGAGATCAATCACTTACTTAAGAG	Genotyping *3V5-DMC1*
LA27	TAGCATCTGAATTTCATAACCAATCTCGATACAC	Genotyping *dmc1-3*
DMC1-genot-compl-F	CATACATTGACACAGAGGGAACC	Genotyping *dmc1-3* in the presence of *3V5-DMC1*
DMC1-genot-compl-R	ATGGAACCCAAAAGAGGAGAC	Genotyping *dmc1-3* in the presence of *3V5-DMC1*
ATH_cecen180F	CATATTCGACTCCAAAACACTAACC	Amplification of pAL universal *CEN180* probe
ATH_cen180R	AGAAGATACAAAGCCAAAGACTCAT	Amplification of pAL universal *CEN180* probe
CEN180-α	CCGCAACAGGATCTTAAAGGCGTAAGAATTTTATTCTGTTAAAAGACACAAAGCCAAAGA	*CEN180* FISH probe
CEN180-β	ATTGAATCTTTGTTAGAAGATACAAAGACAAAGACTCATACGGACTTCGACTACACTATC	*CEN180* FISH probe
CEN180-γ	TTAAACTGCAATTGGATCTTAAAGGCGTAAGAATTGTATCCTTGTTAAAAGACACAAAGC	*CEN180* FISH probe

CEN180-δ	CGCATCTTATAAGCCTAAGTAGTATTTCCTTGTTAGAATACACAAAGTCAAAGACTCATA	*CEN180* FISH probe
CEN180-ε	TCTTATAAGCCTAAGTAGTGTTTCCTTGTTAGAAGACACAAAGCCAATGACTCATATCGC	*CEN180* FISH probe
ATHILA2_GAG_F	GGATCCACTCGACCACCTTG	Amplification of the *ATHILA2* FISH probe
ATHILA2_GAG_R	AACCCTTGAAACGCTCCCAT	Amplification of the *ATHILA2* FISH probe
ATHILA6A6B_GAG_F	GATCCACTCGATCACCTGGAC	Amplification of the *ATHILA6A/6B* FISH probe
ATHILA6A6B_GAG_R	TCCCATGCTTCGCAGAAAGT	Amplification of the *ATHILA6A/6B* FISH probe

Supplementary References:

42. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* (2019), doi:10.1038/s41587-019-0072-8.

43. S. Sato, Y. Nakamura, T. Kaneko, E. Asamizu, S. Tabata, Complete structure of the chloroplast genome of Arabidopsis thaliana. *DNA Res.* **6**, 283–290 (1999).

44. D. B. Sloan, Z. Wu, J. Sharbrough, Correction of Persistent Errors in Arabidopsis Reference Mitochondrial Genomes. *Plant Cell*. **30**, 525–527 (2018).

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

46. A. Morgulis, E. M. Gertz, A. A. Schäffer, R. Agarwala, WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*. **22**, 134–141 (2006).

47. A. Morgulis, G. Coulouris, Y. Raytselis, T. L. Madden, R. Agarwala, A. A. Schäffer, Database indexing for production MegaBLAST searches. *Bioinformatics*. **24**, 1757–1764 (2008).

48. D. Guan, S. A. McCarthy, J. Wood, K. Howe, Y. Wang, R. Durbin, Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. **36**, 2896–2898 (2020).

49. S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L. Salzberg, Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).

50. M. Alonge, S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin, F. J. Sedlazeck, Z. B. Lippman, M. C. Schatz, RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).

51. R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, M. A. DePristo, A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

52. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio.GN]* (2013), (available at http://arxiv.org/abs/1303.3997).

53. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

54. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools. *Gigascience*. **10** (2021), doi:10.1093/gigascience/giab008.

55. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

56. C. Jain, A. Rhie, H. Zhang, C. Chu, B. P. Walenz, S. Koren, A. M. Phillippy, Weighted minimizer sampling improves long read mapping. *Bioinformatics*. **36**, i111–i118 (2020).

57. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

58. F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, M. C. Schatz, Accurate detection of complex structural variations using single-molecule sequencing.

Nat. Methods. **15**, 461–468 (2018).

59. M. Alonge et al, Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*. **182**, 145–161.e23 (2020).

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

61. G. Formenti, A. Rhie, B. P. Walenz, F. Thibaud-Nissen, K. Shafin, S. Koren, E. W. Myers, E. D. Jarvis, A. M. Phillippy, Merfin: improved variant filtering and polishing via k-mer validation, , doi:10.1101/2021.07.16.452324.

62. H. Li, R. Durbin, Fast and accurate sort read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

63. N. Abdennur, L. A. Mirny, Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*. **36**, 311–316 (2020).

64. A. Shumate, S. L. Salzberg, Liftoff: accurate mapping of gene annotations. *Bioinformatics* (2020), doi:10.1093/bioinformatics/btaa1016.

65. S. Ou, W. Su, Y. Liao, K. Chougule, J. R. A. Agda, A. J. Hellinga, C. S. B. Lugo, T. A. Elliott, D. Ware, T. Peterson, N. Jiang, C. N. Hirsch, M. B. Hufford, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

66. N. Buisine, H. Quesneville, V. Colot, Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*. **91**, 467–475 (2008).

67. K. D. Yamada, K. Tomii, K. Katoh, Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees. *Bioinformatics*. **32**, 3246–3251 (2016).

68. D. Ellinghaus, S. Kurtz, U. Willhoeft, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*. **9**, 18 (2008).

69. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

70. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

71. K. Liu, C. R. Linder, T. Warnow, RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One*. **6**, e27731 (2011).

72. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).

73. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 357–359 (2012).

74. F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, T. Manke, deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–91 (2014).

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

76. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

77. B. Bushnell, R. Egan, A. Copeland, B. Foster, A. Clum, H. Sun, Others, BBMap: a fast, accurate, splice-aware aligner. 2014. *Available: sourceforge. net/projects/bbmap* (2019).

78. F. Krueger, Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. https://github.com/FelixKrueger/TrimGalore

79. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. **27**, 1571–1572 (2011).

80. B. A. Rowan, V. Patel, D. Weigel, K. Schneeberger, Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3* . **5**, 385–398 (2015).

81. T. Mandáková, M. A. Lysak, Chromosome Preparation for Cytogenetic Analyses in Arabidopsis. *Curr Protoc Plant Biol*. **1**, 43–51 (2016).

82. J. W. Ijdo, R. A. Wells, A. Baldini, S. T. Reeders, Improved telomere detection using a telomere repeat probe (TTAGGG)n generated by PCR. *Nucleic Acids Res.* **19**, 4780 (1991).

83. K. Nagaki, P. B. Talbert, C. X. Zhong, R. Kelly Dawe, S. Henikoff, J. Jiang, Chromatin Immunoprecipitation Reveals That the 180-bp Satellite Repeat Is the Key Functional DNA Element of Arabidopsis thaliana Centromeres. *Genetics*. **163** (2003), pp. 1221–1225.

84. M. Ravi, S. W. L. Chan, Haploid plants produced by centromere-mediated genome elimination. *Nature*. **464** (2010), pp. 615–618.

85. S. J. Armstrong, G. H. Jones, Meiotic cytology and chromosome behaviour in wild-type Arabidopsis thaliana. *Journal of Experimental Botany*. **54** (2003), pp. 1–10.

86. J. D. Higgins, E. Sanchez-Moran, S. J. Armstrong, G. H. Jones, F. C. H. Franklin, The Arabidopsis synaptonemal complex protein ZYP1 is required for chromosome synapsis and normal fidelity of crossing over. *Genes Dev.* **19**, 2488–2500 (2005).

87. B. Zhu, W. Zhang, T. Zhang, B. Liu, J. Jiang, Genome-Wide Prediction and Validation of Intergenic Enhancers in Arabidopsis Using Open Chromatin Signatures. *Plant Cell*. **27**, 2415–2426 (2015).

88. H. Serra, C. Lambing, C. H. Griffin, S. D. Topp, D. C. Nageswaran, C. J. Underwood, P. A. Ziolkowski, M. Séguéla-Arnaud, J. B. Fernandes, R. Mercier, I. R. Henderson, Massive crossover elevation via combination of HEI10 and recq4a recq4b during Arabidopsis meiosis. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2437–2442 (2018).

89. C. J. Underwood, K. Choi, C. Lambing, X. Zhao, H. Serra, F. Borges, J. Simorowski, E. Ernst, Y. Jacob, I. R. Henderson, R. A. Martienssen, Epigenetic activation of meiotic recombination near Arabidopsis thaliana centromeres via loss of H3K9me2 and non-CG DNA methylation. *Genome Res.* **28**, 519–531 (2018).

90. Genome-wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in Arabidopsis. *Mol. Cell*. **55**, 694–707 (2014).

91. N. Kumekawa, T. Hosouchi, H. Tsuruoka, H. Kotani, The size and sequence organization of the centromeric region of arabidopsis thaliana chromosome 5. *DNA Res.* **7**, 315–321 (2000).

92. N. Kumekawa, T. Hosouchi, H. Tsuruoka, H. Kotani, The size and sequence organization of the

centromeric region of Arabidopsis thaliana chromosome 4. *DNA Res.* **8**, 285–290 (2001).

93. C. Pockrandt, M. Alzamel, C. S. Iliopoulos, K. Reinert, GenMap: ultra-fast computation of genome mappability. *Bioinformatics*. **36**, 3687–3692 (2020).

94. J. M. Keith, *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution* (Humana, 2018).