Article

# Cycles of satellite and transposon evolution in *Arabidopsis* centromeres

https://doi.org/10.1038/s41586-023-06062-z

Received: 17 November 2022

Accepted: 6 April 2023

Published online: 17 May 2023

Check for updates

Piotr Wlodzimierz[1,9], Fernando A. Rabanal[2,9], Robin Burns[1,9], Matthew Naish[1], Elias Primetis[3], Alison Scott[4], Terezie Mandáková[5], Nicola Gorringe[1], Andrew J. Tock[1], Daniel Holland[1], Katrin Fritschi[2], Anette Habring[2], Christa Lanz[2], Christie Patel[1], Theresa Schlegel[2], Maximilian Collenberg[2], Miriam Mielke[2], Magnus Nordborg[6], Fabrice Roux[7], Gautam Shirsekar[2], Carlos Alonso-Blanco[8], Martin A. Lysak[5], Polina Y. Novikova[4], Alexandros Bousios[3]✉, Detlef Weigel[2]✉ & Ian R. Henderson[1]✉

Centromeres are critical for cell division, loading CENH3 or CENPA histone variant nucleosomes, directing kinetochore formation and allowing chromosome segregation[1,2]. Despite their conserved function, centromere size and structure are diverse across species. To understand this centromere paradox[3,4], it is necessary to know how centromeric diversity is generated and whether it reflects ancient trans-species variation or, instead, rapid post-speciation divergence. To address these questions, we assembled 346 centromeres from 66 *Arabidopsis thaliana* and 2 *Arabidopsis lyrata* accessions, which exhibited a remarkable degree of intra- and inter-species diversity. *A. thaliana* centromere repeat arrays are embedded in linkage blocks, despite ongoing internal satellite turnover, consistent with roles for unidirectional gene conversion or unequal crossover between sister chromatids in sequence diversification. Additionally, centrophilic *ATHILA* transposons have recently invaded the satellite arrays. To counter *ATHILA* invasion, chromosome-specific bursts of satellite homogenization generate higher-order repeats and purge transposons, in line with cycles of repeat evolution. Centromeric sequence changes are even more extreme in comparison between *A. thaliana* and *A. lyrata*. Together, our findings identify rapid cycles of transposon invasion and purging through satellite homogenization, which drive centromere evolution and ultimately contribute to speciation.

Until recently, it has been challenging to assemble plant and animal centromeres because they are frequently composed of complex tandem repeat arrays that are the site of kinetochore loading[3,5]. Advances in long-read DNA sequencing, including PacBio High Fidelity (HiFi) and Oxford Nanopore Technology (ONT) sequencing, have now made complex centromere assembly possible, allowing complete telomere-to-telomere reference maps of humans and *Arabidopsis thaliana*[6–8]. However, how intra-species centromeric diversity relates to rapid inter-species evolution is unknown. To investigate centromere evolution on multiple scales, we studied the plant *A. thaliana*, which has megabase-scale satellite arrays that resemble human α-satellite centromeres[6,8,9]. *A. thaliana* is widely distributed across Eurasia, and genetic and epigenetic diversity within the chromosome arms is well documented, with most extant populations resulting from post-glacial spread ~10,000 years ago[10–12]. Refugia 'relict' populations also exist, as do populations in Africa, where the species originated[10–12]. Here we report centromere diversity within *A. thaliana* and compare this with centromeres from the sister species *Arabidopsis lyrata*, which has a sparser, circumpolar distribution[13], to contrast patterns of intra- and inter-species centromere evolution.

## Assembling the *Arabidopsis* pan-centromere

We assembled the genomes of 66 diverse accessions from across the *A. thaliana* native range, with 27–212× depth of PacBio HiFi reads per accession and a read N50 of 12.7–24.4 kb[7,14,15] (Fig. 1a and Supplementary Table 1). Almost all assembled centromere satellite arrays (327/330) were gapless (Supplementary Table 1). For validation, we aligned HiFi reads to their corresponding genomes and quantified primary and secondary allele coverage, which indicated that very few regions were collapsed during assembly (Extended Data Fig. 1 and Supplementary Table 1). Principal-component analysis (PCA) of single-nucleotide polymorphisms (SNPs) in the chromosome arms confirmed the presence of four major genetic groups: Eurasian non-relicts, Iberian non-relicts, Iberian relicts and non-Iberian relicts (Fig. 1a,b and Supplementary Tables 1 and 2). Notably, three pairs of accessions had nearly identical chromosome-arm SNPs,

[1]Department of Plant Sciences, University of Cambridge, Cambridge, UK. [2]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Tübingen, Germany. [3]School of Life Sciences, University of Sussex, Brighton, UK. [4]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany. [5]Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [6]Gregor Mendel Institute, Vienna, Austrian Academy of Sciences, Vienna BioCenter, Vienna, Austria. [7]LIPME, INRAE, CNRS, Université de Toulouse, Castanet-Tolosan, France. [8]Departamento de Genética Molecular de Plantas, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Madrid, Spain. [9]These authors contributed equally: Piotr Wlodzimierz, Fernando A. Rabanal, Robin Burns. ✉e-mail: alex.bousios@sussex.ac.uk; weigel@tue.mpg.de; irh25@cam.ac.uk

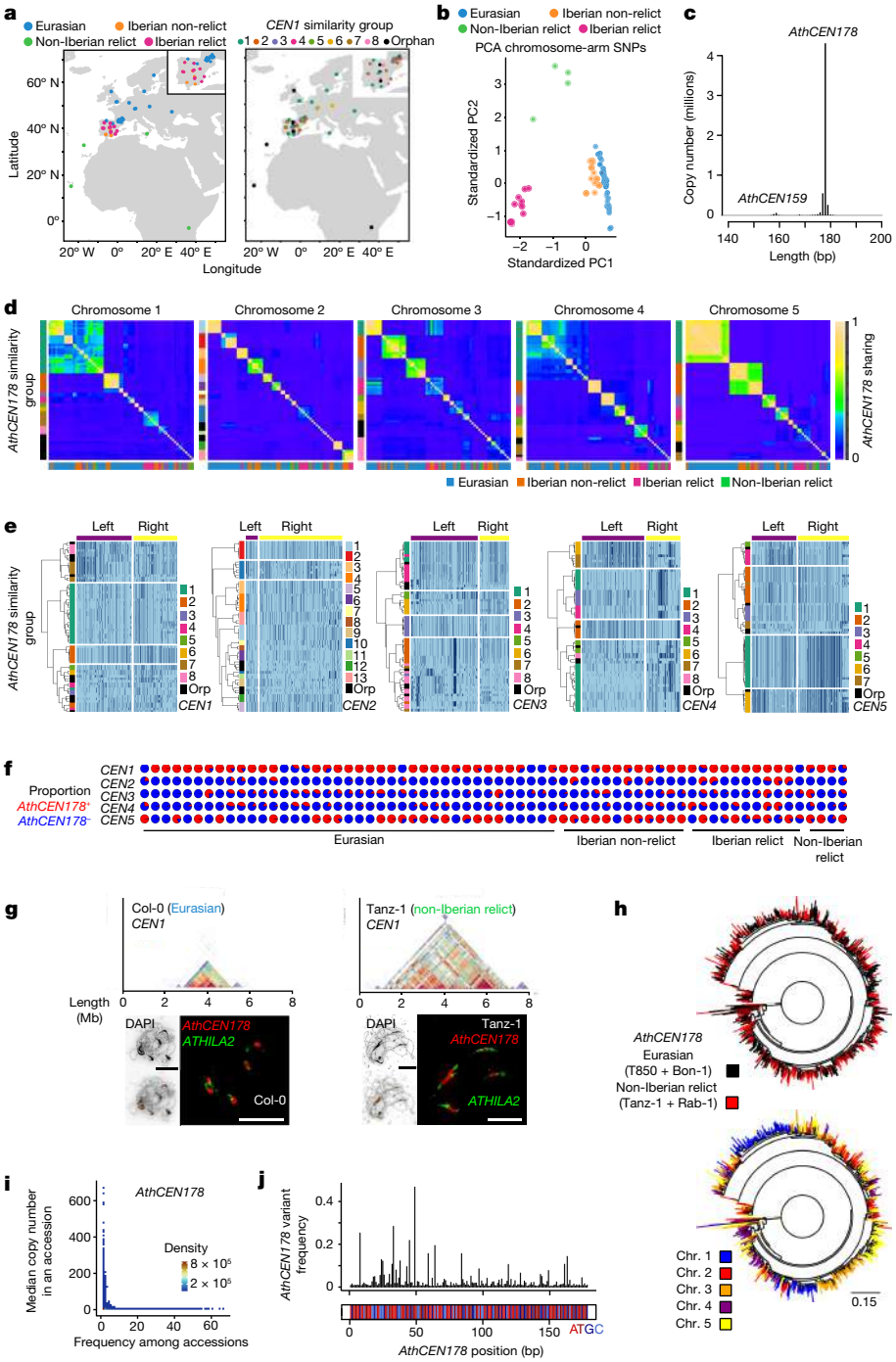



**Fig. 1 | High genetic diversity in the *Arabidopsis* pan-centromere.** **a**, Geographic origin of 66 *A. thaliana* accessions, coloured according to PCA group membership, as shown in **b**. The geographic origin of the *CEN1 AthCEN178* similarity groups is shown to the right. **b**, PCA using chromosome-arm SNPs, highlighting the Eurasian (blue), Iberian non-relict (orange), non-Iberian relict (green) and Iberian relict (pink) accessions. **c**, Satellite repeat length across 66 accessions. **d**, Heat maps showing the percentage of identical *AthCEN178* sequences shared by pairs of chromosomes, with the colour scale on the right. The *x*-axis key shows PCA membership, as in **b**. The *y*-axis key indicates the *AthCEN178* similarity group. **e**, Genetic polymorphisms were plotted in 800-kb regions flanking the Col-0 *AthCEN178* arrays and clustered using Hamming distance. Dendrogram tips are coloured according to *AthCEN178* similarity group, as in **d**. **f**, Pie charts showing *AthCEN178* proportion on the forward (red, +) or reverse (blue, –) strand for each chromosome (rows) and accession (columns). **g**, StainedGlass sequence identity heat maps for Col-0 (Eurasian) and Tanz-1 (relict) *CEN1*. Representative FISH micrographs of pachytene chromosomes of Col-0 and Tanz-1 probed for *AthCEN178* (red) and *ATHILA2* (green) are shown below. For each sample, 50 independent micrographs were analysed. Insets are DAPI-stained images and overlays, of the same cells. Scale bars, 10 µm. **h**, Maximum-likelihood *AthCEN178* phylogenetic tree, sampled from Eurasian (T850 and Bon-1) and non-Iberian relict (Tanz-1 and Rab-1) accessions and rooted using *Capsella rubella* satellites[44]. The same tree is shown with branches coloured by PCA group (top) or by chromosome (bottom). Scale bar, 0.15 substitutions per site. **i**, Median *AthCEN178* copy number per accession plotted against the number of accessions in which they were found. **j**, Frequency of *AthCEN178* sequence variants across the consensus repeat. The colour-coded DNA sequence is shown below (red, A; dark red, T; navy, G; azure, C).

providing an exceptional opportunity to study short-term centromere evolution in natural populations (Supplementary Table 2).

## *Arabidopsis* centromeric repeat diversity

Across our collection of centromeres, we identified 5,345,259 copies of the ~178-bp centromere satellite repeat (*AthCEN178*, formerly *CEN180*), in addition to 137,520 copies of a shorter 159-bp repeat (*AthCEN159*, formerly *CEN160*) (Fig. 1c and Supplementary Table 2). Individual *AthCEN178* sequences were diverse, with almost one quarter (1,234,281) unique in our sample. The number of non-redundant *AthCEN178* sequences began to plateau as more accessions were included, in line with deep sampling of the satellome (Extended Data Fig. 2a). The level of pair-wise *AthCEN178* sequence sharing was considerably higher when considering the same chromosome across accessions (8.02–15.13%), as opposed to different chromosomes (0.01–0.36%), either within or across accessions (Extended Data Fig. 2b). On the basis of pair-wise sharing of *AthCEN178* sequences, we identified discrete similarity groups for each chromosome separately (Fig. 1d and Supplementary Table 2). A high diversity of *AthCEN178* similarity groups was observed that did not geographically relate with the chromosome-arm SNP groups, although satellite sharing decreased with increasing geographic separation ($r = -0.28$, $P \leq 2.2 \times 10^{-16}$) (Fig. 1a and Extended Data Fig. 2c,d). Sharing of *AthCEN178* similarity groups across all five chromosomes was very rare (Extended Data Fig. 2e), in line with independent segregation of centromeres after outcrossing. Genetic variation surrounding the satellite arrays indicated that centromeres belonging to the same *AthCEN178* similarity group are often embedded in centromere-spanning haplotype blocks (Fig. 1e). These 'cenhaps' are consistent with linkage disequilibrium expectations in regions of low meiotic crossover frequency, as observed around human centromeres[16].

The centromeric *AthCEN178* satellite arrays spanned on average 2.91 Mb per chromosome (range, 1.11–6.49 Mb), with extremely high structural diversity within and between chromosomes (Supplementary Table 3 and Supplementary Video 1). While each chromosome exhibited *AthCEN178* strand biases, there were complete strand reversals and inversions, especially on chromosome 5 (Fig. 1f). The SNP-based PCA genetic groups differed significantly in satellite array size, with total *AthCEN178* content higher in the relicts than in the Eurasian accessions (Wilcoxon test, $P = 1.17 \times 10^{-4}$) (Extended Data Fig. 3a and Supplementary Table 2). We independently confirmed this difference using fluorescence in situ hybridization (FISH), observing a significantly greater *AthCEN178* centromeric signal in Tanz-1 (a relict) than in the Col-0 (Eurasian) accession (Wilcoxon test, all $P \leq 1.04 \times 10^{-6}$) (Fig. 1g and Extended Data Fig. 3b,c). *AthCEN178* repeats sampled from contrasting Eurasian and relict backgrounds exhibited stronger phylogenetic clustering by chromosome than by accession (Fig. 1h), in line with limited sequence exchange between chromosomes. Furthermore, *AthCEN178* sequences that were unique to a single accession were frequently at high copy number within their chromosome, whereas *AthCEN178* sequences shared across accessions were typically rare within their genome (Fig. 1i). This is consistent with high turnover and rapid expansion of new satellite variants within chromosomes, causing concerted evolution[17]. We examined the frequency of *AthCEN178* sequence variants along the consensus repeat and observed a marked elevation of variants within the first 50 nucleotides (Fig. 1j). We performed CENH3 chromatin immunoprecipitation and sequencing (ChIP–seq) in the Col-0, Ler-0, Cvi-0 and Tanz-1 accessions and observed that the highest *AthCEN178* variation corresponded to the left edge of CENH3 occupancy within the repeat, in line with an effect of chromatin structure on the mutation rate or persistence of variants (Extended Data Fig. 3e).

## Dynamics of satellite repeat evolution

An important mode of centromere satellite evolution is through tandem duplication, which leads to higher-order repeats (HORs) in plant and vertebrate centromere arrays[5,6,18]. Therefore, we examined *AthCEN178* HORs, defined as tandem repeat duplications of at least three consecutive satellite copies with no more than five substitutions or single-base insertions/deletions for each monomer pair. This approach can assign a single *AthCEN178* sequence to multiple HORs. *AthCEN178* HORs were abundant (total of 98,265,546) in our collection of centromere arrays, with a mean of 1,488,872 per accession (Fig. 2a and Supplementary Tables 2 and 4). HOR lengths showed a negative exponential distribution, with 3-mers (~534 bp) occurring most frequently (Fig. 2a), whose length is within the observed range of meiotic gene conversion in *A. thaliana*[19]. The majority of *AthCEN178* HORs (66.7%) were <0.5 Mb apart (median, 279 kb), although many were more widely spaced (maximum, 7.5 Mb) (Fig. 2a), showing that satellite recombination can also occur over large physical distances. For each chromosome, we derived an *AthCEN178* consensus and calculated each repeat's edit distance from this consensus. For each satellite monomer, we also counted the number of HORs to which it belonged and divided by the total number of *AthCEN178* copies on that chromosome, yielding a HOR score. Across all 330 centromeres, the core *AthCEN178* arrays (between 0.2 and 0.8 of the centromere scaled length) had the greatest HOR scores and lowest edit distances (Fig. 2b and Extended Data Fig. 4a). Within chromosomes, single-copy *AtCEN178* sequences were more frequent at the periphery of the satellite arrays, whereas the core regions were composed of high-copy *AtCEN178* sequences (Extended Data Fig. 4b). This indicates that satellite recombination and HOR formation are frequent, yet spatially focused, within the core of the *A. thaliana* centromere arrays. For example, a ~1-Mb region within the *AthCEN178* array of Ey15-2 chromosome 1 (*CEN1*) featured elevated HOR scores and decreased edit distances (Fig. 2c).

Across our *A. thaliana* sample, chromosomes had heterogeneous *AthCEN178* HOR scores (Extended Data Fig. 4c,d), in line with homogenization occurring independently in different arrays. To detect recent centromere evolution, we compared satellite arrays within *AthCEN178* similarity groups that were embedded in the same cenhap linkage block (Figs. 1d and 2d, Extended Data Fig. 4e and Supplementary Table 2). For example, we considered *CEN1* similarity group 2, which included seven French and two Iberian accessions (Figs. 1d and 2d and Supplementary Table 2). Although flanking sequences indicated membership in a single cenhap (Fig. 1e), the French and Spanish centromeres were differentiated by internal satellite polymorphism (Fig. 2d). For example, a ~1-Mb *AthCEN178* array in BARC-A-17 was related to a smaller ~150-kb array in IP-Ini-0, yet surrounded by highly similar regions (Figs. 1e and 2d and Extended Data Fig. 4f). These accessions also shared a single syntenic *ATHILA* element (10,559 bp) in *CEN1*, located close to the polymorphic region, with only six mutations separating the copies, suggesting a recent common origin. The extensive *AthCEN178* HOR similarities flanking the polymorphic region (Fig. 2d and Extended Data Fig. 4f) indicate that the satellite array has either recently expanded in BARC-A-17 or contracted in IP-Ini-0, despite flanking SNPs being maintained in linkage (Fig. 1e). Similar patterns of internal centromere satellite dynamics, which otherwise occurred embedded in cenhap linkage blocks, were seen for all chromosomes (Extended Data Fig. 4d). Together, these findings are consistent with unidirectional gene conversion or unequal crossover between sister chromatids mediating satellite evolution[20–22].

Although most *AthCEN178* HORs were short (<10 kb) (Fig. 2a), less frequent large-scale (>100 kb) duplications were also observed (Fig. 2e,f). To identify large intra-centromere *AthCEN178* duplications, we screened for HORs that involved >5 monomers, allowing no monomer variants. This identified 94 intra-array duplications that were over 40 kb long (mean length, 367 kb, or ~2,062 *AthCEN178* copies) and present on all chromosomes (Fig. 2e). For example, *CEN5* in MERE-A-13 contained multiple large duplications, including a triplication (Fig. 2f and Extended Data Fig. 4g). This indicates that *A. thaliana AthCEN178* recombination and HOR formation operate at a range of scales, with heterogeneity in rate within and between centromeres.

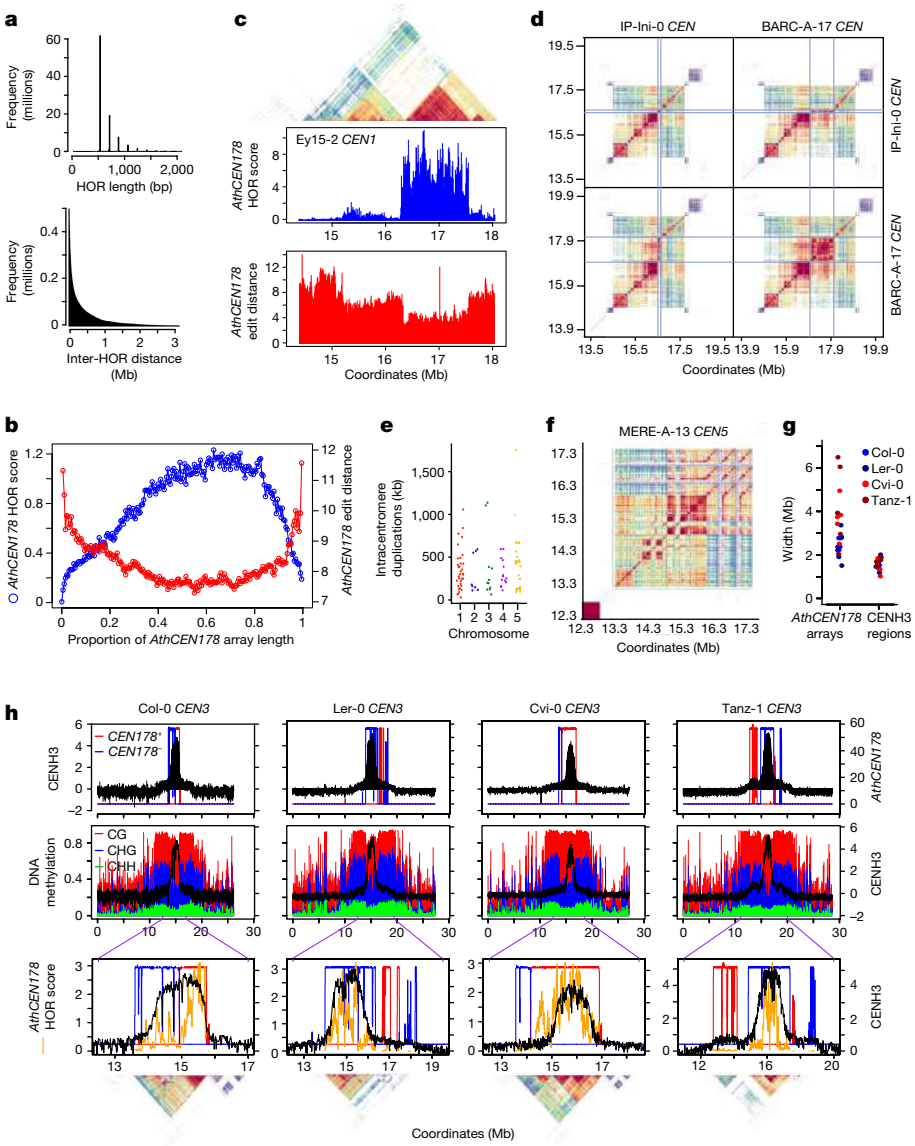



**Fig. 2 | Dynamic genetic and epigenetic evolution of the *Arabidopsis* centromere arrays. a**, Histograms of *AthCEN178* HOR lengths (top) and inter-HOR distances (bottom). **b**, Plot of *AthCEN178* HOR scores (blue) and edit distances (red) across the scaled length of all satellite arrays. **c**, Comparison of *AthCEN178* HOR scores (blue) with edit distances (red) within Ey15-2 *CEN1*. A StainedGlass sequence identity heat map is shown above. **d**, StainedGlass sequence identity heat maps showing sequence identity within and between IP-Ini-0 (Spanish) and BARC-A-17 (French) *CEN1*. Blue lines demarcate a ~1-Mb BARC-A-17 region that is similar to a ~100-kb IP-Ini-0 region. **e**, Intra-centromere duplication sizes per chromosome. **f**, StainedGlass sequence identity heat map of MERE-A-13 *CEN5*. **g**, Size comparisons of *AthCEN178* arrays and CENH3-enriched regions for each chromosome (blue, Col-0; dark blue, Ler-0; red, Cvi-0; dark red, Tanz-1). *AthCEN178* array sizes were significantly greater than those of the CENH3 regions (Wilcoxon test, $P = 2.95 \times 10^{-7}$). **h**, Top, CENH3 ChIP–seq enrichment ($\log_2$(ChIP/input), black) compared to *AthCEN178* density in 10-kb windows on the forward (red) or reverse (blue) strand along chromosome 3 of Col-0, Ler-0, Cvi-0 and Tanz-1. Middle, CENH3 ChIP–seq enrichment (black) plotted against DNA methylation in CG (red), CHG (blue) and CHH (green) sequence contexts, along the whole chromosome. Bottom, close-up views of the centromere regions with *AthCEN178* density (red, blue), CENH3 (black) and HOR scores (orange) plotted. StainedGlass sequence identity heat maps are shown beneath.

As CENH3 defines the functional centromere[1,4], we investigated how CENH3 ChIP–seq enrichment related to the underlying *AthCEN178* satellite arrays in diverse Eurasian and relict accessions (Fig. 2h and Extended Data Fig. 5). We observed, across the Eurasian Col-0 and Ler-0 and relict Cvi-0 and Tanz-1 accessions, that *AthCEN178* array sizes (~1.5–6.5 Mb) were significantly greater than those of the regions of CENH3 ChIP–seq enrichment (~1–2 Mb) (Wilcoxon test, $P = 2.95 \times 10^{-7}$) (Fig. 2g). This suggests homeostasis of CENH3 loading within *AthCEN178* satellite arrays of varying size. At ten centromeres, CENH3 preferentially occupied only one of the distinct *AthCEN178* arrays present (Fig. 2h and Extended Data Fig. 5), in line with an effect of satellite DNA sequence on CENH3 loading. At four centromeres, CENH3 occupied more than one distinct *AthCEN178* array. At the remaining six centromeres, a single *AthCEN178* array was present and occupied by CENH3 (Fig. 2h and Extended Data Fig. 5). Protein diversity in the centromeric histone CENH3 was limited within our *A. thaliana* sample, with the 66 accessions containing only five distinct protein variants, which were not strongly correlated with the *AthCEN178* centromere similarity groups (Extended Data Fig. 6). We additionally profiled centromeric DNA methylation from ONT sequencing reads for the same accessions, at 220–460× depth and with a read N50 of 6.3–26.7 kb (Supplementary Table 1). Despite satellite structural diversity, all CENH3-enriched regions were severely depleted of DNA methylation in the CHG context compared with flanking pericentromeric heterochromatin, with a more modest decrease in the CG context

(Fig. 2h and Extended Data Fig. 5). Together, these data are consistent with complex interactions between genetic and epigenetic information and CENH3 enrichment within the *A. thaliana* centromeres.

## Satellite invasion by *ATHILA* transposons

Assembly of the Col-0 reference strain showed the presence of *ATHILA* retrotransposons within the centromere satellite repeat arrays[6]. Across the accessions studied here, the majority (90%) of non-satellite centromeric sequence was composed of *ATHILA* elements. In total, we detected 9,250 intact *ATHILA* elements and 13,556 soloLTRs, of which 1,357 and 549, respectively, were located within the *AthCEN178* satellite arrays, accounting for 1.7% of their sequence (Supplementary Tables 2 and 5). The intact-to-soloLTR ratio was substantially higher inside the satellite arrays than outside (2.5 versus 0.6) (Supplementary Table 5), in line with increased *ATHILA* integration or reduced soloLTR formation. Of the 14 identified *ATHILA* families, only 5 were frequently observed within the *AthCEN178* arrays, especially *CEN5*, with *ATHILA5* being the most 'centrophilic' (Fig. 3a,b, Extended Data Fig. 7a,b and Supplementary Table 5). The *ATHILA* elements within the *AthCEN178* arrays were significantly younger than those in the chromosome arms (Wilcoxon test, all $P < 1.57 \times 10^{-8}$), with *ATHILA5* elements being the youngest (Fig. 3c and Supplementary Table 5). Using FISH against *ATHILA2* and *ATHILA5*, we confirmed that *ATHILA5* probes produced the most central signal within the DAPI-dense chromocentres (Fig. 3d and Extended Data Fig. 7c). Using the sequences flanking *ATHILA* elements, we mapped the location of 1,367 insertions within the *AthCEN178* consensus repeat. The *ATHILA* elements have integrated throughout the length of the *AthCEN178* consensus, although the highest levels of integration were observed between positions 50 and 100 at the start of the repeat (Extended Data Fig. 7d). The centrophilic *ATHILA* families showed distinct biases in their insertion preferences for different centromeres, but not for different chromosome arms (Extended Data Fig. 7a,b), which may arise at the level of integration or differential post-integration maintenance of *ATHILA* elements within satellite arrays. Of the intact *ATHILA* elements within the *AthCEN178* arrays, 239 encoded all major retrotransposon proteins and the *ATHILA*-specific open reading frame (ORF)[23] (Supplementary Table 5). These *ATHILA* elements are likely to be competent for autonomous transposition and were over-represented in the satellite repeat arrays (21.8% inside versus 12.7% outside). Together, these patterns indicate widespread colonization of the *A. thaliana* satellite arrays by centrophilic *ATHILA*.

In addition to possessing smaller *AthCEN178* satellite arrays than the relicts, the Eurasian accessions displayed a greater level of centrophilic *ATHILA* integration (Wilcoxon test, $P = 1.46 \times 10^{-7}$) (Extended Data Fig. 7e). Furthermore, intact *ATHILA* elements within the *AthCEN178* arrays had significantly higher long terminal repeat (LTR) identity in the Eurasian and Iberian non-relict accessions, compared with the Iberian and non-Iberian relict accessions (Wilcoxon tests, all $P < 1.78 \times 10^{-6}$) (Extended Data Fig. 7f), indicating that there may have been recent transposition activity in Eurasia. To identify recent *ATHILA* invasions, we compared pairs of accessions from the same PCA genetic group with very similar *AthCEN178* repeat arrays. A notable example is provided by the ANGE-B-2 and ANGE-B-10 accessions, which were collected within 5 m of one another in France. These accessions contained a large number of *ATHILA* insertions in *CEN4* and *CEN5* and had very similar *AthCEN178* arrays (Fig. 3e, Extended Data Fig. 8a and Supplementary Tables 2 and 5). All 19 *ATHILA* insertions within the ANGE-B-2 *CEN4* and *CEN5 AthCEN178* arrays could be unambiguously matched to syntenic insertions in ANGE-B-10, on the basis of transposon length, target site duplications (TSDs) and family membership (Fig. 3e, Extended Data Fig. 8a and Supplementary Table 6). In addition, ANGE-B-10 contained 30 unique *ATHILA5* insertions, the majority of which (66.7%) had identical LTRs, in line with very recent integration (Fig. 3e, Extended Data Fig. 8a and Supplementary Tables 2 and 6). Few mutations separated the shared *ATHILA* insertions in the ANGE accession pair (1 difference per 2,000 bp over 172 kb of syntenic sequence), suggesting a recent common ancestry for *CEN4* and *CEN5 ATHILA* insertions (Supplementary Table 6). The new ANGE-B-10 *ATHILA5* elements included non-autonomous derivatives that harboured a ~3-kb deletion of the reverse transcriptase, RNase H and integrase genes (Extended Data Fig. 9). Phylogenetic analysis showed that both the autonomous and non-autonomous *ATHILA5* sub-lineages were present throughout the chromosome-arm PCA groups, reflecting dynamic radiation (Fig. 3f). Our data provide evidence for recent *ATHILA5* invasion of Eurasian *A. thaliana* centromere satellite arrays.

We also observed transposon polymorphism originating through large-scale intra-centromere duplication of the satellite arrays. For example, a ~2-Mb duplication, which copied three intact *ATHILA* elements and one soloLTR, distinguished *CEN5* of FERR-A-8 and FERR-A-12 (Extended Data Fig. 8b–d and Supplementary Table 6). Despite the internal *AthCEN178* duplication, these satellite arrays belong to the same cenhap block, defined by flanking polymorphisms (Fig. 1e), in line with either unequal crossover between sister chromatids or unidirectional gene conversion generating the polymorphism. This analysis showed that *ATHILA* copies may increase by transposition and also through intra-satellite array duplications. However, as the *ATHILA* elements represent a minority of the sequence (~1.7%) within the *AthCEN178* arrays, we speculate that intra-centromere recombination events will more often act to eliminate *ATHILA*. If this were the case, we would expect regions with newly expanded *AthCEN178* HORs to be less likely to harbour *ATHILA*. In agreement with this hypothesis, permutation tests confirmed that *ATHILA* elements were located in regions with significantly lower HOR scores and higher *AthCEN178* divergence ($P < 0.001$) (Fig. 3g). The only exception was the association of *ATHILA5* with reduced HOR scores, which was not significant ($P = 0.128$) (Fig. 3g). As *ATHILA5* is the youngest and most centrophilic family, insufficient time may have occurred for satellite homogenization to have removed the new copies, as seen in ANGE-B-10 (Fig. 3e). Consistent with this, the unique *ATHILA5* insertions in *CEN4* and *CEN5* of ANGE-B-10 feature significantly lower *AthCEN178* edit distances and higher HOR scores in flanking satellites (±2 kb), compared with the elements shared with ANGE-B-2 (Fig. 3h). Alternatively, this may reflect varying spatial integration biases of the centrophilic *ATHILA* families within the *AthCEN178* arrays (Fig. 3b). For Col-0, Ler-0, Cvi-0 and Tanz-1, CENH3 ChIP–seq signal was lower within *ATHILA* elements located in the *AthCEN178* arrays than in the surrounding satellites (Extended Data Fig. 8e), indicating that high levels of *ATHILA* integration may be detrimental to centromere function in *A. thaliana*. Together, these observations support a model in which *AthCEN178* homogenization pathways act to purge *ATHILA* elements within the *A. thaliana* centromere satellite arrays.

## Rapid interspecies centromere evolution

As we observed extreme *A. thaliana* intra-species variation in satellite array size and structure, we sought to compare it to the extent of interspecific centromere divergence. We therefore assembled the genomes of two accessions from the sister species *A. lyrata*. The two accessions, North American MN47 and Siberian NT1, represent sub-specific *A. lyrata* lineages that diverged at least ~90,000 years ago[24–27]. The *A. lyrata* centromere satellite array assemblies were gapless, and sequence comparisons of syntenic centromeres with *A. thaliana* indicated that these regions were significantly different between the species (Fig. 4a and Extended Data Fig. 10a), in line with rapid centromere evolution during or after speciation. The two species share five syntenic centromeres, with three *A. lyrata* centromeres having been lost in *A. thaliana* through chromosome fusions[26,28].

We identified large *A. lyrata* repeat arrays in the expected centromere locations[26,28], which contained two satellite populations: *AlyCEN168*

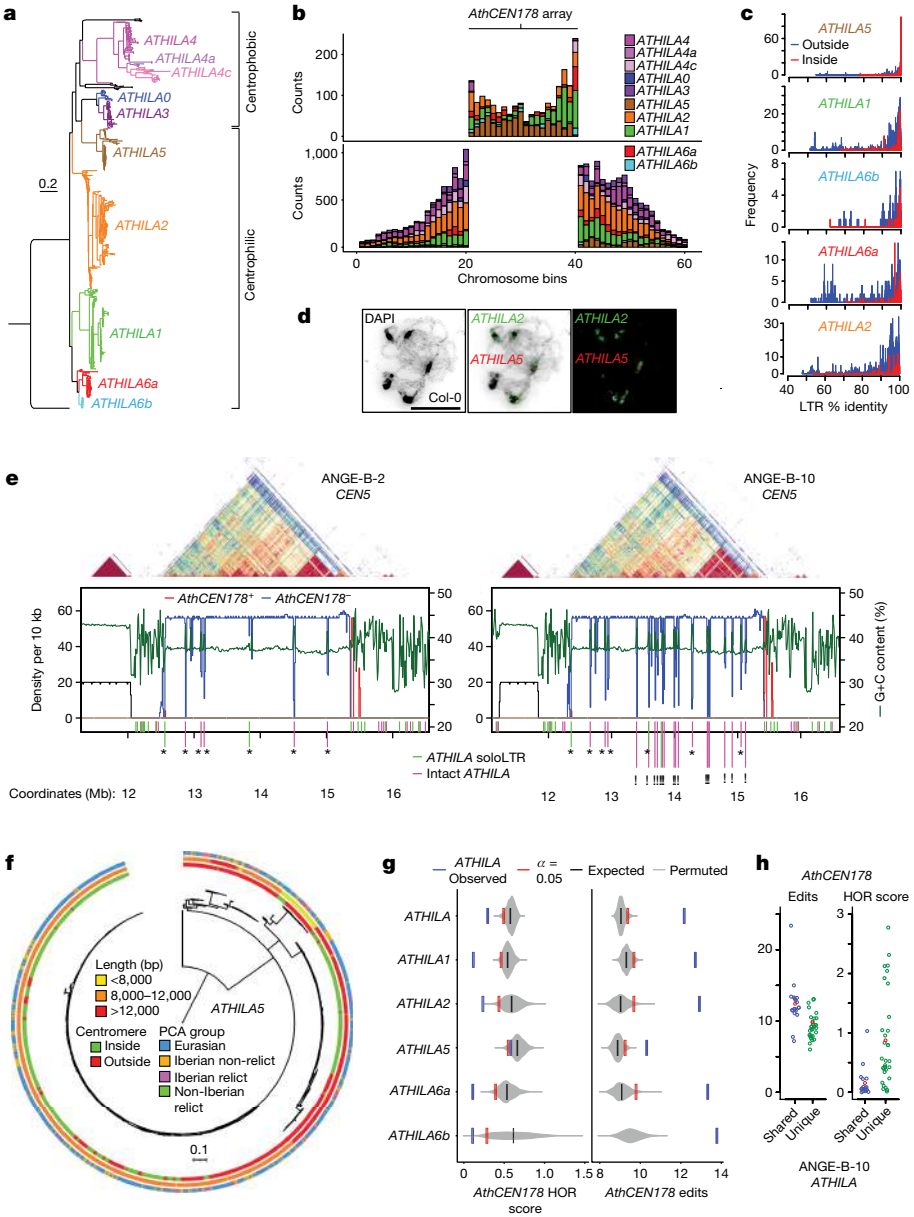



**Fig. 3 | Invasion of the *Arabidopsis* satellite arrays by centrophilic *ATHILA* retrotransposons. a**, Full-length *ATHILA* phylogeny rooted using a maize *Huck-Ty3* element. Centrophobic and centrophilic families are labelled. Scale bar, 0.2 substitutions per site. **b**, Intact elements and soloLTRs for the indicated *ATHILA* families either inside (top) or outside (bottom) the *AthCEN178* arrays. **c**, *ATHILA* LTR percentage sequence identity comparing inside (red) and outside (blue) the *AthCEN178* arrays, according to family. *ATHILA* elements within the arrays were significantly younger than those in the chromosome arms (Wilcoxon test, all $P < 1.57 \times 10^{-8}$). **d**, Representative FISH micrographs with probing for *ATHILA2* (green) and *ATHILA5* (red) on Col-0 pachytene chromosomes (black, DAPI; scale bar, 10 µm). For each sample, 50 independent micrographs were analysed. **e**, ANGE-B-2 and ANGE-B-10 *CEN5* sequence identity heat maps with percentage G+C content (green) and *AthCEN178* density on the forward (red) and reverse (blue) strands plotted below. The *x*-axis ticks indicate intact *ATHILA* (pink) and soloLTR (green) insertions. Asterisks mark *ATHILA* elements shared by ANGE-B-2 and ANGE-B-10; exclamation points mark ANGE-B-10-unique insertions. **f**, *ATHILA5* phylogeny across all 66 accessions. Circumference shading indicates presence inside versus outside the *AthCEN178* arrays, size class and PCA membership. **g**, Analysis of 1-kb regions flanking *ATHILA* elements within the *AthCEN178* arrays across all 66 accessions for HOR score (left) and edit distance (right). Observed mean values (blue) were compared to 1,000 permuted sets of random loci of the same width (grey) within the arrays. Significance thresholds of $\alpha = 0.05$ are indicated (red), as well as means for permuted loci (black). *ATHILA* elements are associated with significantly lower HORs and higher *AthCEN178* divergence ($P < 0.001$), except for *ATHILA5* HORs ($P = 0.128$). Scale bar, 0.1 substitutions per site. **h**, Scatter plots of mean *AthCEN178* edit distance and HOR score in 2-kb windows flanking *ATHILA* elements in ANGE-B-10 *CEN4* and *CEN5*, separated into those shared with ANGE-B-2 (blue) and those unique to ANGE-B-10 (green). Mean values are shown in red.

(~168 bp) and *AlyCEN179* (~179 bp) (Fig. 4a,b). Different centromeres were predominantly composed of a single repeat class, with *AlyCEN168* dominating *CEN2*, *CEN5*, *CEN6* and *CEN8*, whereas *AlyCEN179* dominated *CEN1*, *CEN3*, *CEN4* and *CEN7* (Fig. 4c). These distributions were consistent with FISH analyses, where the pAge and pAa probes distinguished the same chromosome groups[28]. We generated phylogenetic trees and sequence identity heat maps, which showed that *A. thaliana AthCEN178* repeats are more closely related to *AlyCEN168* and *AlyCEN179* repeats, whereas *AthCEN159* repeats have no obvious *A. lyrata* equivalent (Fig. 4d and Extended Data Fig. 10b). Thus, speciation has been accompanied by complete turnover of centromere satellite populations. When analysed phylogenetically, the *A. lyrata* repeats clustered

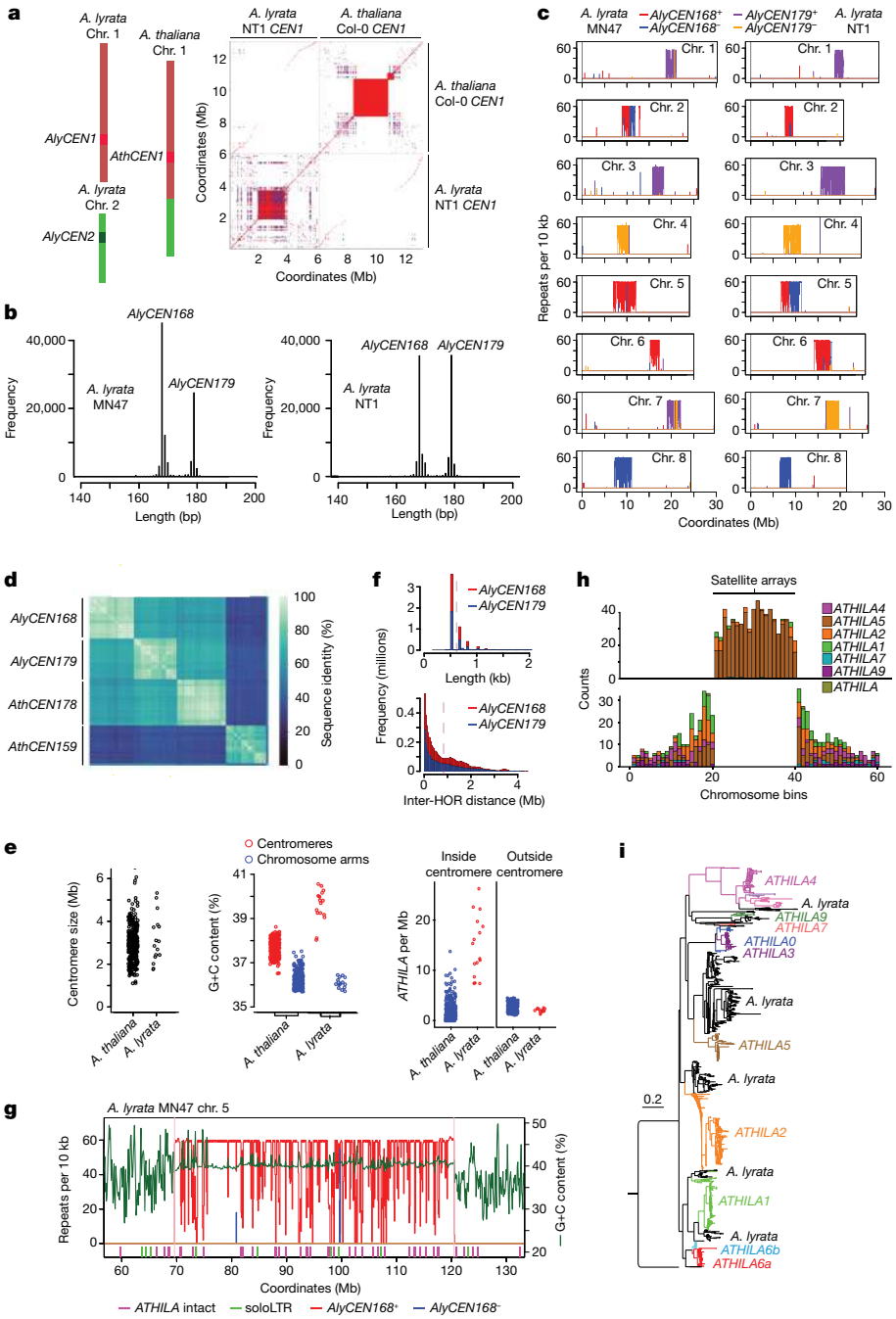



**Fig. 4 | Inter-specific centromere satellite turnover and transposon dynamics between *A. thaliana* and *A. lyrata*.** **a**, Left, diagram of synteny between *A. thaliana* chromosome 1 and *A. lyrata* chromosomes 1 and 2. Right, a sequence identity dot plot comparing *CEN1* in *A. thaliana* Col-0 and *A. lyrata* NT1 using 75-bp windows. Red and blue indicate similarity between the same and opposite strands, respectively. **b**, Tandem repeat lengths in *A. lyrata* MN47 and NT1, with *AlyCEN168* and *AlyCEN179* families labelled. **c**, Satellite repeat density per 10 kb along the *A. lyrata* MN47 and NT1 chromosomes. **d**, Pairwise sequence identity of *AthCEN178* and *AthCEN159* elements from *A. thaliana* Col-0 and *AlyCEN168* and *AlyCEN179* elements from *A. lyrata* NT1. **e**, Scatter plots of satellite array sizes, percentage G+C content of the chromosome arms (blue) and satellite arrays (red), and combined intact and soloLTR *ATHILA* insertions per megabase inside and outside the arrays, from 66 *A. thaliana* and 2 *A. lyrata* accessions. **f**, Satellite HOR lengths (top) and inter-HOR distances (bottom) within the *AlyCEN168* (red)- and *AlyCEN179* (blue)-dominated *A. lyrata* centromeres. **g**, *AlyCEN168* density (red) and percentage G+C content (green) along *CEN5* of *A. lyrata* MN47. Intact *ATHILA* (pink) and soloLTR (green) elements are indicated along the *x* axis. **h**, Combined intact *ATHILA* and soloLTR counts for the indicated families, either inside (top) or outside (bottom) the *A. lyrata* satellite arrays. Centromere arrays were split into 20 bins of varying size, depending on centromere length. **i**, Phylogeny of full-length *ATHILA* elements identified in *A. thaliana* (coloured) and *A. lyrata* (black), rooted using a maize *Huck-Ty3* element. Scale bar, 0.2 substitutions per site.

strongly by chromosome, implying limited inter-chromosome recombination and repeat sharing, as in *A. thaliana* (Extended Data Fig. 10c). The *A. lyrata* and *A. thaliana* arrays were similar in size, and both had higher G+C content than their respective chromosome arms (Fig. 4e). In *A. lyrata*, both *AlyCEN168*- and *AlyCEN179*-dominated centromeres contained HORs of sizes and inter-HOR distances comparable to those in *A. thaliana* (Fig. 4f). Despite MN47 and NT1 chromosomes maintaining populations of distinct satellite classes (*AlyCEN168* versus *AlyCEN179*),

the size and structure of the arrays were polymorphic between the accessions (Fig. 4c and Extended Data Fig. 10a), in line with ongoing intra-species satellite dynamics, as in *A. thaliana*.

We found that centrophilic *ATHILA* behaviour predates the *A. lyrata*–*A. thaliana* split, with *A. lyrata* satellite arrays carrying an even higher insertion load, despite *ATHILA* density not being greater in the chromosome arms (Fig. 4e,g,h and Extended Data Fig. 10d–f). For example, 61% of intact *A. lyrata ATHILA* elements were within the centromere arrays, compared with only 15% in *A. thaliana* (Supplementary Table 5). *ATHILA5* was the most abundant family in the *A. lyrata* centromere satellite arrays (679/994 intact elements), and elements from this family were also more centrophilic (81%) than in *A. thaliana* (56%) and were found to have integrated throughout the length of the arrays (Fig. 4g–i and Supplementary Table 5). Greater retrotransposon content within the *A. lyrata* centromere arrays was not limited to *ATHILA*, as we identified frequent *COPIA* insertions within the satellite arrays (239 *COPIA* elements in 2 *A. lyrata* accessions versus 16 in 66 *A. thaliana* genomes), only six of which were related to the known centrophilic *Tal1* element[29]. This provides evidence for dynamic change in centromeric transposon populations between *A. thaliana* and *A. lyrata*, following speciation.

## Cycles of centromere evolution

Intra-species polymorphism in the *A. thaliana* centromere satellite arrays points to recurrent cycles of expansion and contraction (Fig. 5a). We observed polymorphism in *ATHILA* insertions within the satellite arrays, consistent with ongoing centrophilic retrotransposon invasion. We propose that *ATHILA* integrases have evolved to recognize centromeric DNA sequences or associated chromatin states, which may include CENH3 (ref. 30; Fig. 5a), although we could not detect diagnostic integrase variants that distinguished centrophilic and centrophobic families (Extended Data Fig. 9d). To counter this invasive activity, satellite recombination acts to eliminate *ATHILA* elements (Fig. 5a,b). Evidence for centromeric satellite DNA breakage and recombination exists in several species, including in humans and maize[18,22,31–34]. During meiosis, either a homologue or a sister chromatid could be used for repair, although the latter is most likely to occur during mitosis (Fig. 5b). As haplotype linkage blocks have been maintained across the *AthCEN178* similarity groups, despite internal satellite dynamics, this is consistent with unidirectional allelic or non-allelic gene conversion mediating array evolution (Fig. 5b) or unequal crossover between sister chromatids. As a consequence of homologous recombination, the *Arabidopsis* centromere satellites show signatures of sequence homogenization[17], reminiscent of concerted evolution in diverse other tandem repeat arrays[35–38].

The human kinetochore is thought to recruit recombination factors that promote satellite evolution, termed the kinetochore-associated recombination machine (KARM)[5,39]. We propose that an equivalent activity, 'KARM in *Arabidopsis*' (KARMA), promotes *AthCEN178* recombination, contributing to satellite homogenization, intra-centromere duplications and *ATHILA* purging (Fig. 5a,b). Satellite array size and structure polymorphism between the *A. lyrata* accessions indicates that KARMA is also likely to operate in this species, but may be less efficient at removing transposons, given the higher centromeric load of transposon elements. Variation in these processes may also influence centromeric differentiation between the main genetic lineages of *A. thaliana*, including the larger relict satellite arrays and more frequent *ATHILA* invasion in Eurasian accessions. However, because non-centromeric satellites show concerted evolution[37,38], kinetochore-independent recombination must also exist in tandem repeat arrays located outside the centromere.

Centromere drive during asymmetric female meiosis—as in *A. thaliana*, where only a terminal meiocyte survives[40]—has been proposed to cause rapid evolution of centromere DNA and CENH3 or CENPA protein sequences[41–43]. We propose that recombination, potentially mediated

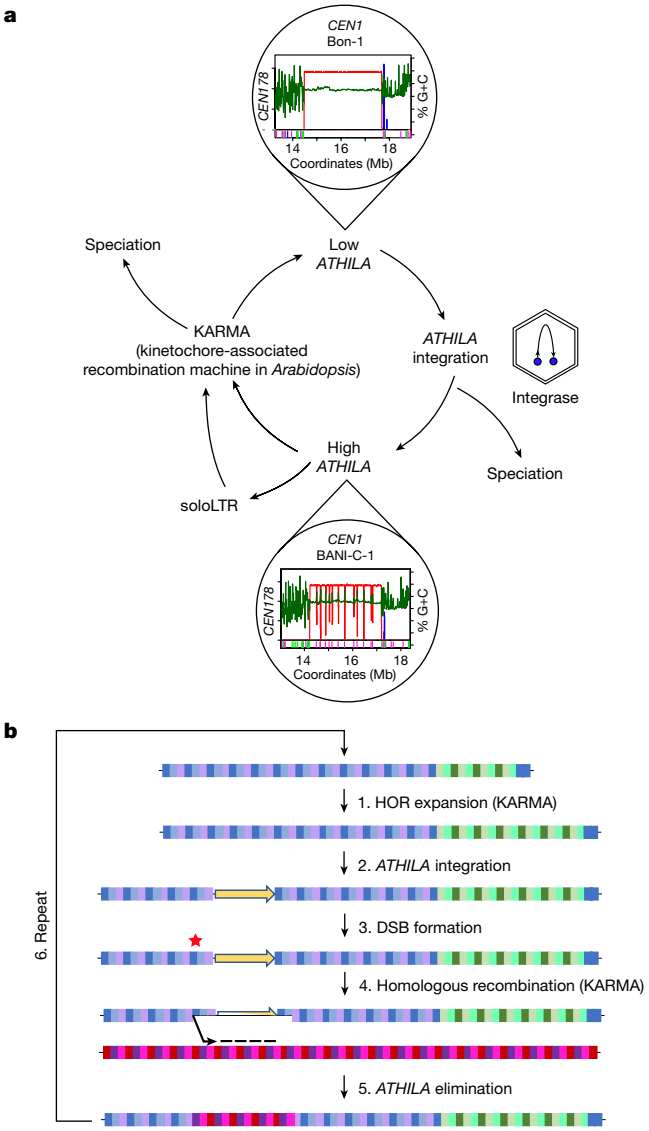



**Fig. 5 | Cycles of satellite homogenization and *ATHILA* retrotransposon invasion drive *Arabidopsis* centromere evolution. a,** *ATHILA* diversity in *A. thaliana* centromere arrays suggests cycles between states with high and low levels of retrotransposons, represented by Bon-1 *CEN1* and BANI-C-1, which belong to *AthCEN178* similarity group 1. *ATHILA* expression leads to virus-like particles where reverse transcription generates new copies with the potential to integrate into the centromeres. We propose that the integrase (blue) of centrophilic *ATHILA* is adapted to recognize centromeric DNA or chromatin states, leading to invasion. *ATHILA* may undergo internal recombination to generate soloLTRs. Simultaneously, satellite arrays undergo homologous recombination, resulting in intra-centromere duplications, satellite homogenization and purging of *ATHILA*. Satellite recombination is proposed to occur through a 'kinetochore-associated recombination machine in *Arabidopsis*' (KARMA) mechanism, by analogy with the human KARM hypothesis[5]. **b**, A representative centromere satellite array is shown, with monomer repeats represented by coloured blocks, where matched colour indicates matched sequence. Through KARMA, satellite HORs can expand (step 1) and arrays can be invaded by *ATHILA* retrotransposons (step 2). The satellite arrays may sustain a DNA double-strand break (DSB, red star) adjacent to an *ATHILA* element during meiosis or mitosis (step 3). The DSB may enter homologous recombination, potentially mediated by KARMA, involving resection and single-stranded DNA invasion of another chromosome, which could be a sister chromosome or a homologue, and template-driven synthesis and repair (step 4). This may result in unidirectional transfer of the satellite sequence (gene conversion) and elimination of the *ATHILA* sequence (step 5). The modified array may then enter further cycles of HOR expansion/contraction and *ATHILA* invasion (step 6).

by KARMA, can promote rapid satellite diversification within *A. thaliana* (Fig. 5). CENH3 location may allow some centromeres to outcompete others during female meiosis and bias segregation into the functional meiocyte, which may also be influenced by *ATHILA* integrations. The high intra-specific diversity of *A. thaliana* satellite arrays implies that specific variants are unable to dominate populations. Instead, conditional, frequency-dependent advantages may generate balanced satellite polymorphism and maintain high intra-species diversity. As centromere drive is widespread in plants and animals[41,42,43], exploring mechanisms of centromere competition within and between *Arabidopsis* species may prove illuminating for eukaryotic genome evolution and speciation.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06062-z.


1. McKinley, K. L. & Cheeseman, I. M. The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16–29 (2016).
2. Talbert, P. B., Masuelli, R., Tyagi, A. P., Comai, L. & Henikoff, S. Centromeric localization and adaptive evolution of an *Arabidopsis* histone H3 variant. *Plant Cell* **14**, 1053–1066 (2002).
3. Melters, D. P. et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).
4. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
5. Miga, K. H. & Alexandrov, I. A. Variation and evolution of human centromeres: a field guide and perspective. *Annu. Rev. Genet.* **55**, 583–602 (2021).
6. Naish, M. et al. The genetic and epigenetic landscape of the centromeres. *Science* **374**, eabi7489 (2021).
7. Rabanal, F. A. et al. Pushing the limits of HiFi assemblies reveals centromere diversity between two *Arabidopsis thaliana* genomes. *Nucleic Acids Res.* **50**, 12309–12327 (2022).
8. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
9. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
10. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
11. Durvasula, A. et al. African genomes illuminate the early history and transition to selfing. *Proc. Natl Acad. Sci. USA* **114**, 5213–5218 (2017).
12. Novikova, P. Y. et al. Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).
13. Schmickl, R., Jørgensen, M. H., Brysting, A. K. & Koch, M. A. The evolutionary history of the *Arabidopsis lyrata* complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evol. Biol.* **10**, 98 (2010).
14. Darwin Tree of Life Project Consortium. Sequence locally, think globally: the Darwin Tree of Life Project. *Proc. Natl Acad. Sci. USA* **119**, e2115642118 (2022).
15. Christenhusz, M. J. M. et al. The genome sequence of thale cress, *Arabidopsis thaliana* (Heynh., 1842). *Wellcome Open Res.* **8**, 40 (2023).
16. Langley, S. A., Miga, K. H., Karpen, G. H. & Langley, C. H. Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. *eLife* **8**, e42989 (2019).
17. Dover, G. Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–117 (1982).
18. Rudd, M. K., Wray, G. A. & Willard, H. F. The evolutionary dynamics of alpha-satellite. *Genome Res.* **16**, 88–96 (2006).
19. Wijnker, E. et al. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *eLife* **2**, e01426 (2013).
20. Smith, G. P. Evolution of repeated DNA sequences by unequal crossover. *Science* **191**, 528–535 (1976).
21. Talbert, P. B. & Henikoff, S. Centromeres convert but don't cross. *PLoS Biol.* **8**, e1000326 (2010).
22. Shi, J. et al. Widespread gene conversion in centromere cores. *PLoS Biol.* **8**, e1000327 (2010).
23. Slotkin, R. K. The epigenetic control of the *Athila* family of retrotransposons in *Arabidopsis*. *Epigenetics* **5**, 483–490 (2010).
24. Mable, B. K., Robertson, A. V., Dart, S., Di Berardo, C. & Witham, L. Breakdown of self-incompatibility in the perennial *Arabidopsis lyrata* (Brassicaceae) and its genetic consequences. *Evolution* **59**, 1437–1448 (2005).
25. Foxe, J. P. et al. Reconstructing origins of loss of self-incompatibility and selfing in North American *Arabidopsis lyrata*: a population genetic context. *Evolution* **64**, 3495–3510 (2010).
26. Hu, T. T. et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
27. Kolesnikova, U. et al. Genome of selfing Siberian *Arabidopsis lyrata* explains establishment of allopolyploid *Arabidopsis kamchatica*. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.24.497443 (2022).
28. Berr, A. et al. Chromosome arrangement and nuclear architecture but not centromeric sequences are conserved between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Plant J.* **48**, 771–783 (2006).
29. Tsukahara, S. et al. Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*. *Genes Dev.* **26**, 705–713 (2012).
30. Malik Harmit, S. & Eickbush, T. H. Modular evolution of the integrase domain in the *Ty3/Gypsy* class of LTR retrotransposons. *J. Virol.* **73**, 5186–5190 (1999).
31. Nijman, I. J. & Lenstra, J. A. Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeats. *J. Mol. Evol.* **52**, 361–371 (2001).
32. Chatterjee, B. & Lo, C. W. Chromosomal recombination and breakage associated with instability in mouse centromeric satellite DNA. *J. Mol. Biol.* **210**, 303–312 (1989).
33. Wolfgruber, T. K. et al. High quality maize centromere 10 sequence reveals evidence of frequent recombination events. *Front. Plant Sci.* **7**, 308 (2016).
34. Mahtani, M. M. & Willard, H. F. Pulsed-field gel analysis of α-satellite DNA at the human X chromosome centromere: high-frequency polymorphisms and array size estimate. *Genomics* **7**, 607–613 (1990).
35. Brown, S. D. & Dover, G. A. Conservation of segmental variants of satellite DNA of *Mus musculus* in a related species: *Mus spretus*. *Nature* **285**, 47–49 (1980).
36. Durfy, S. J. & Willard, H. F. Concerted evolution of primate α satellite DNA. Evidence for an ancestral sequence shared by gorilla and human X chromosome α satellite. *J. Mol. Biol.* **216**, 555–566 (1990).
37. Coen, E., Strachan, T. & Dover, G. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the *melanogaster* species subgroup of *Drosophila*. *J. Mol. Biol.* **158**, 17–35 (1982).
38. Liao, D., Pavelitz, T., Kidd, J. R., Kidd, K. K. & Weiner, A. M. Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the *RNU2* locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. *EMBO J.* **16**, 588–598 (1997).
39. Shepelev, V. A., Alexandrov, A. A., Yurov, Y. B. & Alexandrov, I. A. The evolutionary origin of man can be traced in the layers of defunct ancestral α satellites flanking the active centromeres of human chromosomes. *PLoS Genet.* **5**, e1000641 (2009).
40. Armstrong, S. J. & Jones, G. H. Female meiosis in wild-type *Arabidopsis thaliana* and in two meiotic mutants. *Sex. Plant Reprod.* **13**, 177–183 (2001).
41. Akera, T., Trimm, E. & Lampson, M. A. Molecular strategies of meiotic cheating by selfish centromeres. *Cell* **178**, 1132–1144 (2019).
42. Fishman, L. & Saunders, A. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science* **322**, 1559–1562 (2008).
43. Kursel, L. E. & Malik, H. S. The cellular mechanisms and consequences of centromere drive. *Curr. Opin. Cell Biol.* **52**, 58–65 (2018).
44. Hall, S. E., Luo, S., Hall, A. E. & Preuss, D. Differential rates of local and global homogenization in centromere satellites from *Arabidopsis* relatives. *Genetics* **170**, 1913–1927 (2005).

# Methods

### Genomic DNA extraction and PacBio HiFi sequencing of *A. thaliana*

*A. thaliana* was grown at 23 °C under 16 h of light. High-molecular-weight (HMW) genomic DNA was extracted with a column-based or β-mercaptoethanol (BME) method. For the column-based method, 30-day-old rosettes from multiple individuals were pooled and ground in liquid nitrogen, and ~10 g of powder was resuspended in 500 ml of freshly prepared, ice-cold nuclei isolation buffer (NIB; 10 mM Tris-HCl pH 8.0, 100 mM KCl, 10 mM EDTA, 500 mM sucrose, 4 mM spermidine and 1 mM spermine). The homogenate was filtered through two layers of Miracloth (EMD Millipore, 475855-1R) and distributed into several 50-ml Falcon tubes, to which 1:20 (v/v) NIB containing 20% Triton X-100 was added. Samples were incubated on ice for 15 min and centrifuged at 3,000*g* for 15 min at 4 °C. Pellets were pooled, washed with 35 ml NIB containing 1% Triton X-100 and pelleted once more at 3,000*g* for 15 min at 4 °C. The pellet was gently resuspended in 20 ml of pre-warmed (37 °C) G2 lysis buffer (Qiagen, 1014636) and incubated with 50 μg ml$^{-1}$ RNase A (Qiagen, 19101) at 37 °C for 30 min, followed by treatment with 200 μg ml$^{-1}$ proteinase K (Qiagen, 19133) at 50 °C for 3 h. After centrifugation at 8,000*g* for 15 min at 4 °C, DNA in the supernatant was purified with Genomic-tip 100/G (Qiagen, 10243) using the Genomic DNA Buffer Set (Qiagen, 19060), following the manufacturer's instructions. Then, 0.7 volumes of isopropanol were gently added to the resulting flow-through, and the precipitated DNA was spooled with a glass hook by slow tube rotation and resuspended in EB buffer (Qiagen, 19086) overnight at 4 °C (ref. 7).

For the BME method[45], 300 mg of tissue powder, either from 30-day-old rosettes from multiple plants or from single 26-day-old individual plants, was incubated for 45 min at 55 °C in freshly prepared and pre-heated lysis buffer (1% sodium metabisulfite, 1% PVP40, 0.5 M NaCl, 100 mM Tris-HCl pH 8, 50 mM EDTA pH 8, 1.5% SDS and 2% BME). The next steps were performed at room temperature. First, 60 μl of 20 mg ml$^{-1}$ PureLink RNase A (Thermo Fisher Scientific, 12091021) was added to the lysate and samples were incubated for 10 min. To precipitate proteins, 600 μl of 5 M potassium acetate was added to the samples followed by 2.4 ml of 25:24:1 (v/v/v) phenol:chloroform:isoamyl alcohol (ROTI, A156.1), and samples were incubated for 10 min on a rotator. After centrifuging at 4,400*g* for 10 min, the upper phase was transferred to a new tube and mixed with 24:1 (v/v) chloroform:isoamyl alcohol for 10 min on a rotator. Following a second centrifugation step at 4,400*g* for 10 min, the upper phase was transferred to a new tube, followed by two bead clean-up steps. The first clean-up involved incubation for 30–60 min with 1 volume of a 0.4% solution of SeraMag SpeedBeads Carboxyl Magnetic Beads (GE Healthcare, 65152105050450) on a rotator. After placing the tube on a magnet, the supernatant was discarded and the beads were washed twice with 80% ethanol. Elution was performed with 50 μl EB (Qiagen) after incubation at 37 °C for 15 min. The second clean-up was performed with 0.45 volumes of AMPure PB magnetic beads (PacBio, 100-265-900). After binding for 30 min on a rotator, beads were placed on a magnet and washed twice with 80% ethanol. For elution, 45 μl EB was added and samples were incubated for 10–15 min on a rotator. The HMW DNA extraction method for each accession and whether source tissue was from multiple or single individuals are indicated in Supplementary Table 1.

HiFi library construction was carried out with some variations in the method of shearing HMW DNA and the Binding Kit version for sequencing. For the Megaruptor approach, HMW DNA was sheared in a Megaruptor 2 instrument (Diagenode, B06010002) with the 20-kb or 25-kb setting. Five micrograms of the sheared fraction was used for library construction with the HiFi SMRTbell Express Template Prep Kit 2.0 and size selected with the BluePippin system (SageScience) using the 0.75% DF Marker S1 3–10 kb–Improved Recovery cassette definition (BLF7510, Biozym) under high-pass elution mode and a second elution for 30 min to increase recovery yield. Libraries were sequenced on the Sequel II system (PacBio) using Binding Kit 2.0 (101-842-900). For the gTUBE approach, HMW DNA (120 ng μl$^{-1}$) was sheared twice (back and forth) with a gTUBE (Covaris, 520079) in an Eppendorf Centrifuge 5424 at 4,800 rpm (soft) for 3 × 1 min. Five micrograms of sheared DNA was used for library construction with the HiFi SMRTbell Express Template Prep Kit 2.0 and size selected with the BluePippin system (SageScience) with a 17-kb cut-off in a 0.75% DF Marker S1 High-Pass 6–10 kb v3 gel cassette (BLF7510, Biozym). The resulting libraries were sequenced using Binding Kit 2.2 (101-894-200).

### Genomic DNA extraction and ONT sequencing of *A. thaliana*

For genomic DNA extraction for ONT sequencing, 21-day-old *A. thaliana*, grown on ½ MS medium containing 1% sucrose, were placed in the dark for 48 h before harvesting. Approximately 10 g of flash-frozen seedlings were used per 200 ml of MPD-based extraction buffer pH 6.0 (MEB). Tissue was flash frozen and then ground in liquid nitrogen using a mortar and pestle and resuspended in 200 ml MEB. Ground tissue was thawed in MEB with frequent stirring. The homogenate was forced through four layers of Miracloth and then filtered again through four layers of fresh Miracloth by gravity. Triton X-100 was added to a final concentration of 0.5% on ice, followed by incubation with agitation on ice for 30 min. The suspension was centrifuged at 800*g* for 20 min at 4 °C. The supernatant was removed and the pellet was resuspended using a paintbrush in 10 ml of 2-methyl-2,4-pentanediol buffer pH 7.0 (MPDB). The suspension was centrifuged at 650*g* for 20 min at 4 °C. The supernatant was removed and the pellet was washed with 10 ml MPDB. Washing and centrifugation were repeated until the pellet appeared white, when it was resuspended in a minimal volume of MPDB. From this point onwards, all transfers were performed using wide-bore pipette tips. Five millilitres of CTAB buffer was added to the nuclei pellet and mixed via gentle inversion, followed by incubation at 60 °C until full lysis had occurred, taking between 30 min and 2 h. An equal volume of chloroform was added and samples were incubated on a rocking platform, with a speed of 18 cycles per minute, for 30 min, followed by centrifugation at 3,000*g* for 10 min. An equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added to the lysate, followed by incubation on a rocking platform (18 cycles per minute) for 30 min. The lysate was centrifuged at 3,000*g* for 10 min, and the upper aqueous phase was transferred to a fresh tube. Extraction was repeated using phenol:chloroform:isoamyl alcohol and then only chloroform. One-tenth volume of 3 M 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 −20 °C for 48 h. The precipitated DNA was removed using a glass hook and washed three times in 70% ethanol. The DNA was dissolved in 120 μl of 10 mM Tris-HCl pH 8.5.

Approximately 5 μg of DNA were size selected (>30 kb) with the BluePippin System (Sage Science) and a 0.75% agarose gel cassette (BLF7510, Biozym), using Range mode and BP start set at 30 kb. Library preparation followed that for the ONT SQK-LSK109 protocol kit, using 1.2–1.5 μg of size-selected DNA in a volume of 48 μl. DNA was nick 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 of NEBNext Ultra II End Prep Enzyme Mix (New England Biolabs, E7180S), with incubation for 30 min at 20 °C, followed by 30 min at 65 °C. The sample was cleaned using 1 volume of AMPure XP beads and eluted in 61 μl of nuclease-free water. Adaptors were ligated at room temperature using 25 μl ligation buffer, 10 μl NEBNext T4 DNA ligase and 5 μl adaptor mix for 2 h. The library was cleaned with 0.4 volumes of AMPure XP beads, washed using ONT Long Fragment buffer and eluted in 15 μl elution buffer.

### Genomic DNA extraction and PacBio HiFi sequencing from *A. lyrata*

*A. lyrata* was grown in the greenhouse at 21 °C under 16 h of light, as described[27]. HMW DNA was isolated from 1.5 g of plant material using a NucleoBond HMW DNA kit, as described[27]. Quality was assessed using

a FEMTOpulse device (Agilent), and quantity was measured using a Quantus fluorometer (Promega). HiFi libraries were prepared according to the manual 'Procedure & Checklist—Preparing SMRTbell Libraries using the SMRTbell Express Template Prep Kit 2.0' with an initial DNA fragmentation using gTUBEs (Covaris), and final library size selection was performed using a BluePippin (Sage Science) instrument. Size-selected libraries were sequenced on a Sequel II instrument with Binding Kit 2.0 and Sequel II Sequencing Kit 2.0 for 30 h (PacBio).

### CENH3 ChIP–seq

Approximately 12 g of 2-week-old 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 and 1× protease inhibitor cocktail) and cross-linked in 1% formaldehyde at room temperature for 25 min. The cross-linking reaction was quenched with 125 mM glycine, and samples were incubated at room temperature for a further 25 min. The nuclei were purified from cellular debris by two rounds of filtration through one layer of Miracloth and centrifuged at 2,500*g* for 25 min 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 and 1× protease inhibitor cocktail) and centrifuged at 14,000*g* for 10 min 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 and 1 mM pepstatin A), and chromatin was sonicated using a Covaris E220 device with the following settings: power, 150 V; bursts per cycle, 200; duty factor, 20%; time, 90 s. Sonicated chromatin was centrifuged at 14,000*g*, 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, 1 mM pepstatin A and 1× protease inhibitor cocktail). The chromatin was incubated overnight at 4 °C with 50 µl Protein A magnetic beads (Dynabeads, Thermo Fisher) pre-bound with 2.5 µl of antibody to CENH3 (ref. 2, 1:400 dilution; gift of S. Henikoff, Fred Hutchinson Cancer Research Center, Seattle, WA, USA). 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 and 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 and 1× protease inhibitor cocktail). Immunoprecipitated DNA–protein complexes were eluted from the beads (1% SDS and 0.1 M $NaHCO_3$) at 65 °C for 15 min. Samples were reverse cross-linked 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 by phenol:chloroform:isoamyl alcohol (25:24:1) extraction and ethanol precipitation. Library preparation followed the Tecan Ovation Ultralow System v2 library protocol. ChIP samples were PCR amplified for 12 cycles and sequenced with 150-bp paired-end reads on an Illumina instrument by Novogene.

### FISH

For FISH, mitotic and meiotic (pachytene and diakinesis) chromosome spreads were prepared from young anthers, as described[6]. To identify chromosomes 1, 2 and 4, 5,040-kb (clones F6F9–F23M19), 4,138-kb (clones T9J22–F13P17) and 2,079-kb (clones T6K21–T12H17) BAC contigs were used, respectively. BAC clone T1J24 was used for chromosomal localization of *ATHILA2*. Primers for PCR amplification of the *ATHILA5* probe (ATH5-F3, 5′-GCACAAGGGATGGACAGACT-3′; ATH5-R3, 5′-TGCATCTTATGTCATTGCAGCA-3′) were designed against the consensus sequence of *ATHILA5* LTRs, using Primer3 (v.2.3.7) implemented in Geneious Prime 2022.2. PCR amplification was performed as follows: initial denaturation at 95 °C for 5 min; 35 cycles of denaturation at 95 °C for 20 s, annealing at 58 °C for 20 s and extension at 72 °C for 1 min; and a final extension at 72 °C for 5 min. PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel, 740609.50) and cloned into the pGEM-T Easy Vector System (Promega, A1360), using TOP10 chemically competent cells. Fourteen positive colonies were screened using universal SP6 and T7 primers, and six clones with the expected length of amplicon were Sanger sequenced (Macrogen). Sequences of clones were aligned to the consensus sequence for *ATHILA5* LTRs using MAFFT (v.7.490)[46], and two clones with high similarity to the consensus sequence were selected for cytogenetic experiments. The *AthCEN178* pAL FISH probe, which labels all centromeres, was amplified using primers ATH_cen180F and ATH_cen180R, as reported[6]. PCR amplification was performed as follows: initial denaturation at 95 °C for 5 min; 35 cycles of denaturation at 95 °C for 20 s, annealing at 46 °C for 20 s and extension at 72 °C for 20 s; and a final extension at 72 °C for 5 min. PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). All DNA probes were labelled with biotin–dUTP, digoxigenin–dUTP or Cy3–dUTP using nick translation and were then pooled, ethanol precipitated and pipetted onto pepsin-treated and ethanol-dehydrated slides containing suitable chromosome spreads. The slides were heated to 80 °C for 2 min and incubated at 37 °C for 12 h. The hapten-labelled probes were immunodetected as described[6]. Fluorescence signals were analysed and imaged using a Zeiss AxioImager epifluorescence microscope (Carl Zeiss) with a CoolCube camera (MetaSystems). Images were acquired separately using Isis software (MetaSystems) for all four fluorochromes with appropriate excitation and emission filters (AHF Analysentechnik). The four monochromatic images were pseudo-coloured, merged and cropped using Photoshop CS (Adobe Systems). Fluorescence of *AthCEN178* was measured using ImageJ (National Institutes of Health) and quantified as corrected total fluorescence (CTF = integrated density − (selected area × mean fluorescence of background)).

### Genome assembly from *A. thaliana* HiFi reads

Of the 66 assemblies analysed, 63 were generated for this study and the other 3 have been published[7,14,15]. Sample size was not predetermined. q20 HiFi reads were generated with the PacBio Circular Consensus Sequencing tool ccs (v.6.0.0) (--all), followed by extracthifi (v.1.0) (PacBio tools, https://github.com/PacificBiosciences/pbbioconda). When two or more libraries were sequenced on the same SMRTcell, demultiplexing was carried out with lima (v.2.0.0) (--same --ccs --min-score 70 --min-scoring-regions 2 --min-ref-span 0.8 --peek-guess --split-bam-named). For 49 accessions, de novo assembly was performed with hifiasm (v.0.16.1-r375) (-l0 -f0)[47] and contigs larger than 100 kb were scaffolded into chromosomes with RagTag (v.2.0.1) (scaffold -q 60 -f 30000 -i 0.5 --remove-small)[48], using as the reference genome a version of the TAIR10 assembly in which centromeres, telomeres, organellar nuclear insertions, and 5S and 45S rDNA arrays had been masked[7]. The remaining assembly gaps in the chromosome scaffolds of accessions Rabacal-1 and Tanz-1 were 'patched' with their corresponding continuous long read (CLR)-based assemblies using RagTag (v.2.0.1) (patch --join-only -f10000 --remove-small)[48]. For 14 accessions, de novo assembly was performed with hifiasm v.0.15.4-r343 (-l0 -f0)[47] and contigs larger than 100 kb were scaffolded into chromosomes with RagTag v.2.0.1 (scaffold -q 60 -f 30000 -i 0.5 --remove-small)[48], using as the reference genome a Bionano optical map of IP-Ini-0. The assemblies of accessions Col-0 (ref. 7), Ey15-0 (ref. 7) and Kew-1 (refs. 14,15) were downloaded from the European Nucleotide Archive (ENA) database, under assembly IDs GCA_946499705.1, GCA_946499665.1 and GCA_933208065.1, respectively. All scaffolded assemblies were subjected to manual curation, including inversion, swapping or elimination of some of their underlying contigs. These edits were implemented in the agp files, which were subsequently converted to fasta format with the RagTag agp2fa function[48].

### Quality assessment of *A. thaliana* HiFi assemblies

HiFi reads were aligned to their corresponding chromosomes with pbmm2 (v.1.9.0) (align --sort --log-level DEBUG --preset SUBREAD --min-length 5000), including the mitochondrial and chloroplast

genomes from TAIR10, to reduce the number of organellar reads aligning to the nuclear genome. Unmapped reads and reads aligning to the organelle chromosomes, as well as secondary and supplementary alignments, were filtered out with samtools (v.1.9) (view -b -F 2308 <input.bam> Chr1 Chr2 Chr3 Chr4 Chr5)[49]. The resulting bam file was used to identify collapsed regions within the centromere satellite array coordinates (defined in Supplementary Table 3), using the tool Segmental Duplication Assembler (v.0.1.0)[50]. Similarly, the same bam file was used to determine the coverage of the first and second most common bases in the aligned PacBio HiFi reads across the genome with NucFreq (v.0.1) (NucPlot.py --minobed 2)[50], while a modified version of the script 'hetDetection.R' (ref. 51) was used to identify and quantify heterozygosity stretches within the satellite arrays, on the basis of the bed file generated by NucFreq. In brief, a heterozygosity stretch was defined as a region where the second most common base was present in at least 10% of reads, in at least 5 positions within a 500-bp region (https://github.com/mrvollger/NucFreq). Representative plots of primary and secondary allele coverage are displayed in Extended Data Figs. 1 and 8d.

We identified six accessions (11C1, AUZE-A-5, Ey15-2, GAIL-B-11, IP-Fel-2 and IP-Ini-0) with more than 1.5 Mb of collapsed centromeric regions. Collapsed regions were defined as those with coverage that was 3 s.d. above the genome-wide mean[50]. For the most part, this increase in coverage was uniform over all centromere regions and, except for localized regions (within *CEN3* in 11C1, *CEN1* and *CEN3* in AUZE-A-5, and *CEN2* and *CEN5* in IP-Ini-0), it was not associated with the presence of heterozygosity stretches. Therefore, these patterns might be due to biases introduced during sample preparation or sequencing, as previously suggested[7,8]. Two other assemblies, Lor-16 and Met-16, which were both sequenced using DNA extracted from single individuals, contained several megabase-long regions of heterozygosity within their genomes, including in *CEN4* in Lor-16 and *CEN5* in Met-6. A detailed summary of collapsed, expanded and heterozygous regions within the centromeric satellite arrays for each accession is included in Supplementary Table 1.

### Genome assembly from *A. lyrata* HiFi reads

PacBio reads were assembled using hifiasm in default mode, as described[27,47], choosing the primary contig graph as the assembly. Assembly completeness was assessed using BUSCO and the 'Brassicales_odb10' set[27].

### SNP analysis and PCA

Genome assemblies of each *A. thaliana* accession were aligned to the HiFi assembly of Col-0 using minimap2 (v.2.24)[52], with the parameters 'minimap2 -a -x asm5 --cs -r2k -t 16'. SNPs and short indels were called from the alignment using bcftools (v.1.15.1)[53]. Repetitive regions of the Col-0 HiFi assembly were masked using RepeatMasker for *AthCEN178*, *AthCEN159*, transposable elements, organelle sequences and rDNA. Transposable element sequences were downloaded from the TAIR10 website. SNPs were further filtered to exclude repeat-masked regions of the Col-0 HiFi genome, requiring no missing sites and that SNPs were strictly biallelic, which resulted in 2,011,142 SNPs. PCA was performed using the R package ggbiplot.

### Identification of cenhaps

HiFi reads from each accession were aligned to the HiFi-based assembly of Col-0, including the mitochondrial and chloroplast genomes from TAIR10, with pbmm2 (v.1.9.0) (align --sort --preset CCS) (https://github.com/PacificBiosciences/pbmm2), which is a wrapper of minimap2 (ref. 52). DeepVariant (v.1.3.0) (--model_type=PACBIO)[54] was used to create individual genome call sets, and GLnexus (v.1.4.1)[55] was used to merge calls into a single cohort. We filtered out SNP and indel variants that overlapped centromeres, telomeres, organellar nuclear insertions, and *5S* and *45S* rDNA arrays, as previously defined[7], with GATK (v.4.1.3.0) (VariantFiltration --exclude-intervals <repeats.bed>)[56]. In addition, we performed variant hard-filtering at the population level with bcftools (v.1.15.1) (view -e 'NS < 66 || F_MISSING > 0 || AC = 1 || COV < 2604 || COV > 6685 || AC_Het ≥ 2')[53] and variant hard-filtering at the population level for biallelic variants with no missing data, and we removed sites that were either 0.5× higher or lower than the median coverage, using bcftools (v.1.15.1). Heterozygous sites were permitted, and the minimum minor allele count was set to three individuals.

We examined SNP and indel variants from the HiFi reads to analyse haplotype structures in the ~800-kb flanking regions surrounding the *AthCEN178* centromere arrays. Polymorphisms were filtered for no missing data and to have a minor allele count of at least three. For each chromosome, this amounted to 7,278, 15,016, 4,890, 15,348 and 12,228 SNPs, together with 402, 810, 233, 697 and 593 indels, for chromosomes 1–5, respectively. We phased the polymorphisms using the Beagle phasing algorithm (v.4.0)[57], and we assigned the minor allele as 1 and the major allele as 0. To examine the haplotype structure, we clustered accessions on the basis of the Hamming distance of the polymorphisms flanking the centromeres in Col-0, using the R package cultevo. We generated heat maps to represent the haplotype structure using the R package pheatmap. Samples were coloured on the basis of their *AthCEN178* similarity group (Supplementary Table 2).

### Tandem repeat annotation and analysis

The Tandem Repeat Annotation and Structural Hierarchy (TRASH) pipeline was used to identify *A. thaliana AthCEN178* and *AthCEN159* and *A. lyrata AlyCEN169* and *AlyCEN179* tandem repeats (https://github.com/vlothec/TRASH). Genomic sequences were divided into 1-kb windows that were analysed for their repetitive content using *k*-mer counts (proportion of non-unique *k*-mers). Local concatenation of the windows with high *k*-mer scores produced regions with repetitive sequence content. The periodicity of tandem repeats within these regions was calculated on the basis of distances between identical *k*-mers. This was in turn used for iterative sampling of the region, where each sampled sub-sequence was locally mapped, using matchPattern from the R package Biostrings, and scored for the fraction of sequence covered by the mappings. Mapped regions of the best sub-sequence were used to create a primary consensus with MAFFT (--retree 2) alignment[46]. When a lower-than-expected fraction of the region was covered by the mapped sub-sequences, the remaining sequence was used in a second round of identification, for example, if a region contained continuous runs of repeats of more than one family. Because the primary consensus sequence is a result of a random sampling, repeats of the same family may start at a different relative position and be in a different phase shift. To remedy this, an algorithm using a DNA hash table readjusted the shift by calculating the lowest possible hash score based on each shift iteration, such that closely related sequences had the same relative start position in the annotation. The shifted primary consensus was then used for a second round of mapping, which included a refinement step where short gaps were filled and overlaps were split or removed, and a final annotated set of repeats was reported.

An *AthCEN178* consensus sequence for each accession was calculated on the basis of MAFFT alignment of all repeats (--retree 2), and the most common nucleotides at each position, with a frequency of at least 50%, were extracted. To generate a global consensus sequence, frequency tables from each accession were aligned and averaged and a consensus sequence was extracted, as before. Levenshtein edit distance between each repeat and the accession-specific consensus sequence was calculated using the stringdist function from the stringdist R package[58].

To calculate similarity between the *AthCEN178* sequences of a pair of chromosomes, the number of identical *AthCEN178* instances found within the other chromosome, including duplicated repeats, was calculated reciprocally. These values were normalized by the number of *AthCEN178* repeats in both sets, such that values were in the range of 0–100%, where 0% means no shared repeats and 100% means the chromosomes have the same repeat library. Heat maps were constructed by ordering the chromosome pairs with hierarchical cluster analysis,

using the R function hcluster from the package amap. Centromere similarity groups were visually inferred from the heat maps.

Satellite HORs, understood as repeat block duplications, were identified using the TRASH HOR module. *AthCEN178* repeats from a single chromosome were globally aligned using MAFFT (--kimura 1 --retree 1), and each pair of repeats was scored for the number of disagreements. Each uninterrupted series of repeat pairs with a number of disagreements below the threshold ($n = 5$) was reported as a HOR instance. HORs consisting of fewer than three monomers per block were discarded. Following HOR counting, the count value for each repeat was divided by the number of *AthCEN178* copies on that chromosome and expressed as a percentage, which we term the HOR score. To identify large (kilobase to megabase) intra-centromere duplications within the *A. thaliana AthCEN178* satellite arrays, we ran TRASH HOR identification using stringent settings to identify HORs with an edit distance of 0 between *AthCEN178* instances and involving at least five *AthCEN178* repeats. After filtering for HORs with identical blocks and at least five monomers in each, a high frequency of similar inter-block distances was taken to indicate an intra-centromere duplication. HORs resulting from such duplications were grouped if their length summed to at least 50 kb. Duplication size was estimated on the basis of the first and last HORs in the involved regions.

Maximum-likelihood satellite phylogenies were reconstructed using IQ-TREE with 1,000 bootstraps[59]. IQ-TREE calculates a bootstrap approximation (UFBoot) for each node. The satellites selected to construct each tree were chosen by stratified sampling: satellites were sorted into groups by chromosome and then randomly chosen from these groups. The number of satellites chosen from each group was proportional to the group's size, relative to the total number of satellites in the centromere array. For example, if chromosome 1 contained more satellites than the other chromosomes, proportionally more sequences would be sampled from chromosome 1. In Fig. 1h, 250 *AthCEN178* sequences were sampled from Eurasian (T850 and Bon-1) and non-Iberian relict (Tanz-1 and Rab-1) accessions, respectively. In Extended Data Fig. 10b, 60 *AlyCEN179* and *AlyCEN168* sequences from two *A. lyrata* accessions (NT1 and MN47) and 60 *AthCEN178* and *AthCEN159* sequences from six *A. thaliana* accessions (Bon-1, IP-Bus-0, Rab-1, Tanz-1, IP-Alo-19 and IP-Cas-6) were sampled. In Extended Data Fig. 10c, 450 *AlyCEN168* and 450 *AlyCEN179* satellites from *A. lyrata* MN47 were sampled. Before generating trees, non-identical sequences were removed and the resulting sequences were aligned using MAFFT (v.7.490), with default settings. All trees were rooted with consensus *C. rubella* 160-bp and 175-bp centromeric repeats[44], apart from the trees in Extended Data Fig. 10b,c, which were rooted with 30 *AthCEN178* sequences from the Col-0 *A. thaliana* accession. For Fig. 4d, satellites from accessions were aligned and their percentage identity was scored with Clustal Omega and then visualized using a heat map. One hundred *AthCEN178* and 100 *AthCEN159* satellites from *A. thaliana* Col-0 were sampled, in addition to 100 *AlyCEN179* and 100 *AlyCEN168* satellites from *A. lyrata* NT1.

To assess the presence or absence of *AthCEN178* sequences in the 66 *A. thaliana* accessions, a library of unique *AthCEN178* sequences per chromosome was generated, amounting to 272,174, 225,600, 260,924, 243,509 and 218,246 unique *AthCEN178* sequences from chromosomes 1–5, respectively. The presence or absence of these sequences was codified as two alleles and queried against the *AthCEN178* libraries for each accession generated by TRASH. A binary matrix of *AthCEN178* presence/absence was then generated for each chromosome. Saturation analysis of the unique *AthCEN178* sequences was calculated with the R package vegan, using the function specaccum, with 1,000 permutations of the binary matrix of *AthCEN178* presence/absence. To generate sequence identity heat maps, the StainedGlass package was used[60]. Genome regions of interest were analysed using a 5- or 10-kb window and the mm_f option set as 10 kb.

### *ATHILA* and transposon annotation and analysis

To achieve large-scale yet sensitive and highly accurate discovery and analysis of *ATHILA* LTR retrotransposons, we designed a new tool, ATHILAfinder (https://github.com/eliasprim/ATHILAfinder). The ATHILAfinder pipeline uses sequence motifs that are specific to the junctions between the internal domain and the LTRs of *ATHILA* elements as primary seeds. These motifs were identified using the manually curated and deeply annotated *ATHILA* dataset from the Col-CEN genome assembly[6] and comparison with *ATHILA* and non-*ATHILA* transposable elements in the TAIR10 TE database[61] and the *A. thaliana* TE database developed and maintained by H. Quesneville (https://urgi.versailles.inra.fr/Data/Transposable-elements/Arabidopsis). Using this approach, we designed three seeds located in the 5′ LTR–primer-binding site (PBS) junction that cover all *ATHILA* families and lack homology with other transposon sequences: TATCAATTTGGCGCCGTTGCC covers all *ATHILA* families except the *ATHILA4*/*ATHILA4A*/*ATHILA4C* clade that instead uses TATCAAATTTGGCGCCGTTGCC and *ATHILA8*/*ATHILA8A*/*ATHILA8B* that uses CATCAAGCTTTTTGGCGCCGT. The main difference between the three seeds is the length and composition of the intervening sequence between the terminal pentamer of the 5′ LTR (underlined left) and the PBS (underlined right). We designed an *ATHILA* universal seed in the more conserved polypurine tract (PPT)–3′ LTR junction: CTAAGTTTGGGGGAGTTGATA, with the PPT underlined on the left and the first pentamer of the LTR underlined on the right. We used Vmatch (http://www.vmatch.de/) to identify the motifs with -h 3 or 4, to allow up to three and four mismatches in the 5′ LTR–PBS and PPT–3′ LTR motifs, respectively. We then only kept motifs that were within 2–11 kb and in the same direction and rejected those that contained additional motifs internally or upstream/downstream within a 2-kb window. These new 2- to 11-kb fragments represent the putative internal domain of intact *ATHILA* elements. To examine whether these internal domains are flanked by LTRs, and given that the LTRs of all *ATHILA* families are between 1–2 kb in length, we searched the windows that were found −2 to −1 kb upstream and +1 to +2 kb downstream for two types of 20-mers using Vmatch (-h 4): the 20-mer immediately downstream of the PPT–3′ LTR junction (including the conserved TGATA) was searched for in the upstream window of the 5′ LTR–PBS junction to locate the beginning of the 5′ LTR, and the 20-mer upstream of the 5′ LTR–PBS junction (including the conserved TATCA/CATCA) was searched for in the downstream window of the PPT–3′ LTR junction to locate the end of the 3′ LTR. Intact elements were defined when both 20-mers were found. This approach identified 8,709 *ATHILA* elements in the 66 *A. thaliana* accessions and 885 in the 2 *A. lyrata* accessions; however, additional high-quality intact *ATHILA* elements will be missed by ATHILAfinder when small deletions occur that overlap the seeds or in the case of short indels within the seeds. To retrieve these elements, we included a homology BLASTn-based recovery step that uses as query the intact *ATHILA* dataset and a minimum coverage threshold of 90%. The candidate loci then underwent a similar process as above based on the 20-mers to refine and identify with high precision the external boundaries and the junctions of the internal domain with the LTRs. This approach further identified 541 intact *ATHILA* elements in *A. thaliana* and 109 in *A. lyrata*. We also identified soloLTRs in the *A. thaliana* and *A. lyrata* genomes. We first ran BLASTn using as a query the 5′ LTRs of the intact *ATHILA* elements and applied a minimum coverage threshold of 98%. We then extracted the first and last 50 bp of the target loci and also 50 bp of the upstream and downstream flanking regions to form two 100-bp fragments. These fragments were scanned for the *ATHILA* seeds using Vmatch (h -5) and were classified as soloLTRs if no seeds were identified.

We aligned the flanking pentamers of every intact *ATHILA* and soloLTR using MAFFT and reported TSDs by allowing one mismatch[46]. In total, 6,546 intact *ATHILA* and 8,776 soloLTRs contained a TSD in *A. thaliana* and 806 and 333 contained a TSD in *A. lyrata*. We identified ORFs (minimum of 300 bp) in the internal domain of the intact elements using getorf in EMBOSS[62] and the core domains of GAG and POL ORFs by running HMMER (v.3.3.2) (-E 0.001 --domE 0.001) (http://hmmer.org/), using a collection of hidden Markov models (HMMs) downloaded from Pfam (http://pfam.xfam.org/) that describe the genes of *Ty3*/*Gypsy* LTR retrotransposons: PF03732 for GAG; PF13650, PF08284, PF13975, PF00077 and PF09668 for protease; PF00078 for reverse transcriptase; PF17917, PF17919 and

PF13456 for RNase H; PF00665, PF13683, PF17921, PF02022, PF09337 and PF00552 for integrase; and PF03078 for the *ATHILA*-specific domain. Given that all *ATHILA* families contain non-autonomous sub-lineages that lack the core domains for the reverse transcriptase, RNase H and integrase (Extended Data Fig. 9 and Supplementary Table 5), we used the full-length sequence of the intact elements to examine their phylogenetic relationships. When required, the number of input sequences for multiple-sequence alignments was reduced using CD-HIT[63]. We then ran MAFFT with the G-INS-i or FFT-NS-i parameters depending on the number of input sequences and FastTree (-nt) to generate the maximum-likelihood tree[64]. Phylogenetic trees were visualized and annotated using FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and iTOL[65].

To map the insertion sites of *ATHILA* elements, 20-bp regions flanking each element, including TSDs, were extracted and locally aligned to the *AthCEN178* consensus using the pairwiseAlignment function from the R package Biostrings. Centromeric *ATHILA* elements were filtered on the basis of this score, and only those for which both flanking sequences had a positive alignment score were kept (1,367 of 1,913). Overall, 88.4% of the *ATHILA* elements had their insertion site overlapping by 5 bp, which is expected because of the size of TSDs. The insertion positions were determined as the middle position between the two mapped flanking sequences, and their distribution along the *AthCEN178* consensus sequence was plotted.

We matched and analysed syntenic *ATHILA* insertions, both intact elements and soloLTRs, in closely related pairs of accessions: BARC-A12 and BARC-A-17, BELC-C-10 and BELC-C-12, CAMA-C-2 and CAMA-C-9, SALE-A-10 and SALE-A-17, ANGE-B-2 and ANGE-B-10, the duplicated regions of *CEN5* in FERR-A-8 and the single shared element in *CEN1* of BARC-A-17 and IP-Ini-0. To unambiguously identify, or 'match', syntenic *ATHILA*, we used multiple criteria including phylogeny, orientation, total length, LTR length, flanking sequence (either as TSD or flanking pentamers in the absence of a TSD) and the order of appearance of *ATHILA* elements. The results are shown in Supplementary Table 6, where *ATHILA* elements are classified as 'matched', 'unmatched' (when a centromere or a stretch within a centromere contained different sets of *ATHILA* between the pair; for example, see *CEN2* of the ANGE pair in Supplementary Table 6) or 'new' (when a new *ATHILA* copy was located within a highly syntenic centromere; for example, see *CEN4* and *CEN5* of the ANGE pair in Supplementary Table 6). Matched *ATHILA* were then aligned with MAFFT (--globalpair --maxiterate 1000) to calculate substitutions, indels and alignment length.

We followed two different approaches for retrieval of the complete sequence of the integrase gene, for every family (Extended Data Fig. 9d), which was based on the number of mutations in the coding domain of *ATHILA* elements. When most or all elements of a family contain a large number of mutations, identification of the correct sequence is very challenging, because the *gag-pol* ORF is broken into numerous smaller fragments in different frames. Four families, however, *ATHILA5*, *ATHILA1*, *ATHILA2* and *ATHILA6b*, contained several elements with a complete *gag-pol* ORF (Extended Data Fig. 9a–c). Using this subset of elements, we were able to build the consensus sequence of integrase for these families. We then constructed an HMM profile on the basis of these four consensus sequences and used FraHMMER (https://github.com/TravisWheelerLab/FraHMMER) to identify homologies in the other families, even in the presence of frameshift mutations. We detected complete integrase hits for *ATHILA7*, *ATHILA9*, *ATHILA4c* and *ATHILA0*, but not for the remaining six families (*ATHILA6a*, *ATHILA3*, *ATHILA4*, *ATHILA4a*, *ATHILA8a* and *ATHILA7a*). This was expected on the basis of our previous scans of full-length elements using a collection of HMMs for transposon genes, which failed to retrieve the integrase core domains for these families. Finally, for identification of intact elements of non-*ATHILA* transposons, we ran EDTA with default parameters[66].

### Phylogenetic analysis of CENH3

Protein-coding genes in all assemblies were annotated with Liftoff (v.1.6.2) (-copies -sc 0.98 -polish -cds)[67], using as input annotation file (-g) 'Araport11_GFF3_genes_transposons.201606.gff' downloaded from https://www.arabidopsis.org. The coding DNA sequence (CDS) of locus AT1G01370 (*CENH3*) across all 66 assemblies was extracted with gffread (v.0.12.7)[68], and multiple-sequence alignment was performed with MAFFT (v.7.407) (--reorder --maxiterate 1000 --retree 1)[46]. Subsequently, the fasta alignment was transformed to Nexus format with PGDspider (v.2.1.1.5)[69], and a maximum-likelihood phylogenetic tree was inferred with RAXML-NG (v.0.9.0) (raxml-ng --all --model GTR+G --seed 9 --bs-trees 1000 --bs-metric fbp,tbe)[70]. Tree visualization was performed using the R package ggtree (v.1.16.6)[71], while the R package treeio (v.1.8.2) was used to add the multiple-sequence alignment as a tree object into R[72].

### Mapping DNA methylation using ONT reads

We quantified DNA methylation in the CG, CHG and CHH contexts with DeepSignal-plant (v.0.1.4)[73], which uses a deep-learning method based on a bidirectional recurrent neural network with long short-term memory units to detect 5mC methylation[73]. R9 reads were filtered for length and accuracy using Filtlong (v.0.2.0) (--min_mean_q 90, --min_length 20000). Base-called 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 was called using DeepSignal-plant with the following model:

model.dp2.CNN.arabnrice2-1_120m_R9.4plus_tem.bn13_sn16.both_bilstm.epoch6.ckpt

The scripts 'call_modification_frequency.py' and 'split_freq_file_by_5mC_motif.py' provided in the DeepSignal-plant package[73] were used to generate the methylation frequency at each CG, CHG and CHH site.

### ChIP–seq data alignment and processing

Deduplicated paired-end CENH3 ChIP–seq Illumina reads (2 × 150 bp) from Col-0, Cvi-0, Ler-0 and Tanz-1 were processed with Cutadapt (v.1.18) to remove adaptor sequences and low-quality bases (Phred+33-scaled quality <20)[74]. For each accession, trimmed reads were aligned to the respective genome assembly using Bowtie2 (v.2.3.4.3), with the following settings: --very-sensitive --no-mixed --no-discordant -k 10 --maxins 800 (ref. 75). Up to ten valid alignments were reported for each read pair. Read pairs with Bowtie2-assigned MAPQ of <10 were discarded using samtools (v.1.10)[49]. For retained read pairs that aligned to multiple locations, with varying alignment scores, the best alignment was selected. Alignments with more than two mismatches or consisting of only one read in a pair were discarded. For each dataset, bins per million mapped reads (BPM; equivalent to transcripts per million for RNA-seq data) coverage values were generated in bigWig and bedGraph formats with the bamCoverage tool from deepTools (v.3.5.0)[76]. Reads that aligned to the chloroplast or mitochondrial genome were excluded from coverage normalization. For profiling of CENH3 occupancy within *AthCEN178* repeats, trimmed reads were aligned to their respective genome assembly with Bowtie2 (v.2.3.4.3), using the following settings: --very-sensitive --no-mixed --no-discordant -k 200 --maxins 800. Read pairs with Bowtie2-assigned MAPQ of <2 were discarded using samtools (v.1.10)[49]. For retained read pairs that aligned to multiple locations, with varying alignment scores, the best alignment was selected. Alignments with more than two mismatches or consisting of only one read in a pair were discarded. Alignment bedGraphs were converted to per-base 1-based coverage files and used to calculate $\log_2$(ChIP/input), and custom R scripts were used to plot the average profiles across *AthCEN178* and *ATHILA* elements.

### Evaluation of *AthCEN178* metrics in regions flanking centromeric *ATHILA*

We applied permutation tests to evaluate the relationships between centromeric *ATHILA* and *AthCEN178* metrics (HOR score and edit distance from the *AthCEN178* consensus sequence) across all 66 accessions, using

R (v.4.1.2). For this analysis, mean *AthCEN178* metrics in 1-kb regions flanking centromeric *ATHILA* start and end coordinates were first calculated. The mean of these values was then calculated across all 66 accessions. This was repeated for 1,000 sets of randomly positioned centromeric loci of the same number and width distribution for each accession as for the centromeric *ATHILA*. For each *AthCEN178* metric, values calculated for each set of random loci were compared with the value observed for centromeric *ATHILA* to derive an empiric *P* value. *AthCEN178* metric values for random loci were also used to calculate the significance threshold ($\alpha = 0.05$) and an expected value (the mean value from 1,000 sets).

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

*A. thaliana* and *A. lyrata* accessions used for sequencing are held in the authors' laboratories and seeds are freely available on request. The genome assemblies analysed in this study are available under the following accession numbers: (1) 48 *A. thaliana* HiFi assemblies have been submitted to the ENA under project number PRJEB55353 (ERP140242); (2) 15 *A. thaliana* HiFi assemblies have been submitted to the ENA under project number PRJEB55632 (ERA17524869); (3) 2 *A. thaliana* HiFi assemblies (Col-0 and Ey15-2) are available at the ENA under project number PRJEB50694 (ERP135313)[7]; (4) 1 *A. thaliana* HiFi assembly (Kew-1) from the Darwin Tree of Life is available under project accession PRJEB51511 (refs. 14,15) and can also be accessed at https://portal.darwintreeoflife.org/data/root/details/Arabidopsis%20thaliana; (5) ONT reads from the Ler-0, Cvi-0 and Tanz-0 accessions have been submitted as ArrayExpress accession E-MTAB-12009, while those for the accession Col-0 were previously available as ArrayExpress accession E-MTAB-10272 (ref. 6); (6) CENH3 Illumina ChIP–seq reads from Col-0, Ler-0, Cvi-0 and Tanz-0 have been submitted as ArrayExpress accession E-MTAB-11974; and (7) 2 *A. lyrata* HiFi assemblies are available at the ENA under project number PRJEB50329 (ERP134897)[27].

## Code availability

The TRASH algorithm is available at https://github.com/vlothec/TRASH, the ATHILAfinder algorithm is available at https://github.com/eliasprim/ATHILAfinder and additional custom code associated with the manuscript is available at https://github.com/vlothec/pancentromere.


45. Russo, A. et al. Low-input high-molecular-weight DNA extraction for long-read sequencing from plants of diverse families. *Front. Plant Sci.* **13**, 883897 (2022).
46. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
47. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
48. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
49. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
50. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
51. Mc Cartney, A. M. et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).
52. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
53. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
54. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
55. Yun, T. et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1081 (2021).
56. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
57. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
58. M. P. J.van der Loo The stringdist package for approximate string matching. *R J.* **6**, 111 (2014).
59. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
60. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac018 (2022).
61. Buisine, N., Quesneville, H. & Colot, V. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics* **91**, 467–475 (2008).
62. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
63. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
64. Liu, K., Linder, C. R. & Warnow, T. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS ONE* **6**, e27731 (2011).
65. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
66. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
67. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1016 (2020).
68. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**, 304 (2020).
69. Lischer, H. E. L. & Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 (2012).
70. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
71. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree : an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
72. Wang, L.-G. et al. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37**, 599–603 (2020).
73. Ni, P. et al. Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. *Nat. Commun.* **12**, 5976 (2021).
74. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
75. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
76. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).



**Acknowledgements** We thank S. Henikoff and P. Talbert (Fred Hutchinson Cancer Research Center, USA) for kindly providing anti-CENH3 antibodies. We thank R. Durbin, C. Zhou (University of Cambridge, UK) and the Darwin Tree of Life Project for the *A. thaliana* ddAraThal4 Kew-1 assembly. 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, Marie Curie International Training Network 'MEICOM' and Human Frontier Science Program award RGP0025/2021 to I.R.H.; EMBO long-term postdoctoral fellowship ALTF224-2022 to R.B.; a Human Frontiers Science Program (HFSP) Long-Term Fellowship (LT000819/2018-L) to F.A.R.; the Max Planck Society to D.W.; European Research Council (ERC) Synergy Grant PATHOCOM (951444) from the European Union's Horizon 2020 program to F.R. and D.W.; an ERA-CAPS 1001G+ grant to M. Nordborg and D.W.; Royal Society awards UF160222, URF\R\221024, RGF/R1/180006 and RGF/EA/201030 to A.B.; European Research Council award ERC HOW2DOBLE 101041354 to P.Y.N.; Czech Science Foundation grant no. 21-03909S to M.A.L.; a BBSRC DTP Studentship to N.G.; a Broodbank Fellowship to M. Naish; and grant PID2022-136893NB-I00 from the Ministerio de Ciencia e Innovación of Spain/Agencia Estatal de Investigación/10.13039/50110001103/FEDER, EU, to C.A.-B.


**Author contributions** P.W., F.A.R., M. Naish, G.S., F.R., C.A.-B., M.A.L., P.Y.N., A.B., D.W. and I.R.H. designed the study. Genomic DNA extractions and sequencing were performed by F.A.R., M. Naish, A.S., N.G., K.F., A.H., C.L., T.S., M.C., M.M. and G.S. FISH experiments were performed by T.M. CENH3 ChIP–seq and DNA methylation profiling were performed by M. Naish. P.W., F.A.R., R.B., M. Naish, E.P., A.S., T.M., N.G., A.J.T., C.P., G.S., M.C., M. Nordborg, M.A.L., D.H., P.Y.N., A.B., D.W. and I.R.H. analysed the results. P.W., F.A.R., R.B., P.Y.N., A.B., D.W. and I.R.H. wrote the paper, with input from all authors.

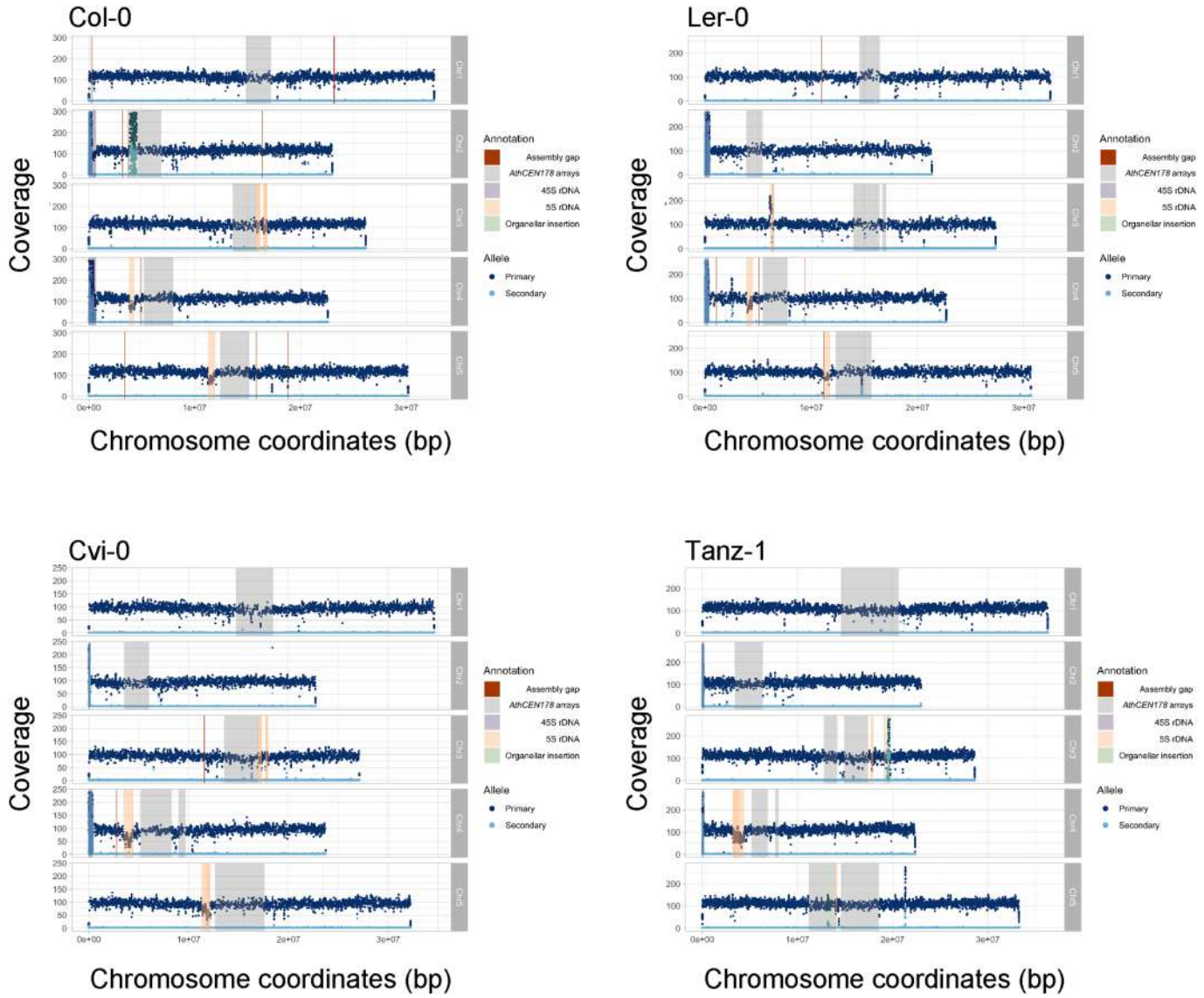
**Competing interests** D.W. holds equity in Computomics, which advises plant breeders. D.W. consults for KWS SE, a plant breeder and seed producer. All other authors declare no competing interests.

**Additional information**

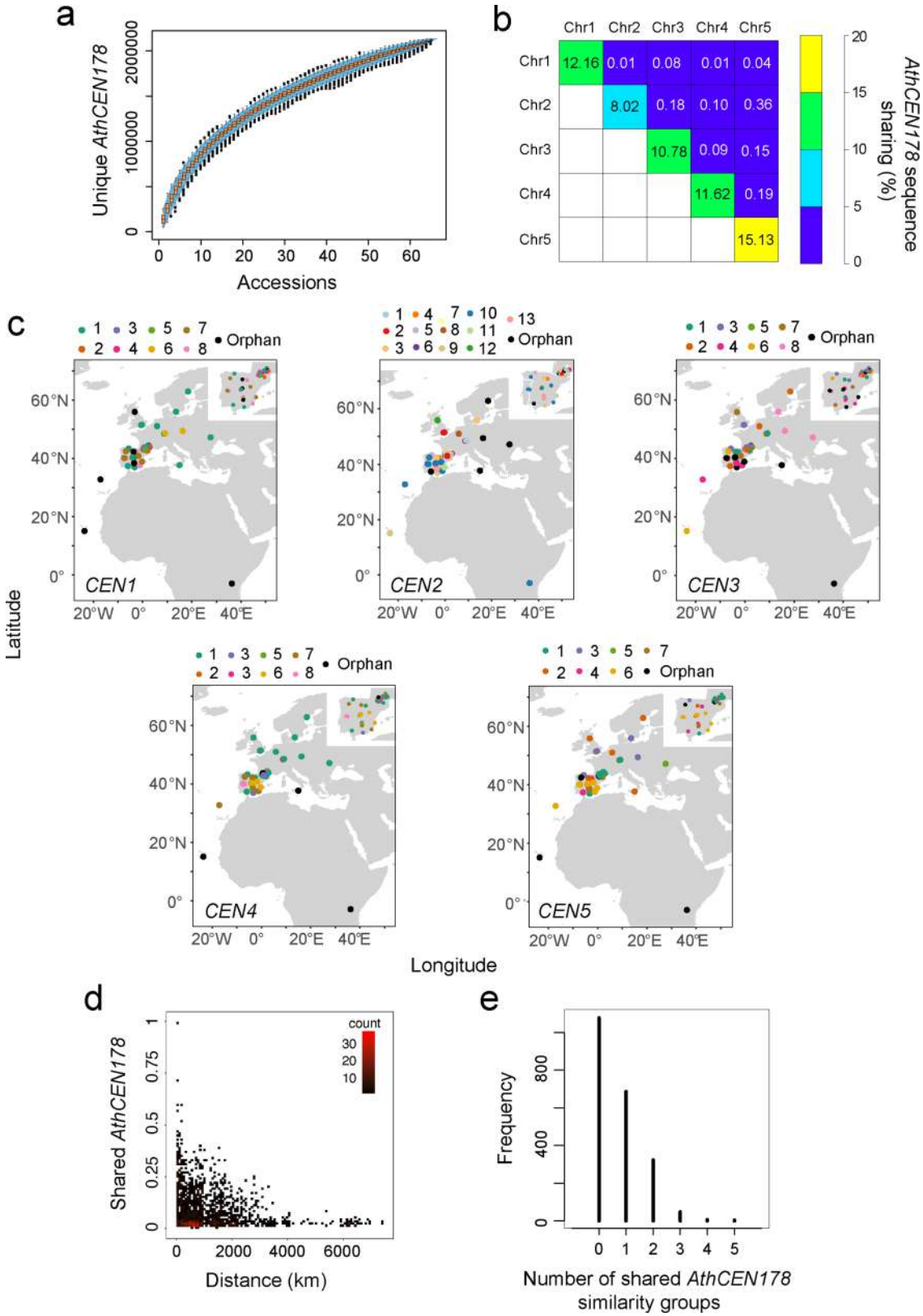
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06062-z.


**Correspondence and requests for materials** should be addressed to Alexandros Bousios, Detlef Weigel or Ian R. Henderson.

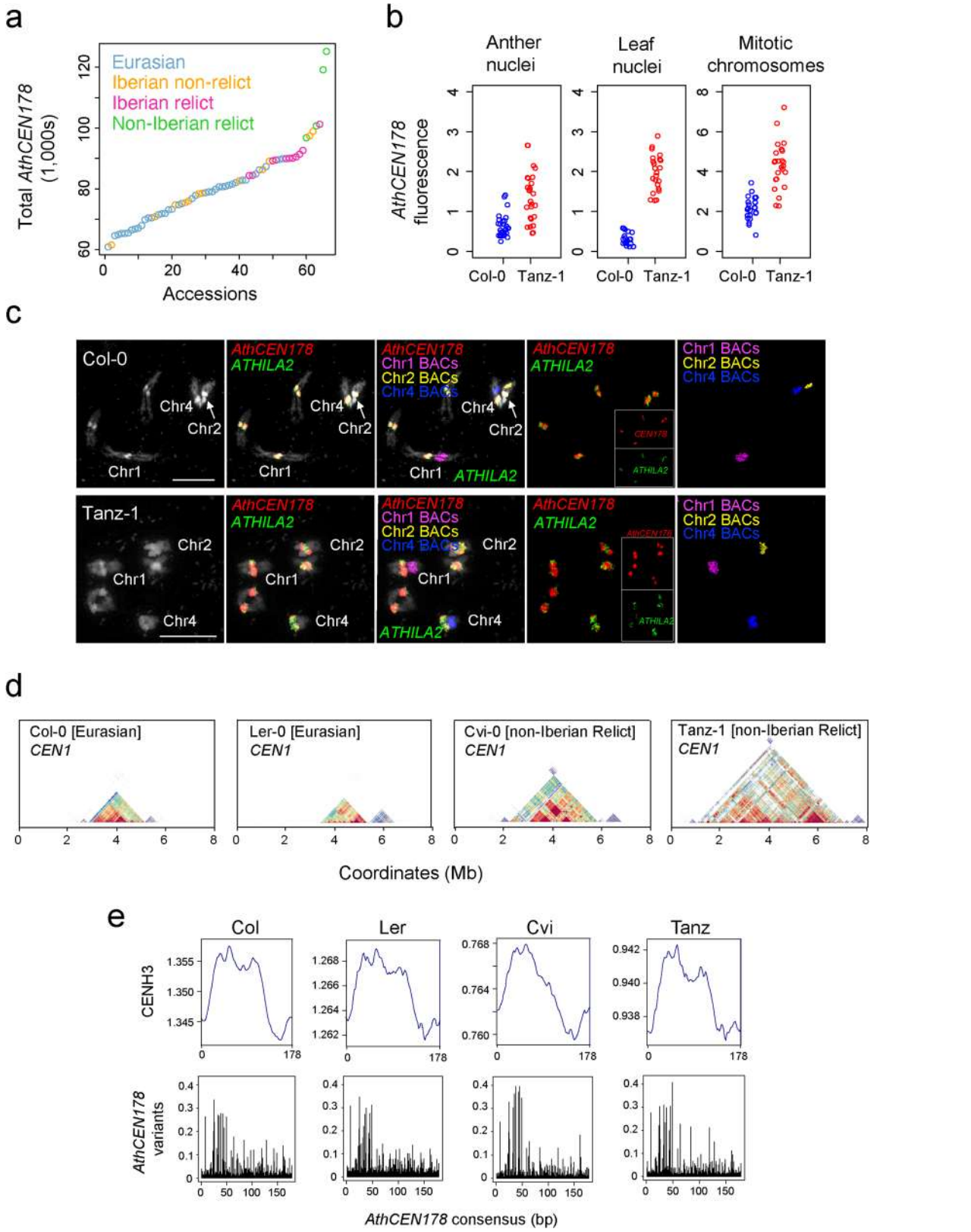

**Peer review information** *Nature* thanks Vincent Colot, Pierre Baduel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



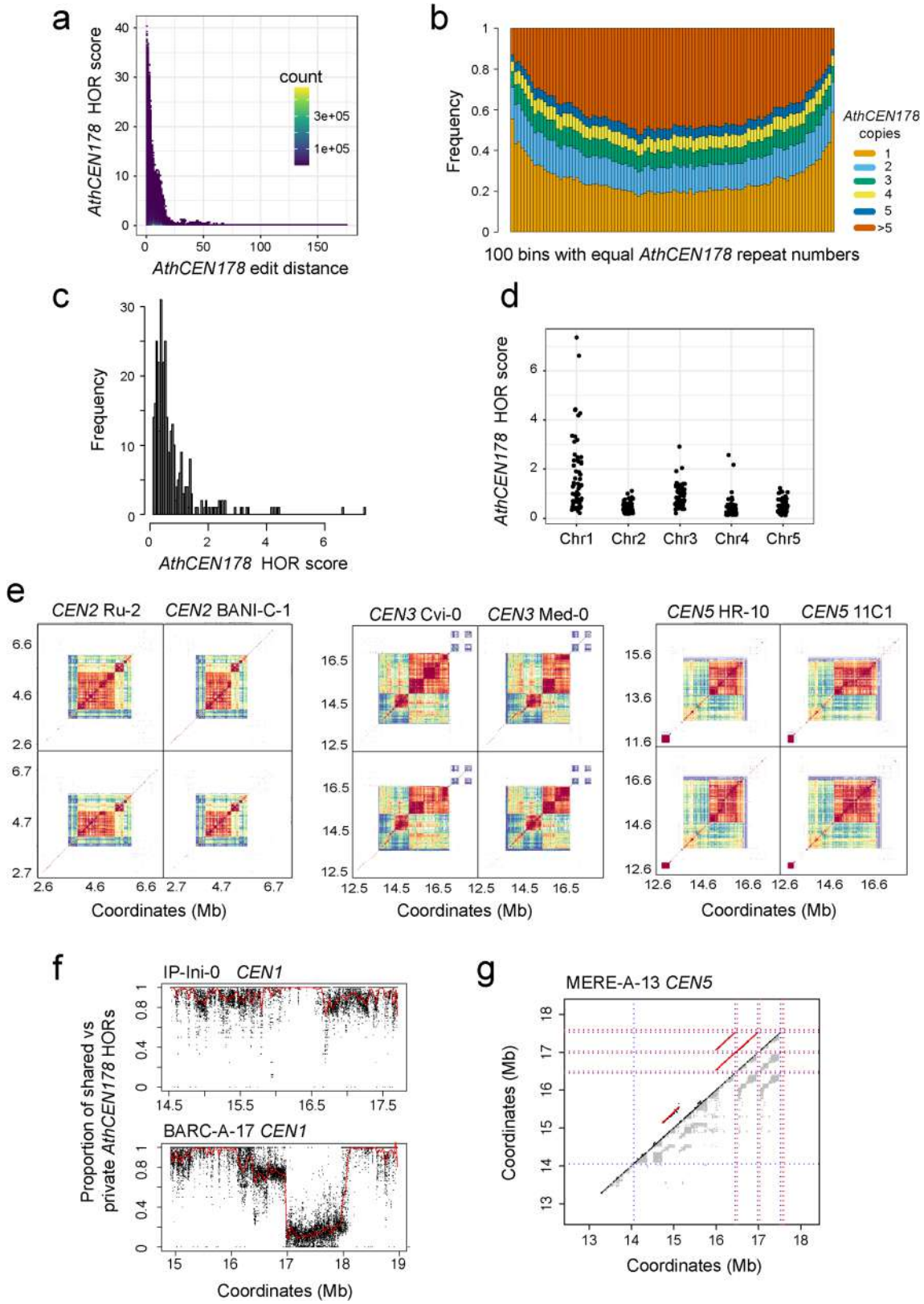

**Extended Data Fig. 1 | Validation of *A. thaliana* centromere assemblies.** The coverage of primary (dark blue) and secondary (light blue) alleles for PacBio HiFi read sets (Col-0, Ler-0, Cvi-0 and Tanz-1) aligned to their corresponding genome assembly. *AthCEN178* arrays coordinates, from Supplementary Table 3, are indicated by grey shading. Assembly gaps are shown by red shading, *45S* rDNA by purple shading, *5S* rDNA by orange shading and organelle insertions by green shading. Equivalent plots for the remaining genome assemblies can be found at the following website: https://github.com/vlothec/pancentromere.




**Extended Data Fig. 2 | Sampling of the *AthCEN178* satellome and geographic distributions of centromere similarity groups. a**, Discovery of non-redundant unique *AthCEN178* variants as a function of sampled accessions, determined with 1,000 permutations. The centre of each boxplot is the median number of non-redundant unique *AthCEN178* variants in the 1,000 permutations. Blue shading = 95% confidence interval. **b**, Heat map showing the average value of exact *AthCEN178* sequence sharing between all pairs of the indicated chromosomes. **c**, Geographic maps are shown with accession origin coloured according to *AthCEN178* similarity group, shown separately for each of the five chromosomes. **d**, Pairwise geographical distance (km) vs. the proportion of shared *AthCEN178* sequences, for all 2,145 accession pairs. **e**, Histogram showing the number of *AthCEN178* similarity groups that are shared when all accession pairs were compared.

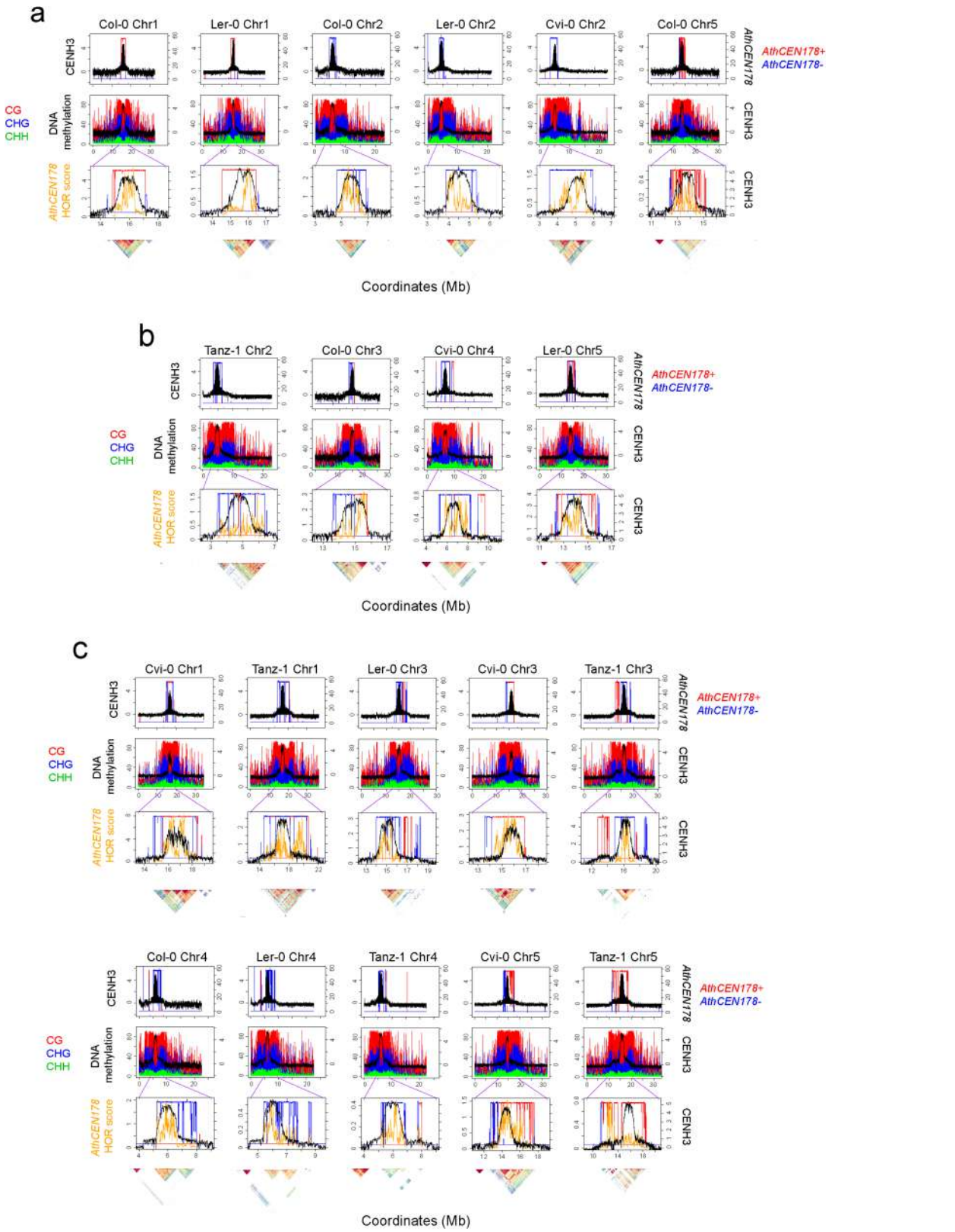



**Extended Data Fig. 3 | Variation in *AthCEN178* copy number and CENH3 occupancy between *A. thaliana* genetic lineages and accessions. a**, Total *AthCEN178* copies per accession, coloured according to Eurasian (blue), Iberian non-relict (orange), non-Iberian relict (green) and Iberian relict (pink) chromosome arm SNP-PCA groups. **b**, Corrected total *AthCEN178* FISH fluorescence intensity from anther nuclei, leaf nuclei, or mitotic chromosomes, in Col-0 (Eurasian, blue) and Tanz-1 (relict, red). All Tanz-1 samples showed significantly greater fluorescence intensity compared to Col-0 (Wilcoxon tests all $P = <1.04\times10^{-6}$). **c**, Representative FISH micrographs for *AthCEN178* (red) and *ATHILA2* (green) on pachytene chromosomes of Col-0 and Tanz-1. Insets on the left are DAPI-stained images of the same cells. Scale bars = 10 µm. **d**, StainedGlass sequence identity heat maps for *CEN1* of Eurasian (Col-0 and Ler-0) and non-Iberian relict (Cvi-0 and Tanz-1) accessions. **e**, CENH3 $\log_2$(ChIP/input) values (upper row) were plotted along all *AthCEN178* repeats in the Col-0, Ler-0, Cvi-0 and Tanz-1 accessions. Beneath (lower row) are plots of *AthCEN178* sequence variants against the consensus repeat for Col-0, Ler-0, Cvi-0 and Tanz-1.

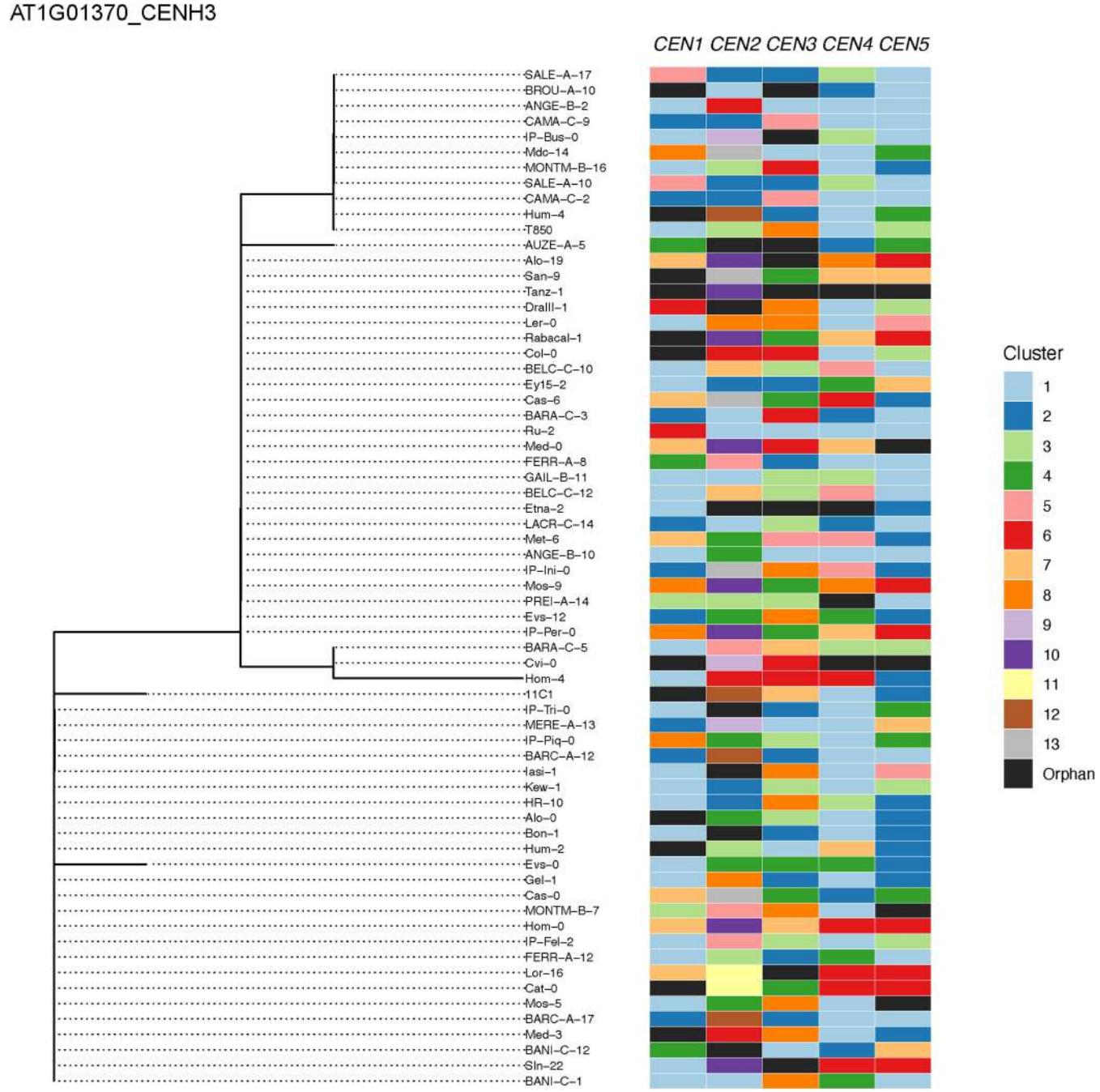



**Extended Data Fig. 4** | See next page for caption.

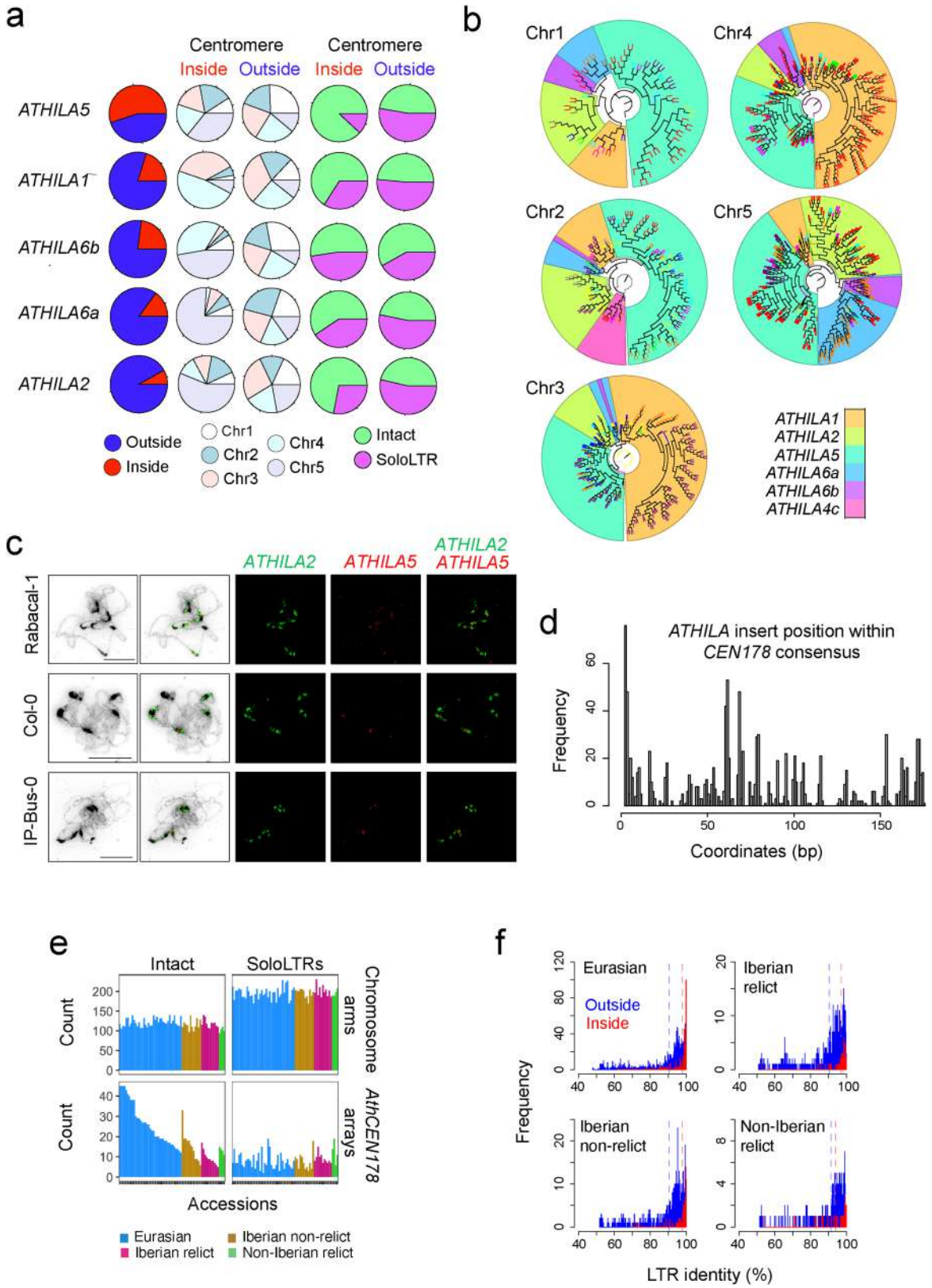
**Extended Data Fig. 4 | *AthCEN178* HORs and dynamic centromere evolution in *A. thaliana*.** **a**, Density plot of *AthCEN178* HOR scores versus edit distances from the chromosome consensus, across all accessions. **b**, The copy number of each *AthCEN178* repeat was calculated within each chromosome individually. For each chromosome, all *AthCEN178* repeats were divided into 100 bins with an equal number of repeats in each bin. The counts of *AthCEN178* with copy numbers of 1, 2, 3, 4, 5 and >5 were divided by the number of repeats per bin, and by the total number of chromosomes. These values were summed for each chromosome to give a total value of 1 per bin. **c**, Histogram of *AthCEN178* HOR scores per centromere. **d**, Scatterplots of *AthCEN178* HOR scores for each of the five chromosomes. **e**, StainedGlass sequence identity heat maps comparing within- and between-accession sequence identity for Ru-2 and BANI-C-1 *CEN2*, Cvi-0 and Med-0 *CEN3* and HR-10 and 11C1 *CEN5*. **f**, Pairwise comparison of the proportion of shared versus private *AthCEN178* HORs between IP-Ini-0 and BARC-A-17 *CEN1*, along the length of each centromere. Red lines represent a smoothing spline. **g**, Dot plot showing intra-centromere duplications (diagonal red lines) detected within MERE-A-13 *CEN5*. Horizontal and vertical dotted lines indicate intact (red) and soloLTR (blue) *ATHILA*.




**Extended Data Fig. 5 | DNA methylation, CENH3 ChIP-seq enrichment, and *AthCEN178* higher-order repeat (HOR) structure within the centromere regions of Col-0, Ler-0, Cvi-0 and Tanz-1. a**, CENH3 ChIP-seq enrichment ($log_2$[ChIP/input], black) compared with *AthCEN178* density in 10 kb windows on forward (red) or reverse (blue) strands along the indicated chromosome and accession. Beneath, CENH3 ChIP-seq enrichment (black) is plotted against DNA methylation (%) in CG (red), CHG (blue) and CHH (green) sequence contexts, along the entire chromosome. Beneath are close-ups of the centromere regions with *AthCEN178* density (red, blue), CENH3 ChIP-seq enrichment (black) and *AthCEN178* HOR score (orange) plotted. A StainedGlass sequence identity heat map is shown at the bottom[60]. The centromeres in (a) are grouped on the basis of having a single *AthCEN178* array that is occupied by CENH3. **b**, As for (a), but showing centromeres that are grouped on the basis of having distinct *AthCEN178* arrays and CENH3 occupying more than one array. **c**, As for (a), but showing centromeres with multiple *AthCEN178* arrays, only one of which is occupied by CENH3.

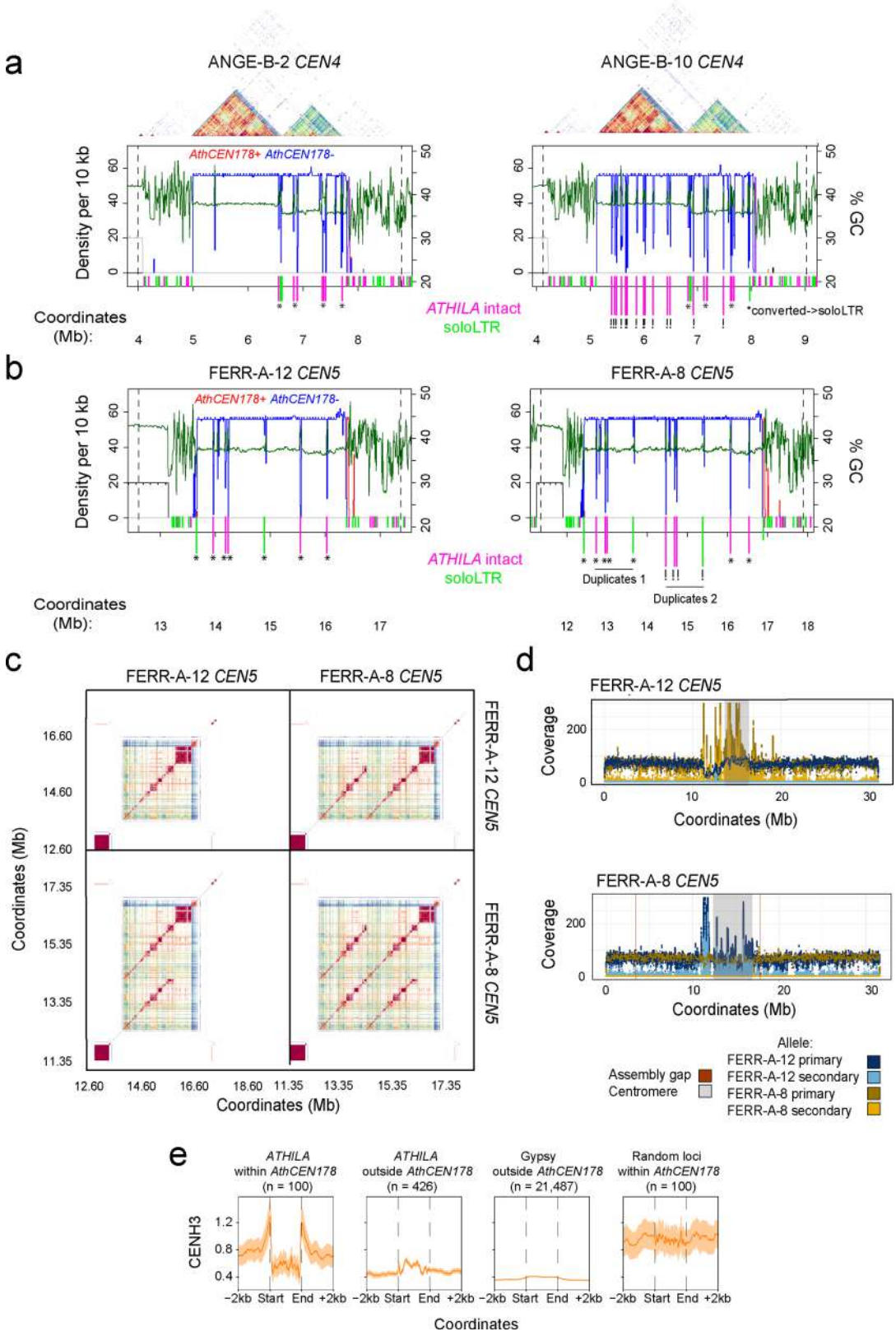



**Extended Data Fig. 6 | Variation in *CENH3* coding sequence in relation to centromere *AthCEN178* similarity groups.** A phylogenetic tree based on *CENH3* nucleotide sequences is shown for the 66 *A. thaliana* accessions (left). To the right of the tree is a coloured key indicating the *AthCEN178* similarity group membership for each of the five chromosomes (*CEN1-CEN5*) for each accession.



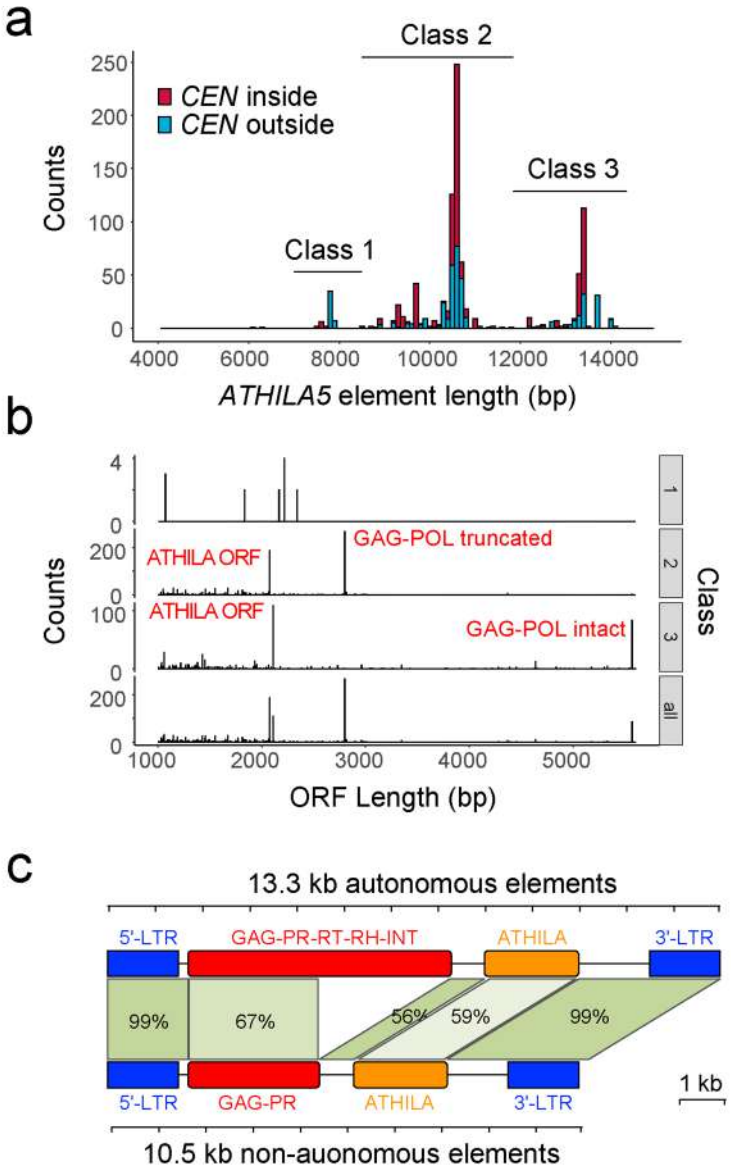

**Extended Data Fig. 7** | See next page for caption.

**Extended Data Fig. 7 | Centrophilic and centrophobic *ATHILA* in *A. thaliana*.** **a**, Pie charts of the proportions of centrophilic *ATHILA* families: (i) inside vs. outside the *AthCEN178* arrays, (ii) inside vs. outside *AthCEN178* arrays by chromosome, and (iii) intact vs. soloLTR located inside or outside the *AthCEN178* arrays. **b**, Phylogenetic trees constructed with full-length centromeric *ATHILA* from each chromosome. The clades representing different *ATHILA* families are indicated by background shading, and the coloured branch tips represent *AthCEN178* similarity groups. **c**, Representative FISH micrographs for *ATHILA2* (green) and *ATHILA5* (red) on pachytene chromosomes in the Col-0, Rab-1 and IP-Bus-0 accessions. Scale bars = 10 µm. **d**, *ATHILA* integration frequency along the length of the *CEN178* consensus repeat. **e**, Counts of intact (left) and soloLTR (right) *ATHILA* located outside (top) or inside (bottom) the *AthCEN178* arrays, ordered by chromosome arm SNP-PCA groups. **f**, Distribution of sequence identity between LTRs of intact *ATHILA* elements, comparing those located inside (red) or outside (blue) the centromeres, according to chromosome arm SNP-PCA group. Intact *ATHILA* within the *AthCEN178* arrays had significantly higher LTR identity in the Eurasians and Iberian non-relicts, compared with the Iberians and non-Iberian relicts (Wilcoxon tests all $P < 1.78 \times 10\text{-}6$).



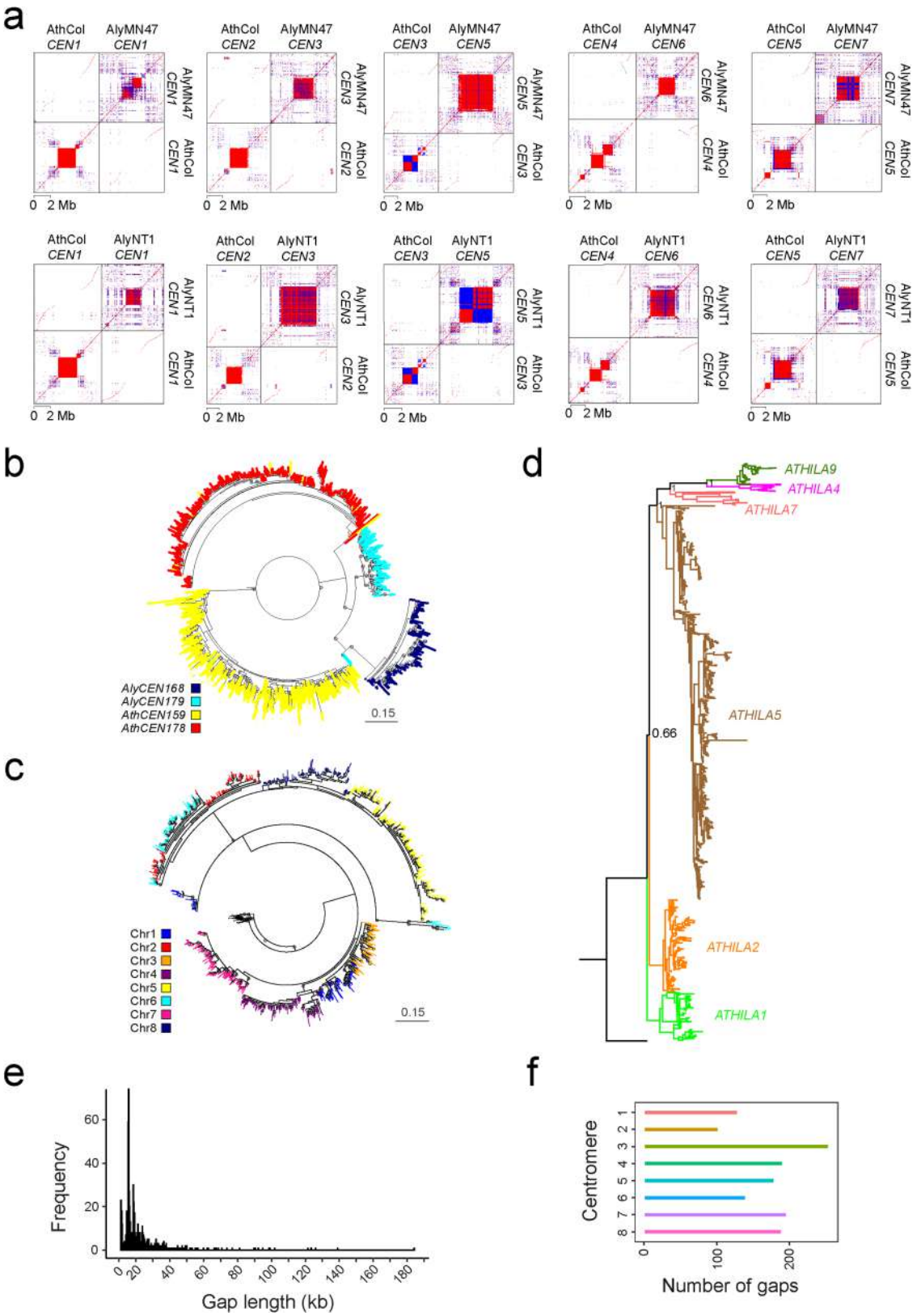

**Extended Data Fig. 8** | See next page for caption.

**Extended Data Fig. 8 | *ATHILA* diversification via *de novo* integration and intra-centromere duplication within *A. thaliana*.** **a**, StainedGlass sequence identity heat maps for ANGE-B-2 and ANGE-B-10 *CEN4*, with % GC content (green) and the density of *AthCEN178* per 10 kb on forward (red) and reverse (blue) strands plotted beneath. X-axis ticks indicate intact *ATHILA* (pink) and soloLTR (green) insertions. '*' marks insertions that are shared between ANGE-B-2 and ANGE-B-10, whereas '!' indicates those unique to ANGE-B-10. **b**, *CEN5* is shown for FERR-A-8 and FERR-A-12 with % GC content (green) and the density of *AthCEN178* per 10 kb on forward (red) and reverse (blue) strands shown beneath. X-axis ticks indicate intact *ATHILA* (pink) and soloLTR (green) insertions. '*' marks insertions that are shared between FERR-A-8 and FERR-A-12, whereas '!' marks those that correspond to post-integration duplications unique to FERR-A-8. **c**, StainedGlass sequence identity heat maps comparing FERR-A-8 and FERR-A-12 *CEN5*. **d**, The coverage of primary FERR-A-12 (dark blue) or FERR-A-8 (brown), and secondary FERR-A-12 (light blue) or FERR-A-8 (orange) alleles of PacBio HiFi reads to the chromosome 5 of the FERR-A-12 (upper), or FERR-A-8 (lower), genome assemblies. *AthCEN178* array coordinates, from Supplementary Table 3, are indicated by grey shading. Assembly gaps are shown by red shading. **e**, CENH3 $\log_2$(ChIP/input) values were plotted over *ATHILA* elements located within the *AthCEN178* arrays of the Col, Ler, Cvi and Tanz accessions (n = 100), in addition to 2 kb flanking regions. Windowed mean values are shown as solid lines, with 95% confidence intervals indicated by the shaded ribbons. This is compared to 100 randomly selected loci within the *AthCEN178* arrays, with the same widths as the *ATHILA*. Also shown are profiles across *ATHILA* elements located outside the *AthCEN178* arrays (n = 426), and Gypsy elements located outside the *AthCEN178* arrays (n = 21,487).




**Extended Data Fig. 9** | See next page for caption.

**Extended Data Fig. 9 | Autonomous and non-autonomous *ATHILA* elements in the collection of *A. thaliana* centromeres. a**, The size distribution of intact *ATHILA5* elements across the 66 *A. thaliana* accessions is plotted. Bar plots are coloured to indicate the number of elements inside (red), or outside (blue) the *AthCEN178* arrays. Three *ATHILA* size classes were defined; Class I for elements <8 kb, Class II between 8–12 kb, and Class III >12 kb. **b**, The distribution of ORF sizes (bp) in the *ATHILA5* elements, in total, or by the indicated size class. Red text indicates the position of the ATHILA-ORF and intact or truncated ORFs for GAG-POL. **c**, A representative diagram of an intact Class III 13.3 kb autonomous *ATHILA5* element, compared to a Class II 10.5 kb non-autonomous derivative. In this example, a single ~2.8 kb fragment that contains the reverse transcriptase, RNaseH and integrase genes is absent in the non-autonomous element. The green shaded areas indicate levels of sequence identity between the matching regions. **d**, Multiple sequence alignment of *ATHILA* integrase amino acid sequence from centrophilic (*ATHILA1, ATHILA2, ATHILA5, ATHILA6b*, rows 1–4) and centrophobic (*ATHILA0, ATHILA4c, ATHILA7, ATHILA9*, rows 5–8) families. The alignment starts immediately downstream of the RNase-H domain (not shown), to ensure that the N-terminus of integrase is included.




**Extended Data Fig. 10** | See next page for caption.

**Extended Data Fig. 10 | Phylogenetic analysis of centromere satellites and *ATHILA* in *A. thaliana* and *A. lyrata*.** **a**, Sequence identity dot plots comparing syntenic centromeres between *A. thaliana* Col (AthCol) and *A. lyrata* MN47 (AlyMN47), or NT1 (AlyNT1), using 80 bp windows. Red and blue indicate strand similarity (red is same, blue is opposite). **b**, Maximum-likelihood phylogenetic tree of *Arabidopsis* satellites, using randomly sampled *AlyCEN168* and *AlyCEN179* from *A. lyrata*, and *AthCEN178* and *AthCEN159* from six *A. thaliana* accessions (Bon-1, IP-Bus-0, IP-Alo-19, IP-Cas-6, Rab-1 and Tanz-1), and using *Capsella rubella* satellites as a root. Branch tips are coloured by satellite repeat family. A grey circle was placed on nodes where UFBoot support value exceeds 95%. **c**, A maximum-likelihood phylogenetic tree of *AlyCEN168* and 450 *AlyCEN179* satellites sampled from *A. lyrata* accession MN47. Thirty *AthCEN178* from the *A. thaliana* Col-0 accession were used as an outgroup. Tree tips are coloured according to chromosome, with the exception of the outgroup sequences, which are shaded in black. A grey circle is placed on nodes where UFBoot support value exceeds 95%. **d**, Phylogenetic tree of full-length *ATHILA* elements identified in *A. lyrata*. Elements were assigned to families based on their relationship to *A. thaliana ATHILA*, as shown in Fig. 4i. The tree was rooted using a maize *Huck Ty3* element. Bootstrap support is shown for key nodes. **e**, Distribution of lengths (kb) of non-satellite sequence gaps within the *A. lyrata* centromere satellite arrays. **f**, Total number of non-satellite gaps between 1 and 200 kb for each chromosome, across the two *A. lyrata* accessions MN47 and NT1.

nature portfolio

Corresponding author(s): Alexandros Bousios, Detlef Weigel and Ian Henderson

Last updated by author(s): Mar 23, 2023

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | The software used and versions are as follows: Primer3 (version 2.3.7), ImageJ (version 2.0.0), PacBio Circular Consensus Sequencing tool ccs (version 6.0.0), extracthifi (version 1.0), lima (version 2.0.0), Hifiasm (version 0.16.1-r375), RagTag (version 2.0.1), pbmm2 (version 1.9.0), samtools (version 1.9), Segmental Duplication Assembler (version 0.1.0), NucFreq (version 0.1), minimap2 (version 2.24), bcftools (version 1.15.1), R (version 4.3.0), DeepVariant (version 1.3.0), GLnexus (version 1.4.1), GATK (version 4.1.3.0), Beagle (version 4.0), TRASH (version 1.0), MAFFT (version 7.505), IQ-TREE (version 2.2.0), StainedGlass (version 0.5), ATHILAfinder (version 1.0), EMBOSS (version 6.5.0), HMMER (version 3.3.2), FastTree (version 2.0), FigTree (version 1.4.4), FraHMMER (version 3.3), EDTA (version 3.0), Liftoff (version 1.6.2), gffread (version 0.12.7), PGDspider (version 2.1.1.5), RAXML-NG (version 0.9.0), DeepSignal-plant (version 0.1.4), Filtlong (version 0.2.0), Tombo (version 1.5.1), Cutadapt (version 1.18), Bowtie2 (version 2.3.4.3) and deepTools (version 3.5.0). |
| Data analysis | The software used and versions are as follows: Primer3 (version 2.3.7), ImageJ (version 2.0.0), PacBio Circular Consensus Sequencing tool ccs (version 6.0.0), extracthifi (version 1.0), lima (version 2.0.0), Hifiasm (version 0.16.1-r375), RagTag (version 2.0.1), pbmm2 (version 1.9.0), samtools (version 1.9), Segmental Duplication Assembler (version 0.1.0), NucFreq (version 0.1), minimap2 (version 2.24), bcftools (version 1.15.1), R (version 4.3.0), DeepVariant (version 1.3.0), GLnexus (version 1.4.1), GATK (version 4.1.3.0), Beagle (version 4.0), TRASH (version 1.0), MAFFT (version 7.505), IQ-TREE (version 2.2.0), StainedGlass (version 0.5), ATHILAfinder (version 1.0), EMBOSS (version 6.5.0), HMMER (version 3.3.2), FastTree (version 2.0), FigTree (version 1.4.4), FraHMMER (version 3.3), EDTA (version 3.0), Liftoff (version 1.6.2), gffread (version 0.12.7), PGDspider (version 2.1.1.5), RAXML-NG (version 0.9.0), DeepSignal-plant (version 0.1.4), Filtlong (version 0.2.0), Tombo (version 1.5.1), Cutadapt (version 1.18), Bowtie2 (version 2.3.4.3) and deepTools (version 3.5.0). |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The genome assemblies analysed in this study are available under the following accession numbers: (i) 48 A. thaliana HiFi assemblies have been submitted to the European Nucleotide Archive (ENA) database under project number PRJEB55353 (ERP140242) (this study), (ii) 15 A. thaliana HiFi assemblies have been submitted to the ENA database under project number PRJEB55632 (ERA17524869), (iii) two A. thaliana HiFi assemblies (Col-0 and Ey15-2) are available at the ENA database under project number PRJEB50694 (ERP135313)7, (iv) one A. thaliana HiFi assembly (Kew-1) from the Darwin Tree of Life is available under project accession PRJEB5151114,15, and can also be accessed at: https://portal.darwintreeoflife.org/data/root/details/Arabidopsis%20thaliana, (v) ONT reads from the Ler-0, Cvi-0 and Tanz-0 accessions have been submitted as ArrayExpress accession E-MTAB-12009 (this study), while those for the accession Col-0 were previously available as ArrayExpress accession E-MTAB-102726, (vi) CENH3 Ilumina ChIP-seq reads from Col-0, Ler-0, Cvi-0 and Tanz-0 have been submitted as ArrayExpress accession E-MTAB-11974 (this study), and (vii) two A. lyrata HiFi assemblies are available at the ENA database under project number PRJEB50329 (ERP134897)27. Arabidopsis thaliana and Arabidopsis lyrata accessions used for sequencing are held in the authors' laboratories and seeds are freely available on request.

## Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

| | |
|---|---|
| Reporting on sex and gender | Not applicable. |
| Population characteristics | Not applicable. |
| Recruitment | Not applicable. |
| Ethics oversight | Not applicable. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | Centromeres are critical for cell division, loading CENH3/CENPA histone variant nucleosomes, directing kinetochore formation and allowing chromosome segregation. Despite their conserved function, centromere size and structure are diverse across species. We assembled 346 centromeres from 66 Arabidopsis thaliana and two A. lyrata accessions, which revealed a remarkable degree of intra- and inter-species diversity. |
| Research sample | We assembled 346 centromeres from 66 Arabidopsis thaliana and two A. lyrata accessions. The choice of accessions was made to sample intra- and inter-species diversity for these species. Arabidopsis thaliana (11C1, ANGE-B-10, ANGE-B-2, AUZE-A-5, BANI-C-1, BANI-C-12, BARA-C-3, BARA-C-5, BARC-A-12, BARC-A-17, BELC-C-10, BELC-C-12, Bon-1, BROU-A-10, CAMA-C-2, CAMA-C-9, Col-0, DraIII-1, Ey15-2, FERR-A-12, FERR-A-8, GAIL-B-11, Gel-1, HR-10, IP-Hum-4, Iasi-1, IP-Fel-2, IP-Piq-0, Kew-1, LACR-C-14, Ler-0, MERE-A-13, MONTM-B-16, MONTM-B-7, PREI-A-14, Ru-2, SALE-A-10, SALE-A-17, T850, IP-Alo-0, IP-Cas-0, IP-Evs-0, IP-Evs-12, IP-Hom-4, IP-Bus-0, IP-Ini-0, IP-Tri-0, IP-Mdc-14, IP-Med-3, IP-Met-6, IP-Mos-5, IP-Alo-19, IP-Cas-6, IP-Cat-0, IP-Hom-0, IP-Hum-2, IP-Per-0, IP-Lor-16, IP-Med-0, IP-Mos-9, IP-San-9, IP-Sln-22, Cvi-0, Etna-2, Rabacal-1 and Tanz-1) and Arabidopsis lyrata (MN47 and NT1) accessions used for sequencing are held in the authors' laboratories and seeds are freely available on request. |
| Sampling strategy | Accessions were sampled from across the native range of Arabidopsis thaliana and A. lyrata. |
| Data collection | Genomic DNA was extracted from these samples and used for long-read DNA sequencing and genome assembly. In addition, FISH, DNA methylation and ChIP-seq experiments were performed on a subset of the sample. |
| Timing and spatial scale | Not applicable. |

| | |
|---|---|
| Data exclusions | No data were excluded. |
| Reproducibility | For FISH experiments 50 independent nuclei were analysed per genotype and per probe set. |
| Randomization | Not applicable. |
| Blinding | Not applicable. |

Did the study involve field work? ☐ Yes ☒ No

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Antibodies

| | |
|---|---|
| Antibodies used | The CENH3 antibody was used at a dilution of 1:400 and this information has been added to the Methods section. This is a polyclonal antibody raised against Arabidopsis CENH3 that was provided as a gift from Prof. Steven Henikoff (Fred Hutchison Institute, Seattle, USA). |
| Validation | Generation and validation of the Arabidopsis CENH3 antibody was first reported in the following publication: Talbert, P. B., Masuelli, R., Tyagi, A. P., Comai, L. & Henikoff, S. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. Plant Cell 14, 1053–1066 (2002). Use of this antibody for ChIP-seq in Arabidopsis is reported in the following publication: Maheshwari S, Ishii T, Brown CT, Houben A, Comai L. Centromere location in Arabidopsis is unaltered by extreme divergence in CENH3 protein sequence.Genome Res. 2017 27:471-478. |

## ChIP-seq

### Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

| | |
|---|---|
| Data access links<br>*May remain private before publication.* | CENH3 Ilumina ChIP-seq reads from Col-0, Ler-0, Cvi-0 and Tanz-0 have been submitted as ArrayExpress accession E-MTAB-11974. |
| Files in database submission | Col_0_CenH3_ChIP_R2.fastq.gz<br>Col_0_CenH3_ChIP_R1.fastq.gz<br>Col_0_CenH3_Input_R1.fastq.gz<br>Col_0_CenH3_Input_R2.fastq.gz<br>Cvi_0_CenH3_ChIP_R1.fastq.gz<br>Cvi_0_CenH3_ChIP_R2.fastq.gz<br>Cvi_0_CenH3_Input_R1.fastq.gz<br>Cvi_0_CenH3_Input_R2.fastq.gz<br>Ler_0_CenH3_ChIP_R1.fastq.gz<br>Ler_0_CenH3_ChIP_R2.fastq.gz<br>Ler_0_CenH3_Input_R2.fastq.gz<br>Ler_0_CenH3_Input_R1.fastq.gz<br>Tanz_1_CenH3_ChIP_R2.fastq.gz<br>Tanz_1_CenH3_ChIP_R1.fastq.gz<br>Tanz_1_CenH3_Input_R1.fastq.gz |

| | |
|---|---|
| | Tanz_1_CenH3_Input_R2.fastq.gz |
| Genome browser session (e.g. UCSC) | No longer applicable. |

## Methodology

| | |
|---|---|
| Replicates | Each genotype was analysed by ChIP and sequencing once. |
| Sequencing depth | All reads were paired end 150 bp.<br>Col-0 ChIP 25,981,323 reads, equivalent to 59.96x coverage<br>Col-0 ChIP 69,421,376 reads, equivalent to 160.20x coverage<br>Ler-0 ChIP 22,764,153 reads, equivalent to 52.53x coverage<br>Ler-0 Input 65,461,600 reads, equivalent to 151.07x coverage<br>Tanz-1 ChIP 24,400,868 reads, equivalent to 56.31x coverage<br>Tanz-1 Input 70,378,959 reads, equivalent to 162.41x coverage<br>Cvi-0 ChIP 17,888,402 reads, equivalent to 41.28x coverage<br>Cvi-0 input 69,092,728 reads, equivalent to 159.44x coverage |
| Antibodies | The CENH3 antibody was used at a dilution of 1:400 and this information has been added to the Methods section. This is a polyclonal antibody raised against Arabidopsis CENH3 that was provided as a gift from Prof. Steven Henikoff (Fred Hutchison Institute, Seattle, USA). |
| Peak calling parameters | Our analysis did not involve peak calling. |
| Data quality | To ensure accuracy in the reported mapping profile across the repetitive centromeric regions, the short read manning parameters were tested using ~26M synthetic 2x150bp illumina reads that were generated from the centromeric array sequences of each accession (dwgsim https://github.com/nh13/DWGSIM). Using the parameters detailed in the methods, the modelled mapping accuracy within the centromere of the synthetic datasets was 97.5%. |
| Software | Deduplicated paired-end CENH3 ChIP-seq Illumina reads (2×150 bp) from Col-0, Cvi-0, Ler-0 and Tanz-1 were processed with Cutadapt (version 1.18) to remove adapter sequences and low-quality bases (Phred+33-scaled quality <20)74. For each accession, trimmed reads were aligned to the respective genome assembly using Bowtie2 (version 2.3.4.3), using the following settings: --very-sensitive --no-mixed --no-discordant -k 10 --maxins 800. Up to 10 valid alignments were reported for each read pair. Read pairs with Bowtie2-assigned MAPQ <10 were discarded using samtools (version 1.10). 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. 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 (version 3.5.0). Reads that aligned to the chloroplast or mitochondrial genomes were excluded from coverage normalisation. For profiling of CENH3 occupancy within AtCEN178 repeats, trimmed reads were aligned to their respective genome assembly using Bowtie2 (version 2.3.4.3), using the following settings: --very-sensitive --no-mixed --no-discordant -k 200 --maxins 800. Read pairs with Bowtie2-assigned MAPQ <2 were discarded using samtools (version 1.10). 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. Alignment bedgraphs were converted to per-base 1-based coverage files, used to calculate log2(ChIP/Input), and custom R scripts were used to plot the average profiles across AtCEN178 and ATHILA elements. |