Rapid cycles of satellite homogenization and retrotransposon invasion drive Arabidopsis centromere evolution

Authors:

Piotr Wlodzimierz[1]*, Fernando A. Rabanal[2]*, Robin Burns[1]*, 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], Max Collenberg[2], Miriam Mielke[2], Magnus Nordborg[6], Fabrice Roux[7], Gautam Shirsekar[2], Carlos Alonso-Blanco[8], Martin A. Lysak[5], Polina Novikova[4], Alexandros Bousios[3†], Detlef Weigel[2†] and Ian R. Henderson[1†]

Affiliations:

[1] Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom

[2] Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Tübingen, Germany

[3] School of Life Sciences, University of Sussex, Brighton, United Kingdom

[4] Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany

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

[6] Gregor Mendel Institute, Vienna, Austrian Academy of Sciences, Vienna BioCenter, Dr. Bohr-Gasse 3, 1030, 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

* Equal contribution.

† **Correspondence:** irh25@cam.ac.uk, weigel@tue.mpg.de and alex.bousios@sussex.ac,uk

Summary Paragraph

Centromeres are critical for cell division, loading CENH3/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 two *A. lyrata* accessions, which revealed a remarkable degree of intra- and inter-species diversity. *Arabidopsis thaliana* centromere repeat arrays are embedded in linkage blocks, despite ongoing internal satellite turnover, demonstrating a role for unidirectional gene conversion in sequence diversification, rather than interhomolog unequal crossover. 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 transposon*s*, consistent with cycles of repeat evolution. Centromeric sequence changes are even more extreme in comparison between *A. thaliana* and *A. lyrata*. Together, our findings reveal rapid cycles of transposon invasion and purging via satellite homogenization, which drive centromere evolution and ultimately contribute to speciation.

Main text

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), have now made complex centromere assembly possible, allowing, for example, complete telomere-to-telomere (T2T) reference maps of human 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 *Arabidopsis thaliana*, which has megabase-scale satellite arrays that resemble human alpha-satellite centromeres[6,8,9]. *Arabidopsis 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 with centromeres from the sister species *A. lyrata*, which has a sparser, circumpolar distribution[13], to contrast patterns of intra- and inter-species centromere evolution.

Towards assembling the Arabidopsis pancentromere

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 an N50 of 12.7-24.4 kbp[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-1b** and **Supplementary Tables 1-2**). Notably, three pairs of

accessions had nearly identical chromosome-arm SNPs, providing an exceptional opportunity to study short-term centromere evolution in natural populations (**Supplementary Table 2**).

Genetic diversity within the Arabidopsis centromere satellite repeats

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 are diverse, with almost one quarter (1,234,281) being unique in our sample. The number of non-redundant *AthCEN178* sequences began to plateau as more accessions were included, consistent with deep sampling of the satellome (**Extended Data Fig. 2a**). The level of pairwise *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**). Based on pairwise 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 were observed that did not geographically relate to the chromosome-arm SNP groups, although satellite sharing decreased with increasing geographic separation (r=-0.28, P=<2.2×10^{-16}) (**Fig. 1a** and **Extended Data Fig. 2c-2d**). Sharing of *AthCEN178* similarity groups across all five chromosomes was very rare (**Extended Data Fig. 2e**), consistent with independent segregation of centromeres after outcrossing. Of the four pairs (of 2,145) that shared *AthCEN178* similarity across all five chromosomes, three had near-identical chromosome-arm SNPs, indicating a very recent common ancestor (**Extended Data Fig. 2e**). 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 span 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 Movie 1**). While each chromosome exhibits *AthCEN178* strand biases, there are complete strand reversals and inversions, especially on chromosome five (**Fig. 1f**). The SNP-PCA genetic groups differed significantly in satellite array size, where total *AthCEN178* content was higher in the relicts than in the Eurasian accessions (Wilcoxon test $P=1.17\times10^{-4}$) (**Extended Data Fig. 3a** and **Supplementary Table 2**). We independently confirmed this difference using FISH, where a significantly greater *AthCEN178* centromeric signal was observed in Tanz-1 (relict), compared to the Col-0 (Eurasian) accession (Wilcoxon tests all $P=<1.04\times10^{-6}$) (**Fig. 1g** and **Extended Data Fig. 3b-3c**). *AthCEN178* repeats sampled from contrasting Eurasian and relict backgrounds revealed stronger phylogenetic clustering by chromosome rather than by accession (**Fig. 1h**), further consistent with limited recombination and satellite sequence sharing between chromosomes. Furthermore, *AthCEN178* sequences that were unique to a single accession were frequently at high copy number within their chromosome, whereas *AthCEN178* 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 ChIP-seq in the Col-0, Ler-0, Cvi-0 and Tanz-1 accessions and compared $\log_2$(ChIP/input) enrichment over the consensus repeat to the frequency of *AthCEN178* sequence variants (**Extended Data Fig. 3e**). The region of high *AthCEN178* variation corresponds to the left edge of CENH3 occupancy within the repeat, consistent with an effect of chromatin structure on mutation rate or persistence of satellite variants.

Satellite higher order repeats and dynamic centromere evolution

An important mode of centromere satellite evolution is via 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, between each monomer pair. This approach can assign a single *AthCEN178* copy to multiple HORs. *AthCEN178* HORs were abundant (total=98,265,546) in our collection of centromere satellite arrays, with a mean of 1,488,872 HORs per accession (**Fig. 2a** and **Supplementary Tables 2** & **4**). HOR lengths showed a negative exponential distribution, with three-mers being most frequent (**Fig. 2a**), consistent with homologous recombination over a distance of ~534 bp. 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 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 it belonged to and divided this by the total number of *AthCEN178* on that chromosome, yielding a HOR score. Across all 330 centromeres, the core *AthCEN178* arrays (between 0.2 and 0.8 of their scaled length) had greatest HOR scores and lowest edit distances (**Fig. 2b** and **Extended Data Fig. 4a**). Within chromosomes, single-copy *AtCEN178* were more frequenct at the periphery of the satellite arrays, whereas the core regions were composed of high-copy *AtCEN178* (**Extended Data Fig. 4b**). Together, this indicates that satellite recombination and HOR formation are frequent, yet spatially focused, within the core of the *A. thaliana* centromere satellite arrays. For example, a ~1 Mb region within the *AthCEN178* array of Ey15-2 chromosome 1 (*CEN1*) featured elevated HOR scores that correlated with decreased edit distances (**Fig. 2c**).

Across our *A. thaliana* sample, chromosomes had heterogeneous *AthCEN178* HOR scores (**Extended Data Fig. 4c-4d**), consistent with homogenization occurring independently in different arrays. To detect recent centromere evolution, we compared satellite arrays within *AthCEN178* similarity groups that were embedded in cenhaps (**Fig. 1d, 2d, Extended Data Fig. 4e** and **Supplementary Table 2**). For example, we considered *CEN1* similarity group 2, which includes seven French and two Iberian accessions (**Fig. 1d, 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 is related to a smaller ~150 kb array in IP-Ini-0, yet surrounded by highly similar regions (**Fig. 1e**, **2d** and **Extended Data Fig. 4f**). These accessions share 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 sharing of an almost identical *ATHILA*, combined with the extensive *AthCEN178* HOR similarities flanking the polymorphic region (**Fig. 2d** and **Extended Data Fig. 4f**), indicates 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 occur embedded in cenhap linkage blocks, were seen for all chromosomes (**Extended Data Fig. 4d**). Together, this supports a unidirectional gene conversion model for centromere satellite evolution, rather than unequal crossovers between homologs[19–21].

Although the majority of *AthCEN178* HORs were short (<10 kb) (**Fig. 2a**), less frequent large-scale (>100 kb) duplications were also observed (**Fig. 2e-2f**). To identify such 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 (mean length=367 kb, or ~2,062 *AthCEN178* copies) and present on all chromosomes (**Fig. 2e**). For example, *CEN5* in MERE-A-13 contains 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.

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 that across the Eurasian Col-0 and Ler-0, and relict Cvi-0 and Tanz-1 accessions, *AthCEN178* array sizes (~1.5-6.5 Mb) were significantly greater than the regions of CENH3 ChIP-seq enrichment (~1-2 Mb) (Wilcoxon test $P=2.95\times10^{-7}$) (**Fig. 2g**). This suggests homeostasis of CENH3 loading within *AthCEN178* satellite arrays of varying sizes. At ten

centromeres, CENH3 preferentially occupied only one of the distinct *AthCEN178* arrays present (**Fig. 2h** and **Extended Data Fig. 5**), consistent 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 is 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 obtained profiles of centromeric DNA methylation from ONT sequencing reads for the same accessions, at 220-460× depth and with an N50 between 6.3-26.7 kb (**Supplementary Table 1**). Despite extensive satellite structural diversity, all CENH3-enriched regions were severely depleted of DNA methylation in the CHG context compared to flanking pericentromeric heterochromatin, with a more modest decrease in the CG context (**Fig. 2h** and **Extended Data Fig. 5**). Together this is consistent with complex interactions between genetic and epigenetic information and CENH3 enrichment within the *A. thaliana* centromeres.

Recent invasion of the satellite repeat arrays by centrophilic *ATHILA* retrotransposons

Assembly of the Col-0 reference strain revealed 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 sequence within the arrays (**Supplementary Tables 2** & **5**). The intact-to-soloLTR ratio was substantially higher inside the satellite arrays than outside (2.5 vs. 0.6) (**Supplementary Table 5**), consistent with increased *ATHILA* integration, or reduced soloLTR formation. Of the 14 identified *ATHILA* families, only five were frequently observed within the *AthCEN178* arrays, especially *CEN5*, with *ATHILA5* being the most 'centrophilic' (**Fig. 3a-3b, Extended Data Fig. 7a-7b** and **Supplementary Table 5**). The *ATHILA*

within the *AthCEN178* arrays were significantly younger than those in the chromosome arms (all Wilcoxon test $P<1.57\times10^{-8}$), with *ATHILA5* being the youngest (**Fig. 3c** and **Supplementary Table 5**). Using FISH against *ATHILA2* and *ATHILA5*, we confirmed that *ATHILA5* probes produced the most centralised signal within the DAPI-dense chromocenters (**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. *ATHILA* have integrated throughout the length of the *AthCEN178* repeat, although highest levels were observed between positions 50 and 100, and 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-7b**), which may arise at the level of integration, or differential post-integration maintenance of *ATHILA* within each satellite array. One of seven intact *ATHILA* elements (n=1,296) encoded all major retrotransposon proteins, of which 1,012 also contained the ATHILA-specific ORF[22] (**Supplementary Table 5**). These complete *ATHILA* elements are likely to be competent for autonomous transposition and are over-represented in the satellite repeat arrays (21.8% inside versus 12.7% outside). Together, these patterns indicate widespread colonisation of the *A. thaliana* satellite arrays by centrophilic *ATHILA*.

In addition to possessing smaller *AthCEN178* satellite arrays than the relicts, the Eurasian accessions have experienced a greater level of centrophilic *ATHILA* integration (Wilcoxon test $P=1.46\times10^{-7}$) (**Extended Data Fig. 7e**). Furthermore, intact *ATHILA* elements within the *AthCEN178* arrays had significantly higher LTR identity in the Eurasian and Iberian non-relict accessions, compared with Iberian and non-Iberian relicts (Wilcoxon tests all $P<1.78\times10^{-6}$) (**Extended Data Fig. 7f**), suggesting 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 are the ANGE-B-2 and ANGE-B-10 accessions, which were collected within five metres of one another in France. They contained a large number of *ATHILA* insertions in *CEN4* and *CEN5* and shared very similar *AthCEN178* arrays (**Fig. 3e, Extended Data Fig. 8a** and **Supplementary Tables 2** & **5**). All 19 *ATHILA* within the *CEN4* and *CEN5 AthCEN178* arrays of ANGE-B-2 could be

unambiguously matched to syntenic insertions in ANGE-B-10, based on 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, consistent with very recent integration (**Fig. 3e, Extended Data Fig. 8a** and **Supplementary Tables 2** & **6**). Few mutations separated the shared *ATHILA* copies between the ANGE accession pair (1 difference per 2,000 bp over 172 kb of syntenic *ATHILA* sequence), suggesting a recent common ancestry of *CEN4* and *CEN5 ATHILA* (**Supplementary Table 6**). The new ANGE-B-10 *ATHILA5* elements included non-autonomous derivatives that harbour a ~3 kb deletion of the reverse transcriptase, RNase-H and integrase genes (**Extended Data Fig. 9**). Phylogenetic analysis revealed that both autonomous and non-autonomous *ATHILA5* sub-lineages are present throughout the chromosome-arm PCA groups, reflecting dynamic radiation within the population (**Fig. 3f**). Our data thus provide evidence for recent *ATHILA5* invasion of the *A. thaliana* centromere satellite arrays in Eurasia.

We also observed transposon polymorphism originating through large-scale intra-centromere duplication of the satellite arrays. For example, a ~2 Mb duplication distinguished *CEN5* of FERR-A-8 and FERR-A-12, which copied three intact *ATHILA* and one soloLTR (**Extended Data Fig. 8b-8d** and **Supplementary Table 6**). Despite the internal *AthCEN178* duplication, these satellite arrays belong to the same cenhap block, defined by flanking polymorphisms (**Fig. 1e**). This is inconsistent with unequal crossover between homologs generating the duplication, and instead favours a unidirectional gene conversion pathway. This analysis reveals that *ATHILA* copies may increase both by transposition, and also via intra-satellite array duplications. However, as the *ATHILA* 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* 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 HOR scores, which was not significant

(*P*=0.128) (**Fig. 3g**). As *ATHILA5* are the youngest and most centrophilic family, insufficient time may have occurred for satellite homogenization to remove the new copies, as seen in ANGE-B-10 (**Fig. 3e**). Consistently, 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 to the elements shared with ANGE-B-2 (**Fig. 3h**). Alternatively this may reflect varying spatial integration biases of 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* located in the *AthCEN178* 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 where *AthCEN178* homogenization pathways act to purge *ATHILA* transposons within the *A. thaliana* centromere satellite arrays.

Rapid interspecies centromere evolution between *A. thaliana* and *A. lyrata*

As we observed extreme *A. thaliana* intra-species variation in satellite array size and structure, we sought to compare the extent of inter-specific 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[23–26]. 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 two species (**Fig. 4a** and **Extended Data Fig. 10a**), consistent 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[25,27].

We identified large *A. lyrata* repeat arrays in the expected centromere locations[25,27], which contain two satellite populations: *AlyCEN168* (~168 bp) and *AlyCEN179* (~179 bp) (**Fig. 4a,b**). Different centromeres were predominantly composed of a single repeat class, with *AlyCEN168* dominating *CEN2*, *5*, *6* and *8*, while *AlyCEN179* dominated *CEN1*, *3*, *4* and *7* (**Fig. 4c**). These distributions are

consistent with FISH analyses, where the pAge and pAa probes distinguished the same chromosome groups[27]. We generated phylogenetic trees and sequence identity heatmaps, which revealed that *A. thaliana AthCEN178* are more closely related to *AlyCEN168* and *AlyCEN179*, while *AthCEN159* has 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 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 are similar in size and both have higher GC 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* vs *AlyCEN179*), the size and structure of the arrays were highly polymorphic between the accessions (**Fig. 4c** and **Extended Data Fig. 10a**), consistent 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 *A. lyrata* chromosome arms (**Fig. 4e, 4g-4h** and **Extended Data Fig. 10d-10f**). For example, 61% of intact *A. lyrata ATHILA* were within the centromere satellite arrays, compared to only 15% in *A. thaliana* (**Supplementary Table 5**). *ATHILA5* is the most abundant family in the *A. lyrata* centromere satellite arrays (679/994 intact elements), which were also more centrophilic (81%) than in *A. thaliana* (56%), and were found to have integrated throughout the length of the centromere satellite arrays (**Fig. 4g-4i** 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* in two *A. lyrata* accessions, versus 16 across all 66 *A. thaliana* genomes), only six of which were related to the known centrophilic *Tal1* element[28]. This provides evidence for dynamic change in centromeric retrotransposon populations between *A. thaliana* and *A. lyrata*, following speciation.

A cyclical model for *Arabidopsis* centromere evolution

Intra-species polymorphism in the *A. thaliana* centromere satellite repeat arrays points to recurrent cycles of expansion and contraction (**Fig. 5a**). We also observed polymorphism in *ATHILA* insertions within the satellite arrays, consistent with ongoing centrophilic retrotransposon invasion. We hypothesise that *ATHILA* integrases have evolved to recognise centromeric DNA sequences, or associated chromatin states that may include CENH3[29] (**Fig. 5a**), even though we could not detect diagnostic variants in integrase that distinguished centrophilic and centrophobic families (**Extended Data Fig. 9d**). To counter this invasive activity, satellite recombination acts to eliminate *ATHILA* (**Fig. 5a-5b**). Evidence for centromeric satellite DNA breakage and recombination exists in several species, including humans and maize[18,21,30–33]. During meiosis, either a homolog or a sister chromosome could be used as a repair template, while the sister chromosome is most likely 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 centromere array change (**Fig. 5b**). Alternatively, unequal crossover between sister chromatids could create satellite polymorphism, without a detectable effect on linkage of flanking markers. As a consequence of homologous recombination, the Arabidopsis centromere satellites show signatures consistent with sequence homogenization[17], reminiscent of concerted evolution in diverse other satellite and tandem repeat arrays[34–37].

The human kinetochore is hypothesised to recruit recombination factors that promote satellite evolution, termed the Kinetochore Associated Recombination Machine (KARM)[5,38]. We propose that an equivalent activity, 'KARM in Arabidopsis' (KARMA), promotes *AthCEN178* recombination, contributing to centromere homogenization, large-scale intra-centromere duplications and purging of *ATHILA* (**Fig. 5a-5b**). The differences in size and structure of the centromere satellite arrays between the *A. lyrata* accessions indicates that KARMA is likely to operate in this species as well, but may be less effective at removing transposons, given the higher *ATHILA* load and the presence of *A. lyrata*-specific centrophilic *COPIA* elements. Variation in these processes is also likely to influence

centromeric differentiation between the main genetic lineages of *A. thaliana*, including larger satellite repeat arrays in the relict accessions, and greater invasion of the satellite arrays by *ATHILA* in Eurasia. However, since non-centromeric satellites show concerted evolution[36,37], kinetochore-independent recombination pathways must also exist in tandem repeat arrays located outside of the centromere.

Centromere drive during asymmetric female meiosis – as in *A. thaliana*, where only a terminal meiocyte survives[39] – has been proposed to cause rapid evolution of centromere DNA and CENH3/CENPA protein sequences[40–42]. Within *A. thaliana*, we observe high satellite array diversity, yet limited CENH3 amino acid variation. Hence, we do not consider competition between CENH3 amino acid variants as a cause of the extreme intra-species satellite diversity. Instead, recombination events, potentially mediated via KARMA, could promote rapid satellite evolution within *A. thaliana* (**Fig. 5**). We propose that CENH3 location will 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 intraspecific diversity of centromere satellite arrays implies that specific variants are unable to dominate *A. thaliana* populations. Instead, conditional, frequency-dependent advantages may generate balanced satellite polymorphism and maintain high levels of intraspecies diversity. As centromere drive is widespread in plants and animals[40–42], exploring mechanisms of centromere competition within and between *Arabidopsis* species may prove to be illuminating for eukaryotic genome evolution and speciation.

References

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 in. *Proc. Natl. Acad. Sci. U. S. A.* **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. U. S. A.* **119**, (2022).
15. Christenhusz MJM, Twyford AD, Hudson A, Royal Botanic Gardens Kew Genome Acquisition Lab, Royal Botanic Garden Edinburgh Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium. The genome sequence of thale cress, Arabidopsis thaliana

(Heynh., 1842). *Wellcome Open Res 2023,* **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**, (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. Smith, G. P. Evolution of Repeated DNA Sequences by Unequal Crossover. *Science* vol. 191 528–535 Preprint at https://doi.org/10.1126/science.1251186 (1976).
20. Talbert, P. B. & Henikoff, S. Centromeres convert but don't cross. *PLoS Biol.* **8**, e1000326 (2010).
21. Shi, J. *et al.* Widespread gene conversion in centromere cores. *PLoS Biol.* **8**, e1000327 (2010).
22. Slotkin, R. K. The epigenetic control of the Athila family of retrotransposons in Arabidopsis. *Epigenetics* **5**, 483–490 (2010).
23. 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).
24. 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).
25. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
26. Kolesnikova, U. *et al.* Genome of selfing Siberian Arabidopsis lyrata explains establishment of allopolyploid Arabidopsis kamchatica. *bioRxiv* 497443 (2022).
27. 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).
28. Tsukahara, S. *et al.* Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata. *Genes Dev.* **26**, 705–713 (2012).
29. Malik Harmit S. & Eickbush Thomas H. Modular Evolution of the Integrase Domain in the Ty3/Gypsy Class of LTR Retrotransposons. *J. Virol.* **73**, 5186–5190 (1999).

30. Nijman, I. J. & Lenstra, J. A. Mutation and Recombination in Cattle Satellite DNA: A Feedback Model for the Evolution of Satellite DNA Repeats. *Journal of Molecular Evolution* **52,** 361–371 (2001).

31. Chatterjee, B. & Lo, C. W. Chromosomal recombination and breakage associated with instability in mouse centromeric satellite DNA. *Journal of Molecular Biology* **210,** 303–312 (1989).

32. Wolfgruber, T. K. *et al.* High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events. *Front. Plant Sci.* **7**, 308 (2016).

33. 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).

34. 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).

35. Durfy, S. J. & Willard, H. F. Concerted evolution of primate alpha satellite DNA. Evidence for an ancestral sequence shared by gorilla and human X chromosome alpha satellite. *J. Mol. Biol.* **216**, 555–566 (1990).

36. 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).

37. 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).

38. 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 alpha satellites flanking the active centromeres of human chromosomes. *PLoS Genet.* **5**, e1000641 (2009).

39. 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).

40. Akera, T., Trimm, E. & Lampson, M. A. Molecular Strategies of Meiotic Cheating by Selfish

Centromeres. *Cell* **178**, 1132–1144.e10 (2019).

41. Fishman, L. & Saunders, A. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science* **322**, 1559–1562 (2008).
42. Kursel, L. E. & Malik, H. S. The cellular mechanisms and consequences of centromere drive. *Curr. Opin. Cell Biol.* **52**, 58–65 (2018).
43. 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).

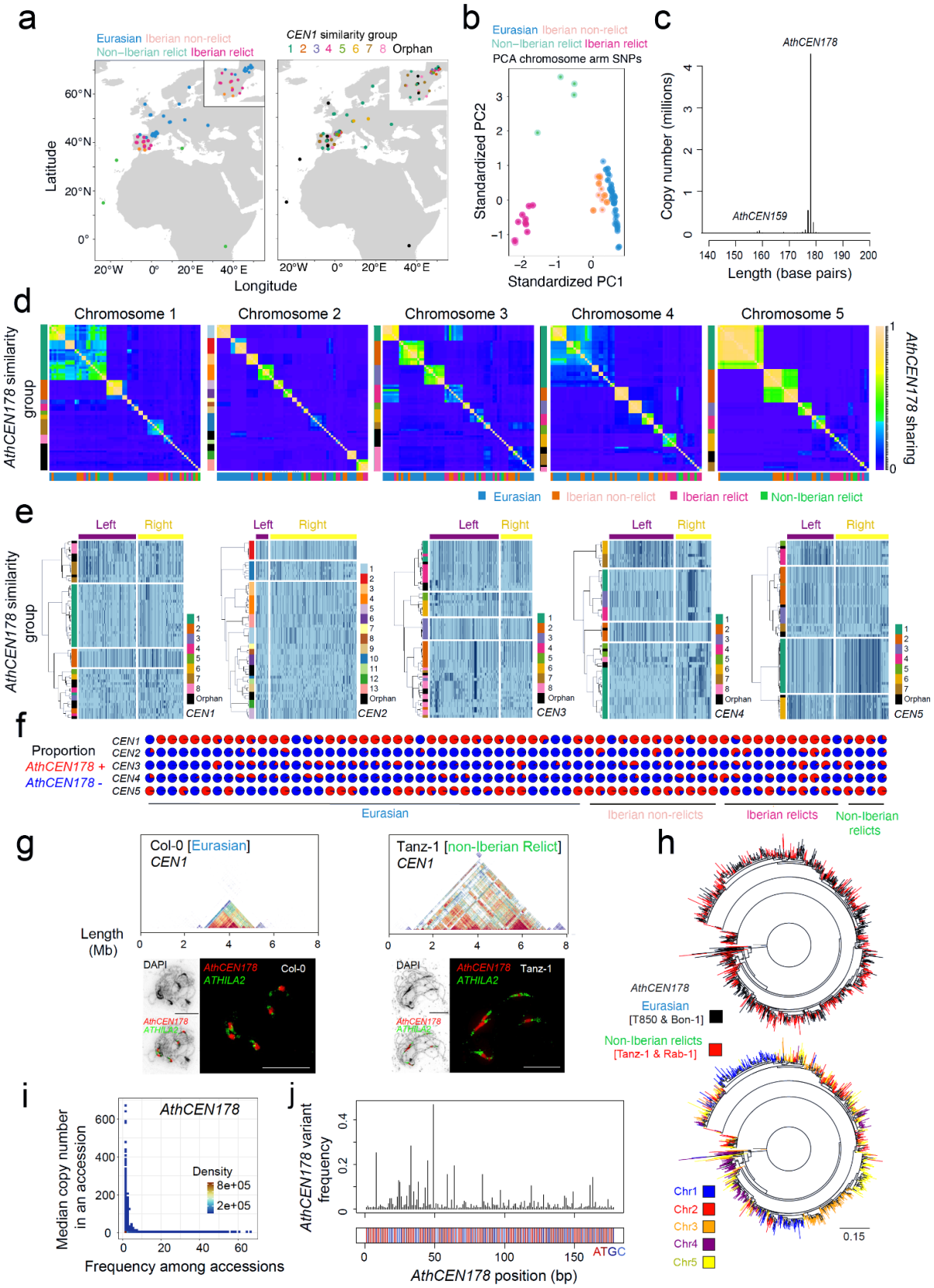

a
Eurasian Iberian non-relict
Non−Iberian relict Iberian relict
CEN1 similarity group
1 2 3 4 5 6 7 8 Orphan
Latitude
Longitude
b
Eurasian Iberian non-relict
Non−Iberian relict Iberian relict
PCA chromosome arm SNPs
Standardized PC2
Standardized PC1
c
AthCEN178
AthCEN159
Copy number (millions)
Length (base pairs)
d
Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
AthCEN178 similarity group
AthCEN178 sharing
Eurasian
Iberian non-relict
Iberian relict
Non-Iberian relict
e
Left
Right
AthCEN178 similarity group
Orphan
CEN1
CEN2
CEN3
CEN4
CEN5
f
Proportion
AthCEN178 +
AthCEN178 -
Eurasian
Iberian non-relicts
Iberian relicts
Non-Iberian relicts
g
Col-0 [Eurasian]
CEN1
Tanz-1 [non-Iberian Relict]
Length (Mb)
DAPI
AthCEN178
ATHILA2
Col-0
Tanz-1
h
AthCEN178
Eurasian
[T850 & Bon-1]
Non-Iberian relicts
[Tanz-1 & Rab-1]
Chr1
Chr2
Chr3
Chr4
Chr5
0.15
i
AthCEN178
Median copy number in an accession
Density
8e+05
2e+05
Frequency among accessions
j
AthCEN178 variant frequency
ATGC
AthCEN178 position (bp)

Figure 1. High genetic diversity in the *A. thaliana* pancentromere. a, Geographic origin of 66 *A. thaliana* accessions, coloured according to principle component analysis (PCA) group membership, as shown in (b). The geographic origin of *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 (bp) across 66 accessions. The *AthCEN178* and *AthCEN159* families are labelled. **d,** Heat maps showing % of identical *AthCEN178* repeats shared between each pair of chromosomes, with the colour scale on the right. The x-axis key shows PCA group 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. The dendrogram tips are coloured according to *AthCEN178* similarity group, as in (d). **f,** Pie charts showing proportion of *AthCEN178* on forward (red) or reverse (blue) strands for each chromosome (rows) and accession (columns). **g,** StainedGlass sequence identity heatmaps for Col-0 (Eurasian) and Tanz-1 (Relict) *CEN1*. Representative FISH micrographs probed for *AthCEN178* (red) and *ATHILA2* (green) on pachytene chromosomes of Col-0 and Tanz-1 are shown beneath. Inset are DAPI-stained images, and overlays, of the same cells. Scale bars=10 μm. **h,** Maximum likelihood *AthCEN178* phylogenetic tree, sampled equally from Eurasian (T850 and Bon-1) and non-Iberian relict (Tanz-1 and Rab-1) accessions, and rooted using *Capsella rubella* satellites[43]. The same tree is shown with branches coloured according to PCA group (upper), or by chromosome (lower). Scale=0.15. **i,** Median *AthCEN178* copy number in an accession plotted against the number of accessions in which they are found. **j,** Frequency of *AthCEN178* sequence variants across the consensus repeat. A colour-coded DNA sequence is shown beneath (A=red, T=dark red, G=navy blue, C=azure blue).

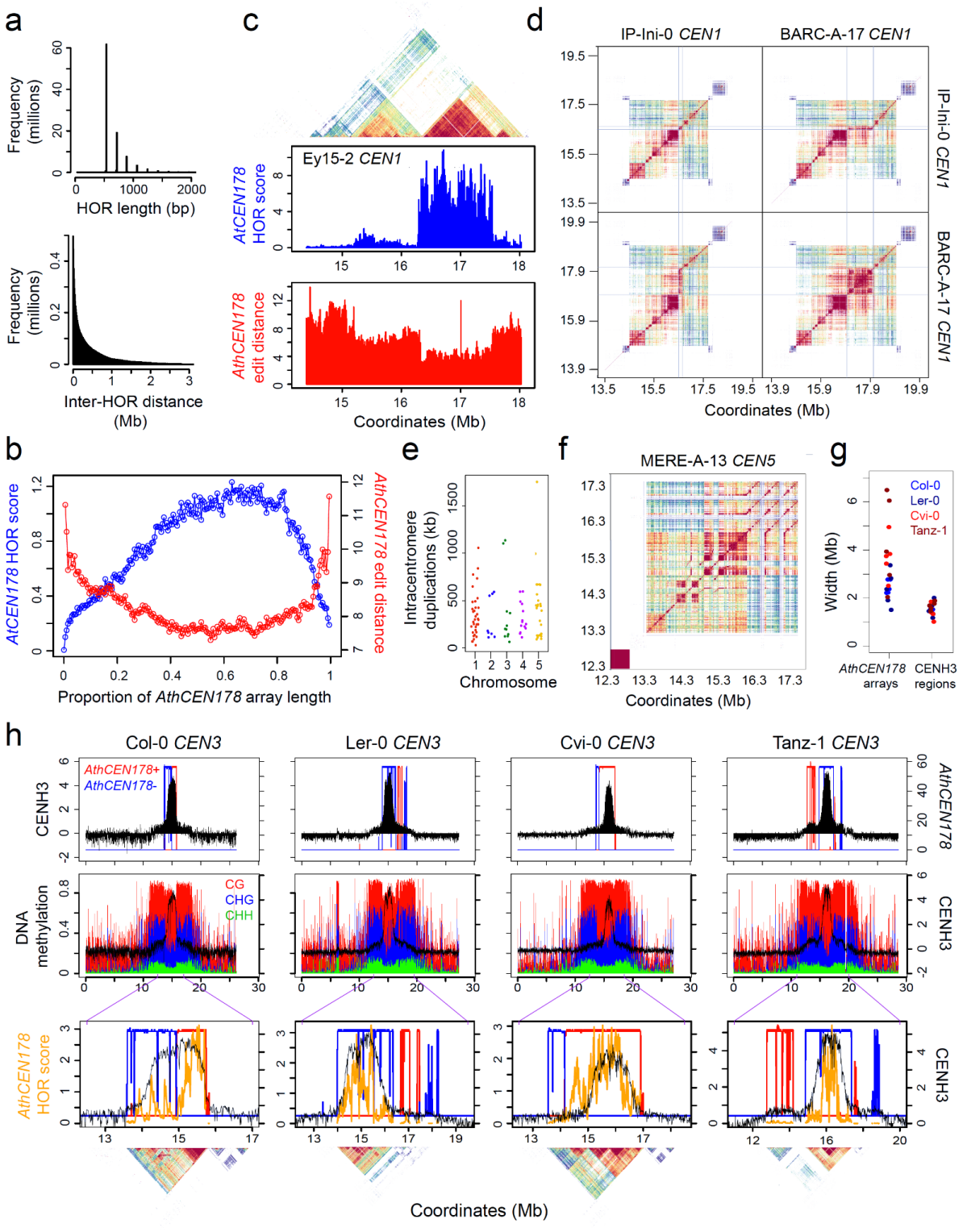

a
Frequency (millions)
HOR length (bp)
Inter-HOR distance (Mb)
c
Ey15-2 CEN1
AtCEN178 HOR score
AthCEN178 edit distance
Coordinates (Mb)
d
IP-Ini-0 CEN1
BARC-A-17 CEN1
Coordinates (Mb)
b
AtCEN178 HOR score
AthCEN178 edit distance
Proportion of AthCEN178 array length
e
Intracentromere duplications (kb)
Chromosome
f
MERE-A-13 CEN5
Coordinates (Mb)
g
Col-0
Ler-0
Cvi-0
Tanz-1
Width (Mb)
AthCEN178 arrays
CENH3 regions
h
Col-0 CEN3
Ler-0 CEN3
Cvi-0 CEN3
Tanz-1 CEN3
AthCEN178+
AthCEN178-
CENH3
DNA methylation
CG
CHG
CHH
AthCEN178 HOR score
AthCEN178
Coordinates (Mb)

Figure 2. Dynamic genetic and epigenetic evolution of the *A. thaliana* centromere satellite arrays. a, Histograms showing the frequency of *AthCEN178* HOR lengths (bp, top), and the frequency of inter-HOR distances (Mb, bottom). **b,** Quantification of *AthCEN178* HOR scores (blue) and edit distances (red) across the scaled length of all satellite arrays. **c,** Comparison of HOR scores (blue) with *AthCEN178* edit distances (red) within Ey15-2 *CEN1*. A StainedGlass sequence identity heat map of the same region is shown above. **d,** StainedGlass sequence identity heat maps comparing within- and between-accession sequence identity for IP-Ini-0 (Spanish) and BARC-A-17 (French) *CEN1*. Grey lines demarcate the ~1 Mb region in BARC-A-17 that is similar to a ~100 kb region in IP-Ini-0. **e,** Intra-centromere duplication sizes (kb) 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. Data points are coloured according to accession (Col-0=blue, Ler-0=dark blue, Cvi-0=red, Tanz-1=dark red). **h,** 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 chromosome 3 of Col-0, Ler-0, Cvi-0 and Tanz-1. Beneath, CENH3 ChIP-seq enrichment (black) is plotted against DNA methylation (%) in CG (red), CHG (blue) and CHH (green) sequence contexts, along the whole chromosome. Beneath are close-ups of the centromere regions with *AthCEN178* density (red, blue), CENH3 ChIP-seq enrichment (black) and *AthCEN178* HOR scores (orange) plotted. A StainedGlass sequence identity heat map is shown at the bottom.

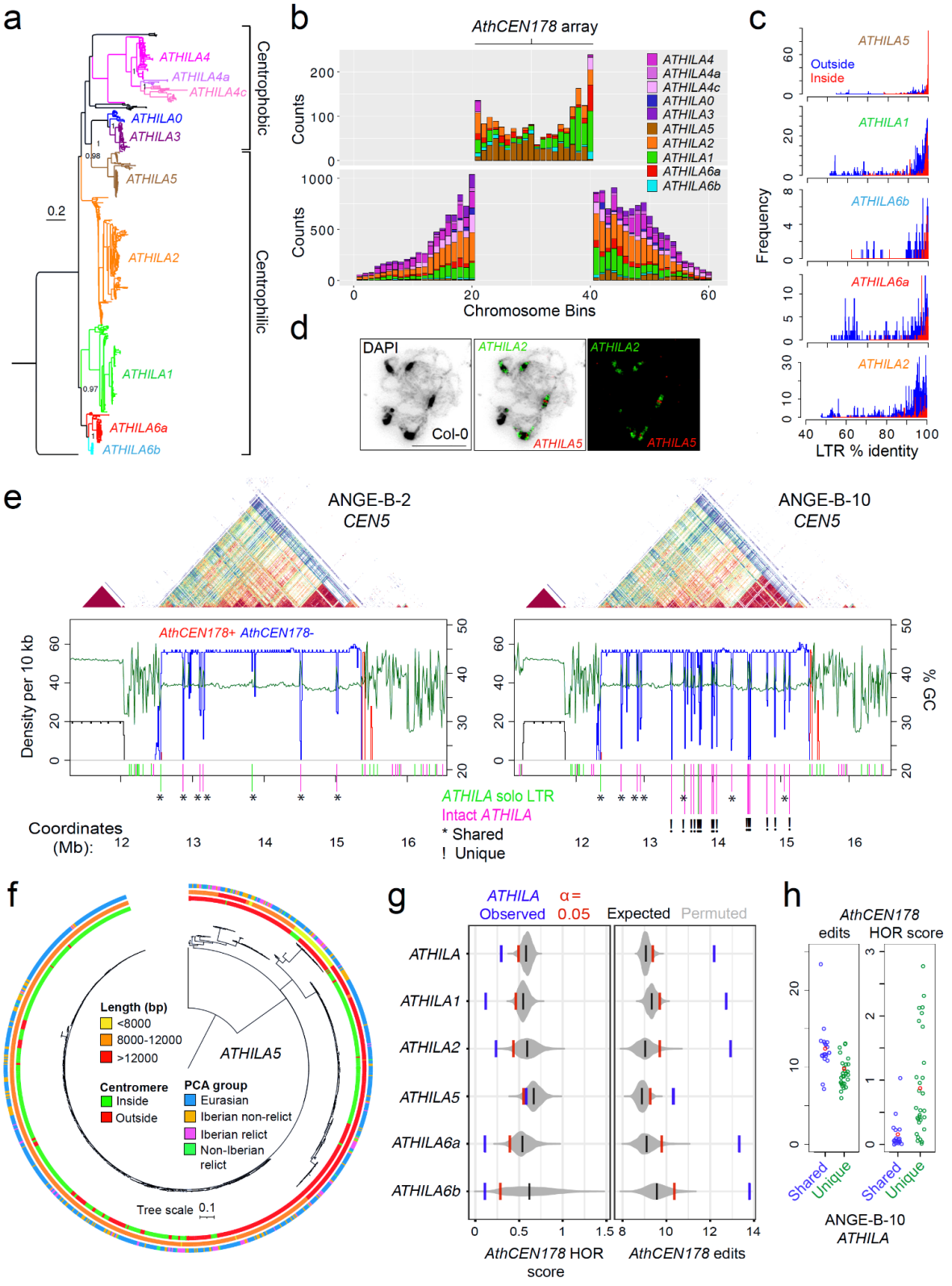

a
ATHILA4
ATHILA4a
ATHILA4c
ATHILA0
ATHILA3
ATHILA5
ATHILA2
ATHILA1
ATHILA6a
ATHILA6b
Centrophobic
Centrophilic
0.2
b
AthCEN178 array
Counts
Chromosome Bins
c
Outside
Inside
Frequency
LTR % identity
d
DAPI
Col-0
e
ANGE-B-2
CEN5
ANGE-B-10
CEN5
AthCEN178+ AthCEN178-
Density per 10 kb
% GC
Coordinates (Mb):
ATHILA solo LTR
Intact ATHILA
* Shared
! Unique
f
Length (bp)
<8000
8000-12000
>12000
Centromere
Inside
Outside
PCA group
Eurasian
Iberian non-relict
Iberian relict
Non-Iberian relict
Tree scale 0.1
g
ATHILA
Observed
α = 0.05
Expected
Permuted
AthCEN178 HOR score
AthCEN178 edits
h
AthCEN178
edits
HOR score
Shared
Unique
ANGE-B-10
ATHILA

Figure 3. Centrophilic invasion of the *A. thaliana* satellite arrays by *ATHILA* retrotransposons. **a,** Phylogeny of full-length *ATHILA* rooted using a maize *Huck-Ty3* element. Centrophobic and centrophilic families are labelled, and bootstrap support is shown. **b,** Count of intact elements and soloLTRs for the indicated *ATHILA* families, either inside (top) or outside (bottom) the *AthCEN178* arrays. *AthCEN178* arrays were split into 20 bins of varying sizes, depending on array length. For upstream and downstream flanking regions, the average bin size (141 kb) across all arrays was used. **c,** % sequence identity between *ATHILA* LTRs for elements inside (red) or outside (blue) the *AthCEN178* arrays, according to family. **d,** Representative FISH micrographs probed for *ATHILA2* (green) and *ATHILA5* (red) on Col-0 pachytene chromosomes. DAPI staining is shown in black. Scale bars=10 μm. **e,** StainedGlass sequence identity heat maps for ANGE-B-2 and ANGE-B-10 *CEN5*, with GC content (%, green) and *AthCEN178* density per 10 kb on forward (red) and reverse (blue) strands plotted beneath. X-axis ticks indicate intact *ATHILA* (pink) and soloLTR (green) insertions. '*' marks shared *ATHILA* between ANGE-B-2 and ANGE-B-10, whereas '!' marks ANGE-B-10-unique insertions. **f,** Phylogeny of 950 *ATHILA5* across all 66 accessions. Shading around the circumference indicates *ATHILA* presence inside versus outside the *AthCEN178* arrays, element length size class (bp) (**Extended Data Fig. 9**), and PCA group membership (**Fig. 1b)**. **g,** Analysis of 1 kb regions flanking *ATHILA* within the satellite arrays across all 66 accessions for *AthCEN178* HOR scores (left) and edit distances (right). Observed mean values (blue) were compared to 1,000 permuted sets of random loci of the same widths (grey) within the *AthCEN178* arrays. The α=0.05 significance thresholds are indicated (red), and the means of the permuted loci (black). Empirical *P*-values from all comparisons were <0.001, except for *ATHILA5* HOR scores (*P*=0.128). **h,** Scatter plots of mean *AthCEN178* edit and HOR scores in 2 kb windows flanking *ATHILA* 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.

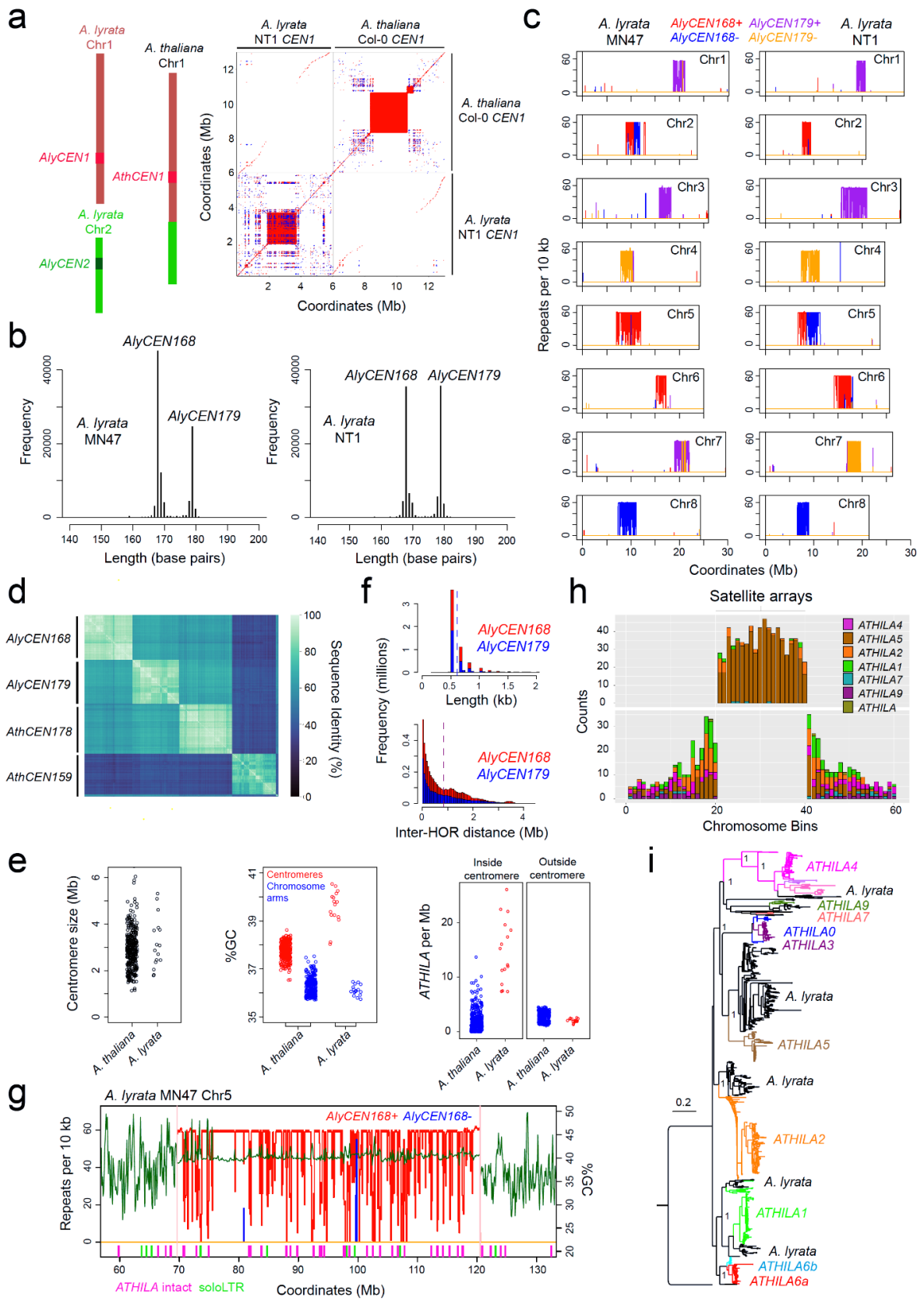


Figure 4. Inter-specific centromere satellite turnover and transposon dynamics between *A. thaliana* and *A. lyrata*. a, Diagram of synteny between *A. thaliana* chromosome 1 and *A. lyrata*

chromosomes 1 and 2 on the left. A sequence identity dot plot comparing *A. thaliana* Col and *A. lyrata* NT1 *CEN1* using 75 bp windows is shown on the right. Red and blue indicate the same, or opposite strand similarity. **b,** Distribution of tandem repeat lengths (bp) in *A. lyrata* MN47 and NT1, with the *AlyCEN168* and *AlyCEN179* families labelled. **c,** Satellite repeat density per 10 kb along *A. lyrata* MN47 and NT1 chromosomes, with the colour code at the top. **d,** Pairwise sequence identity of 100 sampled *AthCEN178* and *AthCEN159* from *A. thaliana* Col-0, and *AlyCEN168* and *AlyCEN179* from *A. lyrata* NT1. **e,** Scatter plots of satellite array sizes (Mb), GC content of the chromosome arms (blue) and satellite arrays (red), and combined intact and soloLTR *ATHILA* insertions per Mb inside and outside the satellite arrays, from 66 *A. thaliana* and two *A. lyrata* accessions. **f,** Satellite HOR lengths (bp) and inter-HOR distances (Mb) within the *AlyCEN168* (red) and *AlyCEN179* (blue) dominated *A. lyrata* centromeres. **g,** *AlyCEN168* density (red) and GC content (%, green) along *CEN5* of *A. lyrata* MN47. Intact *ATHILA* (pink) and soloLTR (green) positions are indicated on the x-axis. **h,** Combined intact *ATHILA* element and soloLTR counts for the indicated families, either inside (top) or outside (bottom) the *A. lyrata* satellite arrays. Centromere satellite arrays were split into 20 bins of varying sizes, depending on centromere length. For the 20 bins upstream and downstream, the average bin size (158 kb) across all centromeres was used. **i,** Phylogeny of full-length *ATHILA* elements identified in *A. thaliana* (coloured) and *A. lyrata* (black), rooted using a maize *Huck Ty3* element, with bootstrap support shown for key nodes.

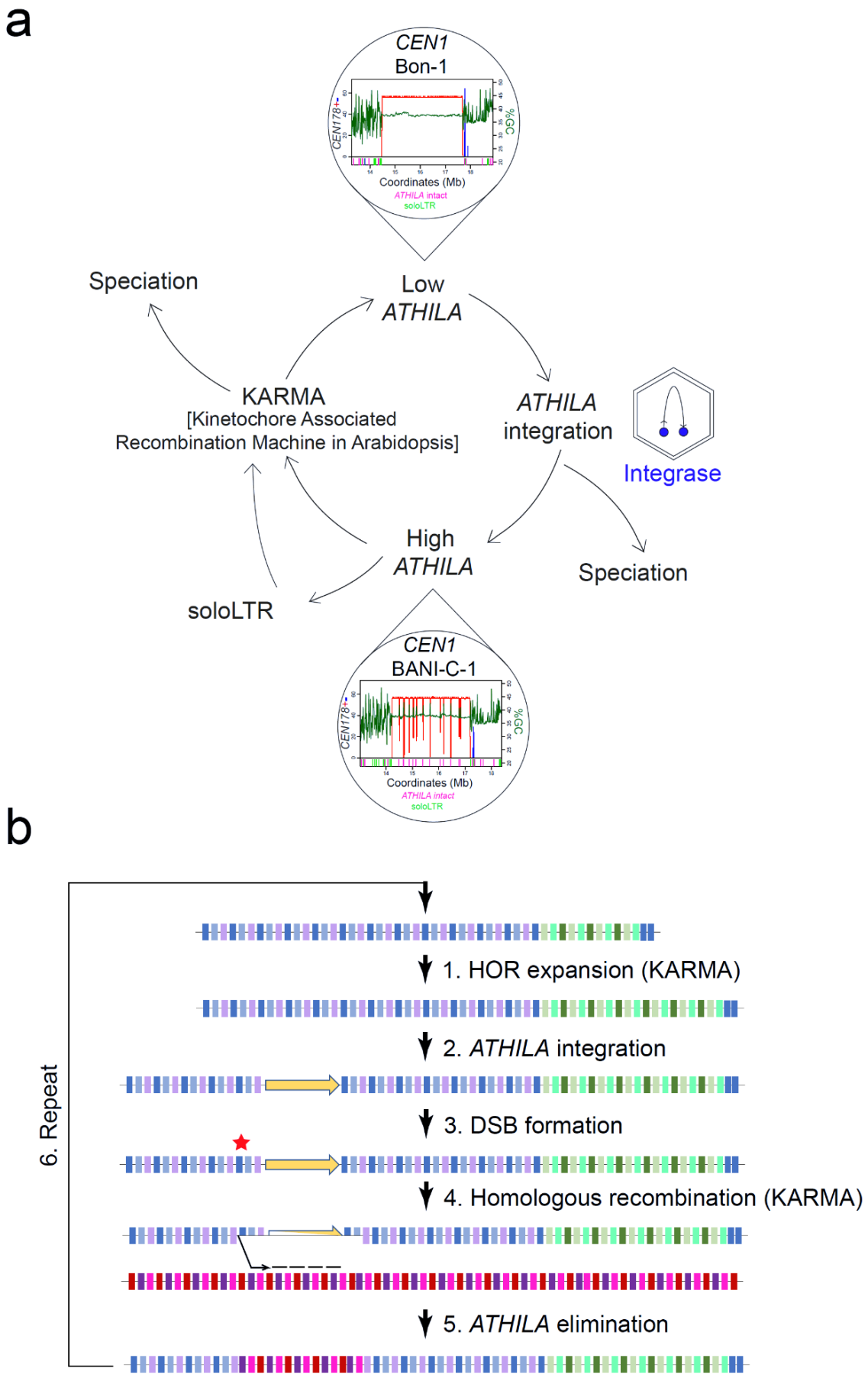

a
CEN1
Bon-1
CEN178
%GC
Coordinates (Mb)
ATHILA intact
soloLTR
Speciation
Low
ATHILA
KARMA
[Kinetochore Associated
Recombination Machine in Arabidopsis]
ATHILA
integration
Integrase
High
ATHILA
Speciation
soloLTR
CEN1
BANI-C-1
b
6. Repeat
1. HOR expansion (KARMA)
2. ATHILA integration
3. DSB formation
4. Homologous recombination (KARMA)
5. ATHILA elimination

Figure 5. Cycles of satellite repeat and centrophilic *ATHILA* retrotransposon evolution in *Arabidopsis* centromeres. a, *ATHILA* diversity in *A. thaliana* centromere satellite arrays suggests cycles between high and low retrotransposon states, represented by Bon-1 *CEN1* (zero) and BANI-C-1 (13 *ATHILA*), which are members of *AthCEN178* similarity group 1. *ATHILA* transcription leads to the formation of virus-like particles (VLPs) where reverse transcription generates new *ATHILA* copies with the potential to integrate into the chromosomes. We propose that the integrase (blue) of centrophilic *ATHILA* is adapted to recognise centromeric DNA or chromatin states, leading to invasion. Intact *ATHILA* may undergo internal homologous recombination to generate soloLTRs. Simultaneously, satellite arrays undergo phases of homologous recombination that can result in intra-centromere duplications, satellite homogenization and purging of *ATHILA*. Recombination within the satellites is hypothesised to be promoted by the kinetochore and associated factors, termed the 'Kinetochore Associated Recombination Machine in Arabidopsis' (KARMA), by analogy with the human KARM hypothesis[5]. The dynamic polymorphism observed in *A. thaliana* and *A. lyrata* satellites and retrotransposons are consistent with a role in driving centromere evolution during speciation. **b,** A representative centromere satellite array is shown, with monomer repeats represented by coloured blocks, and matched colours indicate matched sequence. Through KARMA, satellite HORs can expand (step 1), and arrays can be invaded by *ATHILA* (step 2). The satellite arrays may sustain a DNA double strand break (DSB, red star) adjacent to an *ATHILA*, which could occur during meiosis or mitosis (step 3). The DSB may then enter homologous recombination, potentially mediated via KARMA, involving resection and ssDNA invasion of another chromosome, which could be a sister, or a homolog, where it undergoes template-driven synthesis and repair (step 4). This results in unidirectional transfer of satellite sequence (gene conversion) and elimination of the *ATHILA* sequence (step 5). The modified satellite array may then enter further cycles of HOR expansion/contraction and *ATHILA* invasion (step 6).

Methods

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

Arabidopsis thaliana was grown at 23ºC under 16 hours of light[7]. High-molecular weight (HMW) genomic DNA was extracted with a column-based or β-mercapto-ethanol (BME) method. For the column-based method, 30-day-old rosettes from multiple individuals were pooled, ground in liquid nitrogen, and ~10 grams of powder were 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, 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) of NIB containing 20% Triton-X-100 was added. Samples were incubated on ice for 15 minutes and centrifuged at 3,000 *g* at 4°C for 15 minutes. Pellets were pooled, washed with 35 ml NIB containing 1% Triton-X-100, and pelleted once more at 3,000 *g* at 4°C for 15 minutes. The pellet was gently resuspended in 20 ml of pre-warmed (37°C) G2 lysis buffer (Qiagen, #1014636), incubated with 50 µg/ml RNaseA (Qiagen; #19101) at 37°C for 30 minutes, followed by 200 µg/ml proteinase K treatment (Qiagen; #19133) at 50°C for 3 hours. After centrifugation at 8,000 *g* at 4°C for 15 minutes, 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. 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.

For the BME method[44], 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 minutes at 55°C in freshly prepared and pre-heated lysis buffer (1% sodium metabisulfite, 1% PVP40, 0.5 M NaCl, 100 mM Tris HCl pH8, 50 mM EDTA pH8, 1.5% SDS, 2% β-mercapto-ethanol [BME]). The next steps were performed at room temperature. 60 µl of 20 mg/ml PureLinkTM RNaseA (Thermo Fisher Scientific, #12091021) was added to the lysate and incubated for 10 minutes. 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 incubated for 10 minutes on a rotator. After centrifuging at 4,400 *g* for 10 minutes, the upper phase was transferred to a new tube and mixed with 24:1 (v/v) chloroform:isoamyl alcohol for 10 minutes on a rotator. Following a second centrifugation step at 4,400 *g* for 10 minutes, the upper phase was transferred to a new tube, followed by two bead cleanups. The first cleanup involved incubation for 30-60 minutes with 1×volume of 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 beads were washed twice with 80% ethanol. Elution was performed with 50 µl EB (Qiagen) after incubation at 37°C for 15 minutes. The second cleanup was performed with 0.45×volume of AMPure PB magnetic beads (Pacific Biosciences, P/N 100-265-900). After a binding time of 30 minutes on a rotator, beads were placed on a magnet and washed two times with 80% ethanol. For elution, 45 µl EB was added and incubated for 10-15 minutes on a rotator. 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 SA, #B06010002) with 20 kb or 25 kb settings. Five µg of the sheared fraction was used for library construction with the HiFi SMRTbell Express Template Prep Kit 2.0, size-selected with the BluePippin system (SageScience) using the 0.75% DF Marker S1 3-10 kbp - Improved Recovery cassette definition (BLF7510, Biozym) under high pass elution mode and a second elution for 30 minutes to increase recovery yield. Libraries were sequenced on the Sequel II system (Pacific Biosciences) using Binding Kit 2.0 (P/N 101-842-900). For the gTUBE approach, HMW-DNA (120 ng/µl) was sheared twice (back and forth) with a gTUBE (Covaris, P/N 520079) in a Eppendorf Centrifuge 5424 at 4,800 rpm (soft) for 3×1 minute. Five µg of sheared DNA was used for library construction with the HiFi SMRTbell Express Template Prep Kit 2.0, size-selected with the BluePippin system (SageScience) with 17 kb cutoff in a 0.75% DF Marker S1 High-Pass 6-10 kb vs3

gel cassette (BLF7510, Biozym), and the resulting libraries were sequenced using Binding Binding Kit 2.2 (P/N 101-894-200).

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

For genomic DNA extraction for Oxford Nanopore Technologies (ONT) sequencing, 21 day-old *A. thaliana*, grown on ½ MS media containing 1% sucrose, were placed in the dark for 48 hours prior to harvesting. Approximately 10 grams 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 pestle and mortar, and resuspended in 200 ml MEB. Ground tissue was thawed in MEB with frequent stirring. The homogenate was forced through 4 layers of Miracloth, and then filtered again through 4 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 minutes. The suspension was centrifuged at 800 *g* at 4°C for 20 minutes. The supernatant was removed and the pellet resuspended using a paintbrush in 10 ml 2-methyl-2,4 pentanediol buffer pH 7.0 (MPDB). The suspension was centrifuged at 650 *g* at 4°C for 20 minutes. The supernatant was removed and the pellet was washed with 10 ml of MPDB. Washing and centrifugation was 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 ml CTAB buffer was added to the nuclei pellet and mixed via gentle inversion, followed by incubation at 60ºC until full lysis had occurred, taking between 30 minutes and 2 hours. An equal volume of chloroform was added and incubated on a rocking platform, with a speed of 18 cycles per minute, for 30 minutes, followed by centrifugation at 3000 *g* for 10 minutes. An equal volume of phenol/chloroform/isoamyl alcohol (PCI, 25:24:1) was added to the lysate, followed by incubation on a rocking platform (18 cycles per minute) for 30 minutes. The lysate was centrifuged at 3,000 *g* for 10 minutes and the upper aqueous phase was transferred into a fresh tube. The PCI extraction was then repeated. The extraction was repeated using only chloroform. One-tenth volume of 3M Sodium Acetate was added to the lysate and mixed by gentle inversion. Two volumes of ice cold ethanol were added and mixed by inversion. DNA was precipitated at -20ºC for 48 hours. The

precipitated DNA was removed using a glass hook and washed three times in 70% ethanol. The DNA was dissolved in 120 µl of 10 mM Tris-HCl pH 8.5.

Approximately 5 µg of DNA was size selected (>30 kb) using the BluePippin System (Sage Science) and the 0.75% agarose gel cassette (BLF7510, Biozym), using Range mode and BP start set at 30 kb. Library preparation followed 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 NEBNext Ultra II End Prep Enzyme Mix (New England Biolab, E7180S), with incubation for 30 minutes at 20°C, followed by 30 minutes at 65°C. The sample was cleaned using 1×volume AMPure XP beads and eluted in 61 µl of nuclease-free water. Adapters were ligated at room temperature using 25 µl Ligation Buffer, 10 µl NEBNext T4 DNA Ligase and 5 µl Adapter Mix for 2 hours. The library was cleaned with 0.4×volume AMPure XP beads, washed using ONT Long Fragment buffer and eluted in 15 µl elution buffer.

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

Arabidopsis lyrata was grown in the greenhouse at 21°C under 16 hours of light[26]. HMW DNA was isolated from 1.5 g of plant material using a NucleoBond HMW DNA kit[26]. Quality was assessed using a FEMTOpulse device (Agilent) and quantity 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 g-TUBEs (Covaris), and final library size selection 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 hours (Pacific Biosciences).

CENH3 chromatin immunoprecipitation and sequencing (ChIP-seq)

Approximately 12 grams 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 crosslinked in 1% formaldehyde at room temperature for 25 minutes. The crosslinking reaction was quenched with 125 mM glycine and incubated at room temperature for a further 25 minutes. The nuclei were purified from cellular debris via two rounds of filtration through one layer of Miracloth and centrifuged at 2,500 *g* at 4°C for 25 minutes. 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* at 4°C for 10 minutes. 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 Evolution device with the following settings: power=150 V, bursts per cycle=200, duty factor=20% and time=90 seconds. 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 α-CENH3[2] (gift of Prof. Steven Henikoff). 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 minutes. Samples were reverse crosslinked by incubating with 0.24 M NaCl at 65°C overnight. Proteins and RNA were digested with Proteinase K treatment, and RNase A, and DNA was purified 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.

Fluorescence *in situ* hybridization (FISH)

For fluorescence *in situ* hybridization (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′ and ATH5-R3 5′-TGCATCTTATGTCATTGCAGCA-3′) were designed against the consensus sequence of *ATHILA*5 LTRs, using Primer3 (version 2.3.7) implemented in Geneious Prime 2022.2. PCR amplification was performed as follows: initial denaturation at 95°C for 5 minutes; 35 cycles of denaturation at 95°C for 20 seconds, annealing at 58°C for 20 seconds and extension at 72°C for 1 minute; and a final extension at 72°C for 5 minutes. PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel; #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/T7 primers and six clones with the expected length of amplicon were Sanger sequenced (Macrogen). Sequences of clones were aligned to the consensus sequence of *ATHILA5* LTRs using MAFFT (version 7.490)[45], 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 minutes; 35 cycles of denaturation at 95°C for 20 seconds, annealing at 46°C for 20 seconds and extension at 72°C for 20 seconds; and a final extension at 72°C for 5 minutes. PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). 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 minutes and incubated at 37°C for 12 hours. The hapten-labelled probes were immuno-detected, 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 the Isis software (MetaSystems) for all four fluorochromes

using appropriate excitation and emission filters (AHF Analysentechnik). The four monochromatic images were pseudocoloured, merged, and cropped using Photoshop CS (Adobe Systems). 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

From the 66 assemblies analysed, 63 were generated for this study, and the other three have been published[7,14,15]. q20 High Fidelity (HiFi) reads were generated with the PacBio Circular Consensus Sequencing tool ccs (version 6.0.0) (--all), followed by extracthifi (version 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 (version 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* assemblies were performed with hifiasm (version 0.16.1-r375) (-l0 -f0)[46], and contigs larger than 100 kb were scaffolded into chromosomes with RagTag (version 2.0.1) (scaffold -q 60 -f 30000 -i 0.5 --remove-small)[47], 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 (version 2.0.1) (patch --join-only -f 10000 --remove-small)[47]. For 14 accessions, *de novo* assemblies were performed with hifiasm v0.15.4-r343 (-l0 -f0)[46], and contigs larger than 100 kb were scaffolded into chromosomes with RagTag v2.0.1 (scaffold -q 60 -f 30000 -i 0.5 --remove-small)[47], using as reference genome a Bionano optical map of IP-Ini-0. The assemblies of accessions Col-0[7], Ey15-0[7] and Kew-1[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 the 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 RagTag agp2fa function[47].

Quality assessment of *A. thaliana* HiFi assemblies

HiFi reads were aligned to their corresponding chromosomes with pbmm2 (version 1.9.0) (align --sort --log-level DEBUG --preset SUBREAD --min-length 5000), including the mitochondrion and chloroplast genomes from TAIR10, to reduce the amount of organellar reads aligning to the nuclear genome. Unmapped reads, reads aligning to the organelle chromosomes, as well as secondary and supplementary alignments were filtered out with samtools (version 1.9) (view -b -F 2308 <input.bam> Chr1 Chr2 Chr3 Chr4 Chr5)[48]. 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 (version 0.1.0)[49]. Similarly, the same bam file served to determine the coverage of the first and second most common bases in the aligned PacBio HiFi reads genome-wide with NucFreq (version 0.1) (NucPlot.py --minobed 2)[49], while a modified version of the script 'hetDetection.R'[50] was used to identify and quantify heterozygosity stretches within the satellite arrays, based on the bed file emitted by NucFreq. Briefly, a heterozygosity stretch was defined as a region where the second most common base is 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 Fig. 1** and **Extended Data Fig. 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 three standard deviations above the genome-wide mean[49]. For the most part, this increase in coverage is uniform over all centromere regions and, except for localised regions (within *CEN3* in 11C1, *CEN1* and *CEN3* in AUZE-A-5, and *CEN2* and *CEN5* in IP-Ini-0), it is 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 from DNA extracted from single individuals, contained several Mb-long regions of heterozygosity within their genomes, including *CEN4* in Lor-16 and *CEN5* in Met-6. A detailed summary of collapsed, expanded and heterozygous regions within the centromeric satellite arrays per accession is included in **Supplementary Table 1**.

Genome assembly from *A. lyrata* HiFi reads

PacBio reads were assembled using Hifiasm in default mode[26,46], choosing the primary contig graph as the assembly. Assembly completeness was assessed using BUSCO and the 'Brassicales_odb10' set[26].

Single nucleotide polymorphism analysis and PCA

Genome assemblies of each *A. thaliana* accession were aligned to the HiFi assembly of Col-0 using minimap2 (version 2.24)[51], using the parameters 'minimap2 -a -x asm5 --cs -r2k -t 16'. SNPs and short indels were called from the alignment using bcftools (version 1.15.1)[52]. 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 being strictly biallelic, which resulted in 2,011,142 SNPs. PCA was performed using the R package ggbiplot.

Identification of centromere-spanning haplotypes 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 (version 1.9.0) (align --sort --preset CCS) (https://github.com/PacificBiosciences/pbmm2), which is a wrapper of minimap2[51]. DeepVariant (version 1.3.0) (--model_type=PACBIO)[53] was used to create individual genome call sets, and GLnexus (version 1.4.1)[54] 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 (version 4.1.3.0) (VariantFiltration --exclude-intervals <repeats.bed>)[55]. In addition, we performed variant hard-filtering at the population level with bcftools (version 1.15.1) (view -e 'NS < 66 || F_MISSING > 0 || AC = 1 || COV < 2604 || COV > 6685 || AC_Het >= 2')[52], and variant hard-filtering at the population level, for biallelic variants with no missing data, and we removed sites that were either 0.5x higher or lower than the median coverage,

using bcftools (version 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 ~800 kb flanking regions surrounding the *AthCEN178* centromere arrays. Polymorphisms were filtered for no missing data and to have a minor allele count of 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 one to five, respectively. We phased the polymorphisms using the Beagle phasing algorithm (version 4.0)[56], and we assigned the minor allele as 1 and the major allele as 0. To examine the haplotype structure, we clustered accessions based on the Hamming distance of the polymorphisms flanking the centromeres in Col-0, using the R package cultevo. We generated heatmaps to represent the haplotype structure using the R package pheatmap. Samples were coloured based on 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 *Arabidopsis thaliana AthCEN178* and *AthCEN159*, and *Arabidopsis 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 based on 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 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, using MAFFT (--retree 2) alignment[45]. 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. Since 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 readjusts the shift by calculating the lowest possible hash score based on each shift iteration, so that closely related sequences have 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 reported.

An *AthCEN178* consensus sequence for each accession was calculated based on MAFFT alignment of all repeats (--retree 2), and the most common nucleotides per position, with a frequency of at least 50%, were extracted. To generate a global consensus sequence, frequency tables from each accession were aligned, averaged and a consensus sequence 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[57].

To calculate similarity between *AthCEN178* sequences of a pair of chromosomes, the number of identical *AthCEN178* instances found within the other chromosome, without excluding duplicated repeats, was calculated reciprocally. These values were normalised by the number of *AthCEN178* repeats in both sets, so that values are in range of 0-100%, where 0 means no shared repeats and 100% means the chromosomes share the same repeat library. Heat maps were constructed by ordering the chromosome pairs using hierarchical cluster analysis, using the R function hcluster from the package amap. Centromere similarity groups were visually inferred from the heat maps.

Satellite higher-order repeats (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 lower than threshold number of disagreements (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 is divided by the number of

AthCEN178 on that chromosome, and expressed as a percentage, which we term the HOR score. To identify large (kb - Mb) 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 5 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 up to at least 50 kb. Duplication size was estimated based on the first and last HORs in the involved regions.

Maximum-likelihood satellite phylogenies were reconstructed using IQ-TREE with 1,000 bootstraps[58]. 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 *CEN178* 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* from two *A. lyrata* accessions (NT1 and MN47) and 60 *AthCEN178* and *AthCEN159* from 6 *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 aligned using MAFFT (version 7.490), using default settings. All trees were rooted with consensus *Capsella rubella* 160 bp and 175 bp centromeric repeats[43], apart from **Extended Data Fig. 10b-c**, which was rooted with 30 *AthCEN178* from the Col-0 *A. thaliana* accession. For **Fig. 4d**, satellites from accessions were aligned and their percentage identity scored with Clustal Omega, then visualised using a heat map. 100 *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 *Arabidopsis 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 Chr1, Chr2, Chr3, Chr4 and Chr5, 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 per chromosome. Saturation analysis of the unique *AthCEN178* sequences was calculated using 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[59]. 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 (TE) in the TAIR10 TE database[60], and the *A. thaliana* TE database developed and maintained by Hadi Quesneville (https://urgi.versailles.inra.fr/Data/Transposable-elements/Arabidopsis). Using this approach, we designed three seeds located on 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 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 2-11 kb fragments represent the putative internal domain of intact *ATHILA* elements. To examine if 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 are 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; the 20-mer upstream of 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 two *A. lyrata* accessions; however, additional high-quality intact *ATHILA* 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 undergo 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* in *A. thaliana* and 109 in *A. lyrata*. We also identified soloLTRs in the *A. thaliana* and *A. lyrata* genomes. We first run 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 target site duplications (TSDs) by allowing one mismatch[45]. 6,546 intact *ATHILA* and 8,776 soloLTRs contained a TSD in *A. thaliana*, and 806 and 333 in *A. lyrata*. We identified open reading frames (minimum 300 bp) in the internal domain of the intact elements using getorf in EMBOSS [61], and the core domains of GAG and POL by running HMMER (version 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 of 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 were reduced using CD-HIT[62]. We then run 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[63]. Phylogenetic trees were visualised and annotated using FigTree (http://tree.bio.ed.ac.uk/software/figtree/), and iTOL[64].

To map the insertion sites of *ATHILA* elements, 20 bp regions flanking each element, including target site duplications (TSDs), were extracted and locally aligned to the *AthCEN178* consensus using the pairwiseAlignment function from the R package Biostrings. Centromeric *ATHILA* were filtered based on this score and only those with both flanking sequences having positive alignment scores were kept (1,367 out of 1,913). 88.4% of the *ATHILA* had their insertion site overlapping by 5 bp, which is expected due to 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 was plotted.

We matched and analysed syntenic *ATHILA* insertions, both intact elements and soloLTRs, in closely related pairs of accessions: BARC-A12 & BARC-A-17, BELC-C-10 & BELC-C-12, CAMA-C-2 & CAMA-C-9, SALE-A-10 & SALE-A-17 and ANGE-B-2 & 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 of this pipeline are shown in **Supplementary Table 6**, where *ATHILA* were 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**), and '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 the 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 contained a large number of mutations, identification of the correct sequence is very challenging, because the gag-pol ORF was 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-9c**). Using this subset of elements, we were able to build the consensus sequence of integrase for these families. We then constructed a Hidden Markov Model profile based on 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 6 families (*ATHILA6a*, *ATHILA3*, *ATHILA4*, *ATHILA4a*, *ATHILA8a*, and *ATHILA7a*). This was expected based on 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 the identification of intact elements of non-*ATHILA* transposons, we ran EDTA with default parameters[65].

Phylogenetic analysis of CENH3

Protein coding genes in all assemblies were annotated with Liftoff (version 1.6.2) (-copies -sc 0.98 -polish -cds)[66], 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 (version 0.12.7)[67], and a multiple sequence alignment was performed with MAFFT (version 7.407) (--reorder --maxiterate 1000 --retree 1)[45]. Subsequently, the fasta alignment was transformed to Nexus format with PGDspider (version 2.1.1.5)[68], and a maximum likelihood phylogenetic tree was inferred with RAXML-NG (version 0.9.0) (raxml-ng --all --model GTR+G --seed 9 --bs-trees 1000 --bs-metric fbp,tbe)[69]. Tree visualisation was performed using the R-package ggtree (version 1.16.6)[70], while the R package treeio (version 1.8.2) was used to add the multiple sequence alignment as a tree object into R[71].

Mapping DNA methylation using ONT reads

We quantified CG, CHG and CHH context DNA methylation with DeepSignal-plant (version 0.1.4)[72], which uses a deep-learning method based on a bidirectional recurrent neural network (BRNN) with long short-term memory (LSTM) units to detect 5mC methylation[72]. R9 reads were filtered for length and accuracy using Filtlong (version 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 (version 1.5.1). Methylation prediction for the CG, CHG, and CHH contexts were called using DeepSignal-plant with the 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[72] 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 (version 1.18) to remove adapter sequences and low-quality bases (Phred+33-scaled quality <20)[73]. 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[74]. 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)[48]. 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)[48,75]. 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)[48]. 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 $\log_2$(ChIP/Input), and custom R scripts were used to plot the average profiles across *AtCEN178* 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) across all 66 accessions, using R (version 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 per 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 empirical *P* value. *AthCEN178* metric values for random loci were also used to calculate the significance threshold (α=0.05), and an expected value (the mean value from 1,000 sets).

Code availability statement

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 are available at https://github.com/vlothec/pancentromere.

Data availability statement: *Arabidopsis thaliana* and *Arabidopsis 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: (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) (as part of a separate study by G.S. and D.W.), (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 PRJEB51511[14,15], 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-10272[6], (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)[26].

Methods references

44. 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).
45. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
46. 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).
47. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
49. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
50. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).
51. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
52. 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).
53. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
54. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btaa1081.
55. 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).

56. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

57. Loo, M. J. vander, Mark, J., P., van & Loo, D. The stringdist Package for Approximate String Matching. *The R Journal* **6,** 111 (2014).

58. 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).

59. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: Interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* (2022) doi:10.1093/bioinformatics/btac018.

60. 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).

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

62. 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).

63. 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).

64. 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).

65. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

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

67. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res.* **9**, (2020).

68. Lischer, H. E. L. & Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 (2012).

69. 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).

70. 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).

71. 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).

72. Ni, P. *et al.* Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. *Nat. Commun.* **12**, 5976 (2021).

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

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

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

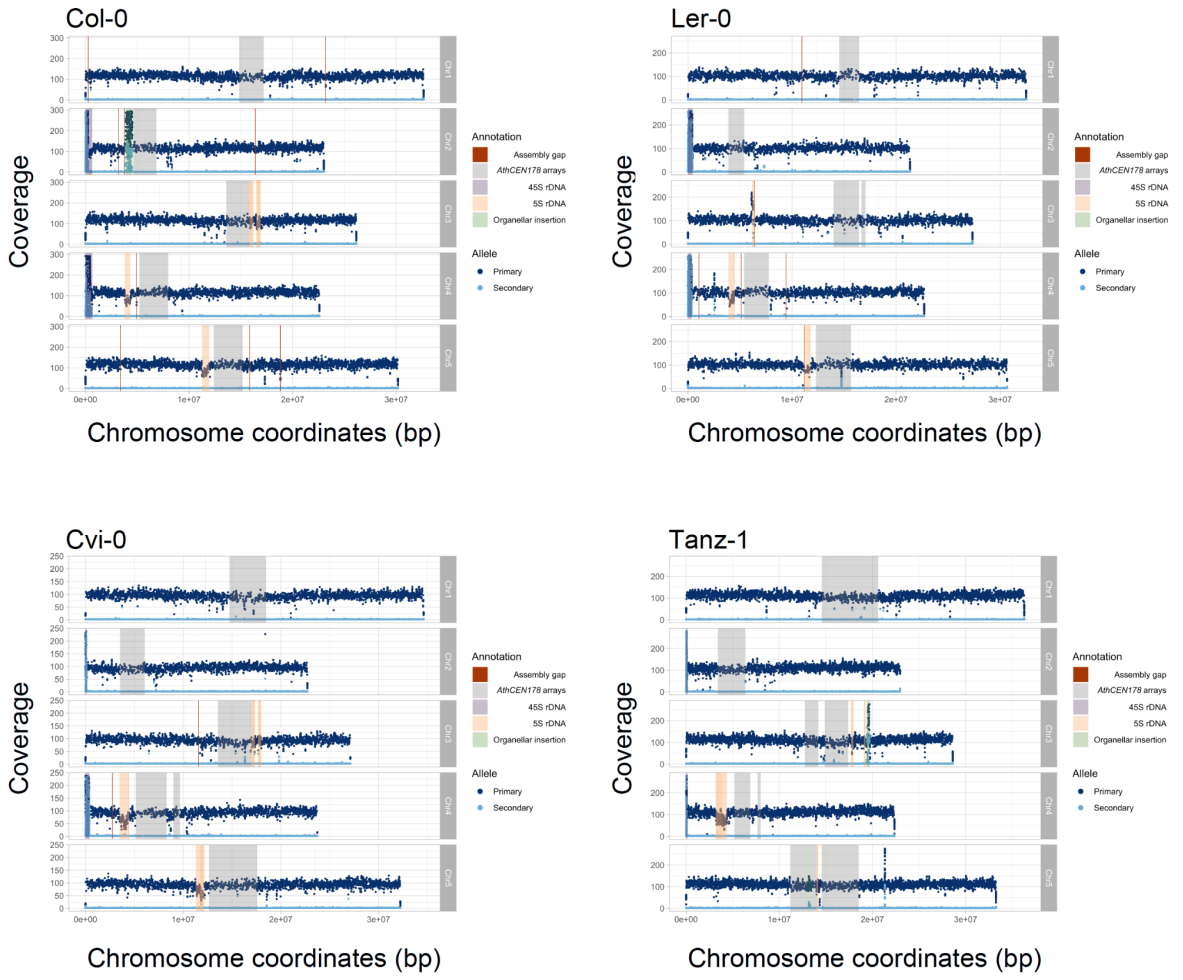
76. Weng, M.-L. *et al.* Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in Arabidopsis thaliana. *Genetics* **211**, 703–714 (2019).

Acknowledgements: We thank Steven Henikoff and Paul Talbert for kindly providing anti-CENH3 antibodies. We thank Richard Durbin, Chenxi Zhou and the Darwin Tree of Life Project for the *A. thaliana* ddAraThal4 Kew-1 assembly.

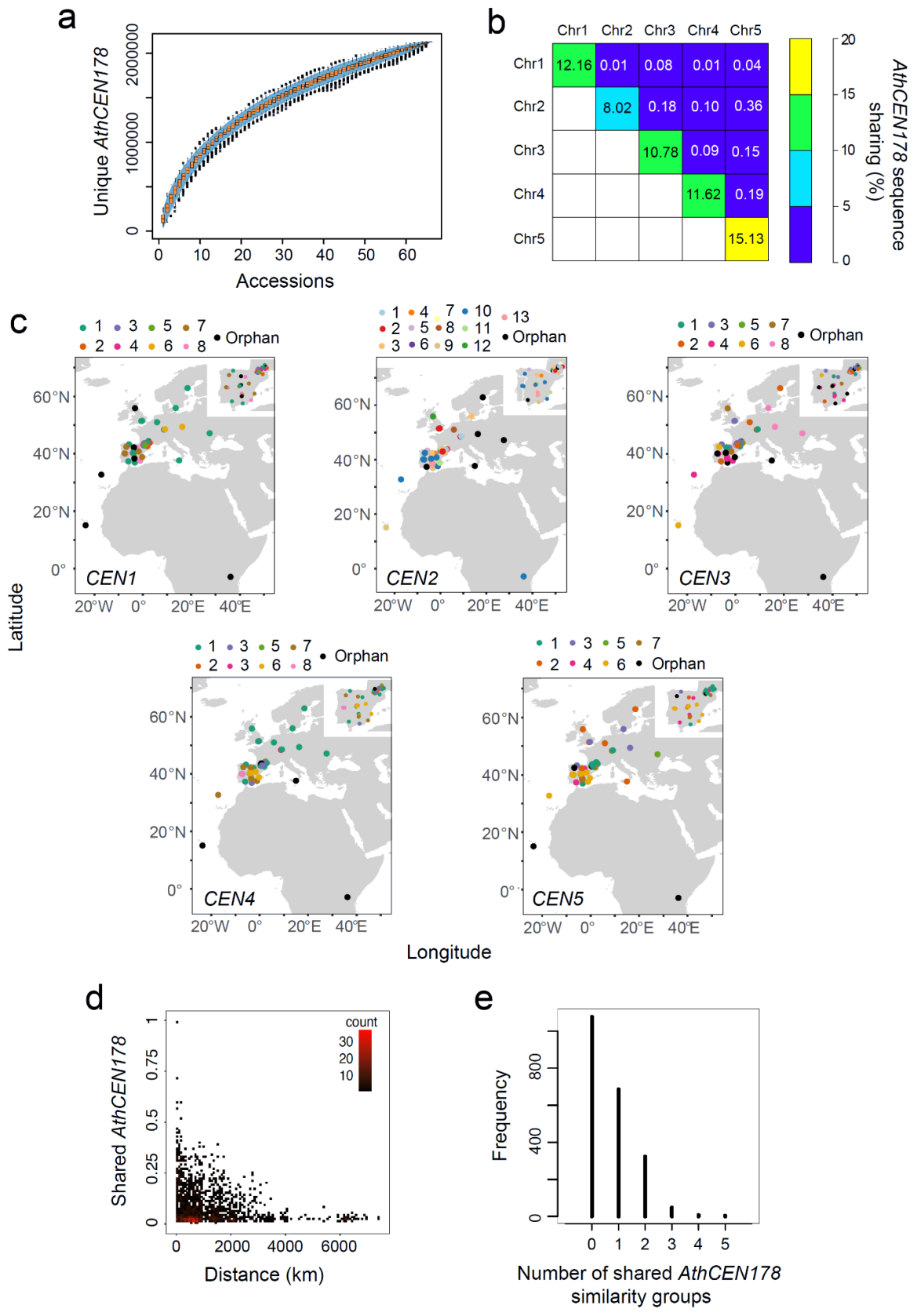
Author Contributions: PW, FAR, MNa, FR, CAB, MAL, PN, AB, DW and IH designed the study. FAR, MN, AS, TM, NG, DH, KF, AH, CL, GS, TS, MC, MM and PN performed the experiments. PW, FAR, RB, MNa, EP, AS, TM, NG, AT, MNo, MAL, DH, PN, AB, DW and IH analysed the results. PW, FAR, RB, PN, AB, DW and IH wrote the paper, with input from all authors.

Competing Interests Declaration: DW holds equity in Computomics, which advises plant breeders. DW consults for KWS SE, a plant breeder and seed producer. All other authors have no competing interests.

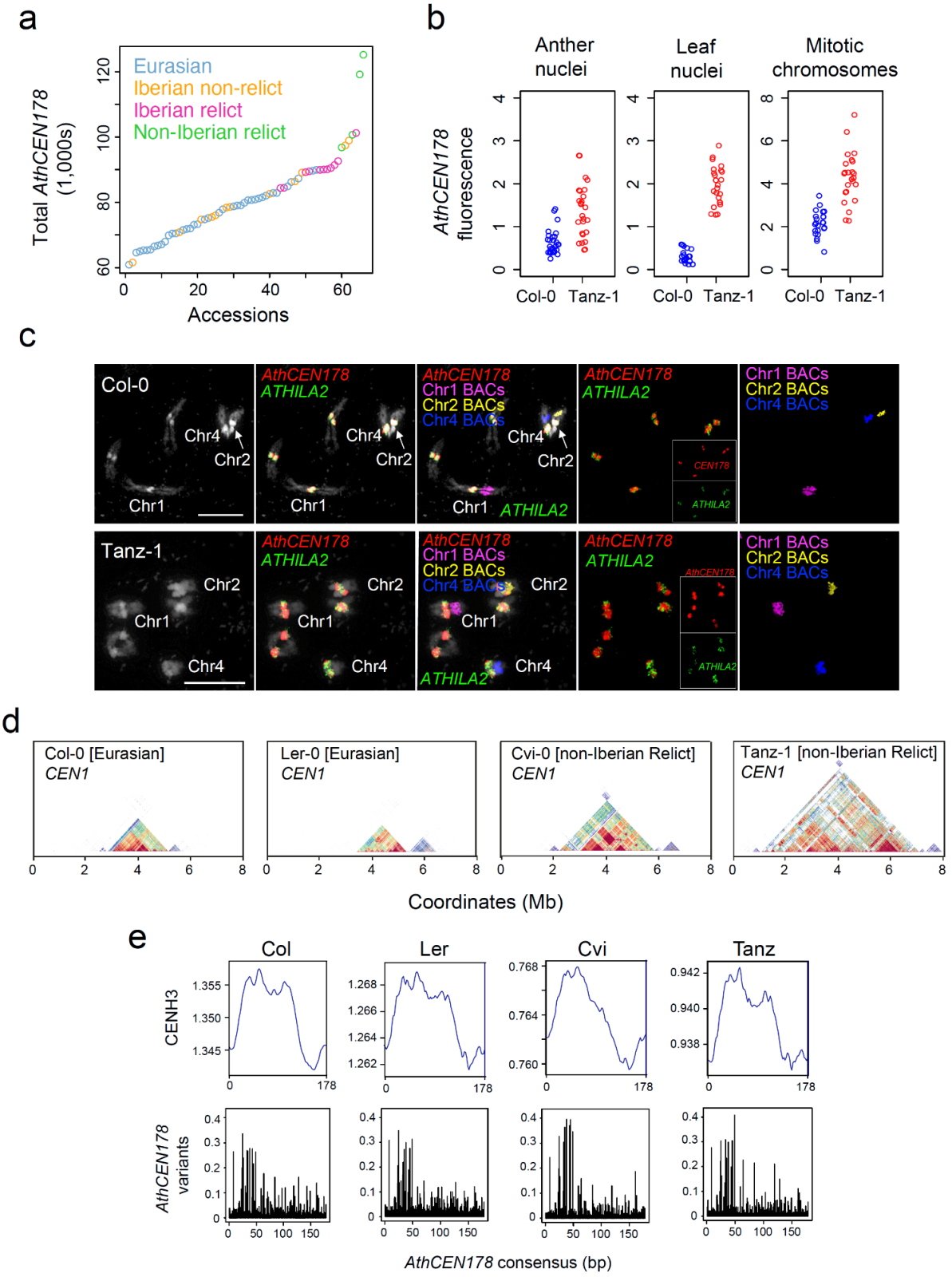
Funding: This work was supported by BBSRC grants BB/S006842/1, BB/S020012/1 and BB/V003984/1, European Research Council Consolidator Award ERC-2015-CoG-681987, Marie Curie International Training Network 'MEICOM' and Human Frontier Science Program award RGP0025/2021 to IRH; EMBO long term postdoctoral fellowship ALTF224-2022 to RB; Human Frontiers Science Program (HFSP) Long-Term Fellowship (LT000819/2018-L) to FAR; Max Planck Society to DW; ERA-CAPS 1001G+ grant to MNo and DW; Royal Society awards UF160222, URF\R\221024, RGF/R1/180006 and RGF/EA/201030 to AB; European Research Council Award ERC HOW2DOBLE 101041354 to PYN; Czech Science Foundation grant number 21-03909S to ML; a BBSRC DTP Studentship to NG; a Broodbank Fellowship to MN; and grant PID2022-136893NB-I00 from the Ministerio de Ciencia e Innovación of Spain/Agencia Estatal de Investigación/10.13039/501100011033/FEDER, EU, to CAB.



Extended Data Figure 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.

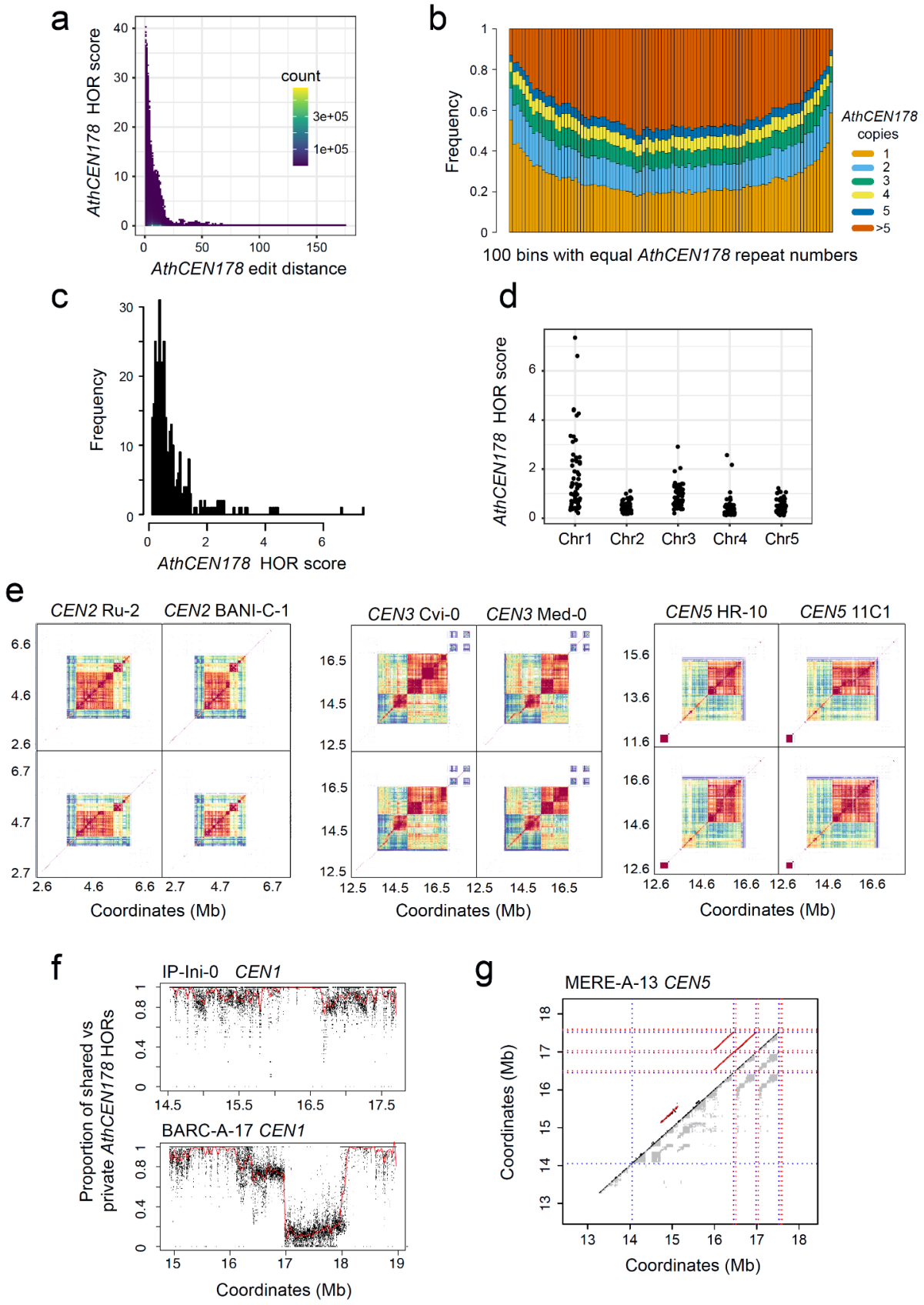

a
Unique AthCEN178
Accessions
b
Chr1 Chr2 Chr3 Chr4 Chr5
12.16 0.01 0.08 0.01 0.04
8.02 0.18 0.10 0.36
10.78 0.09 0.15
11.62 0.19
15.13
AthCEN178 sequence sharing (%)
c
Orphan
CEN1
CEN2
CEN3
CEN4
CEN5
Latitude
Longitude
d
Shared AthCEN178
Distance (km)
count
e
Frequency
Number of shared AthCEN178 similarity groups

Extended Data Figure 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. 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.



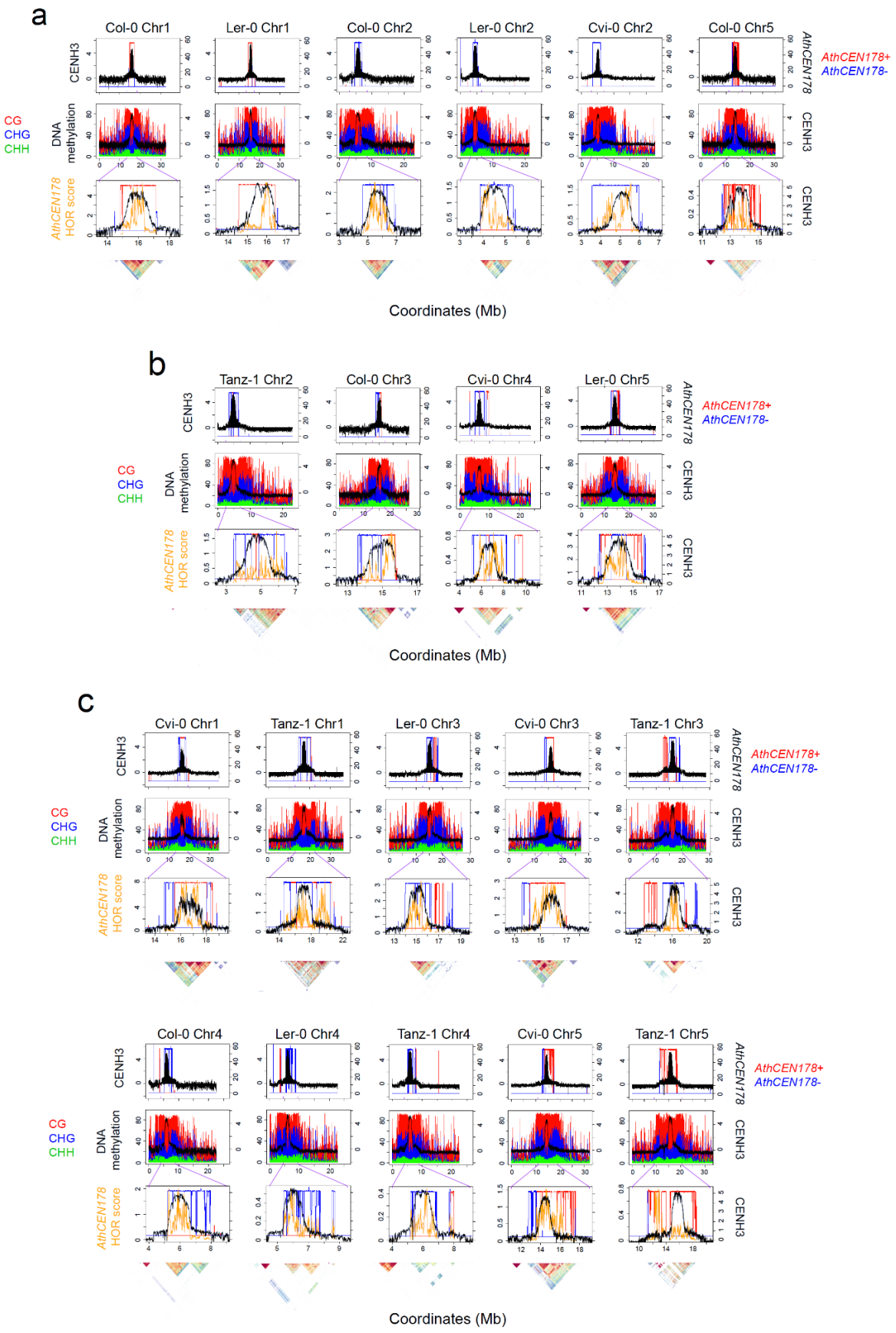
Extended Data Figure 3. Variation in *AthCEN178* copy number and CENH3 occupancy between *A. thaliana* genetic lineages and accessions. a, Total *AthCEN178* copies per accession, coloured according to Eurasian (blue), Iberian non-relict (orange), non-Iberian relict (green) and Iberian relict

(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). **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.

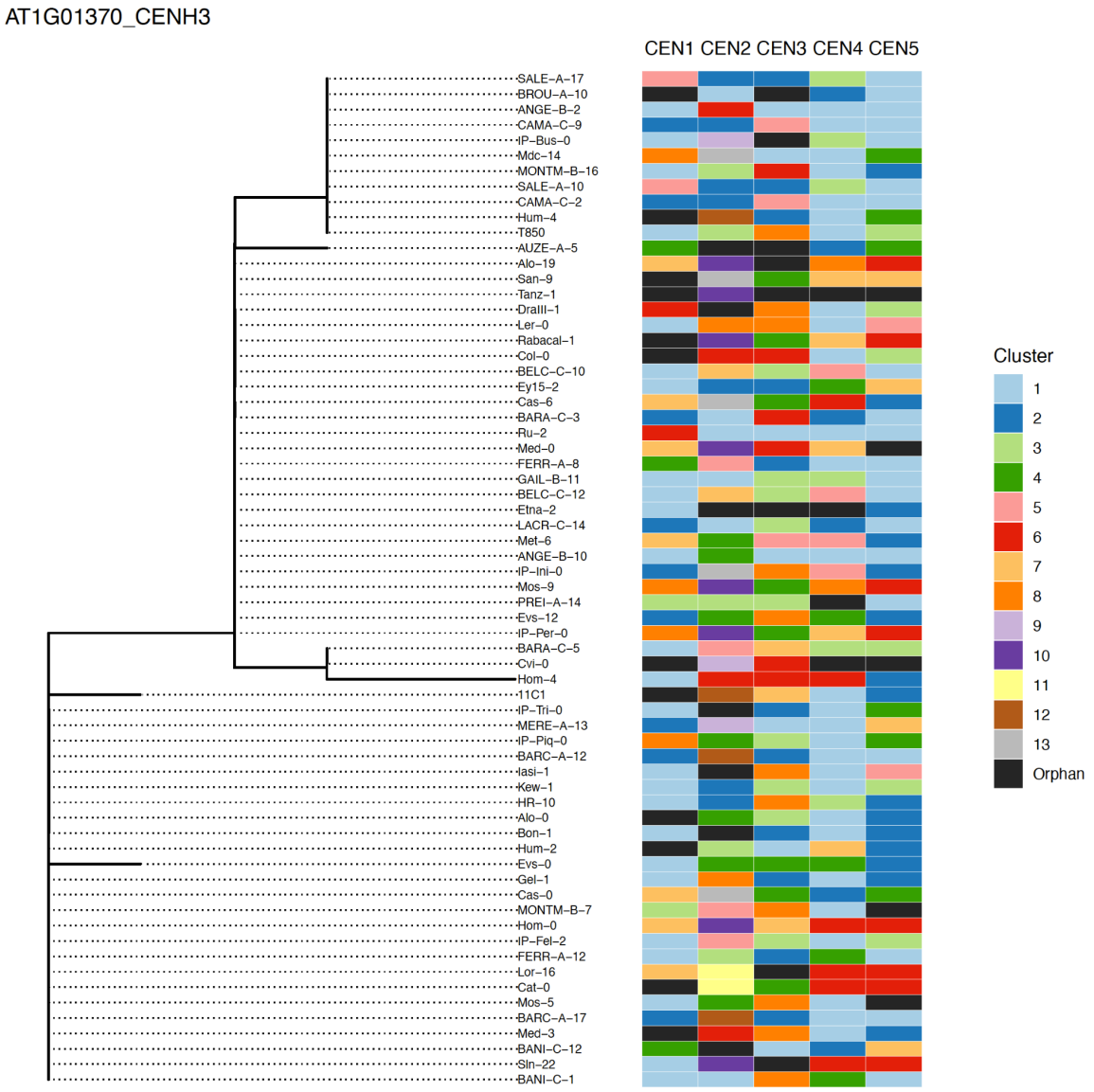


Extended Data Figure 4. *AthCEN178* HORs and dynamic centromere evolution in *A. thaliana*. a, Density plot of *AthCEN178* HOR scores versus edit distances from the chromosome consensus, across

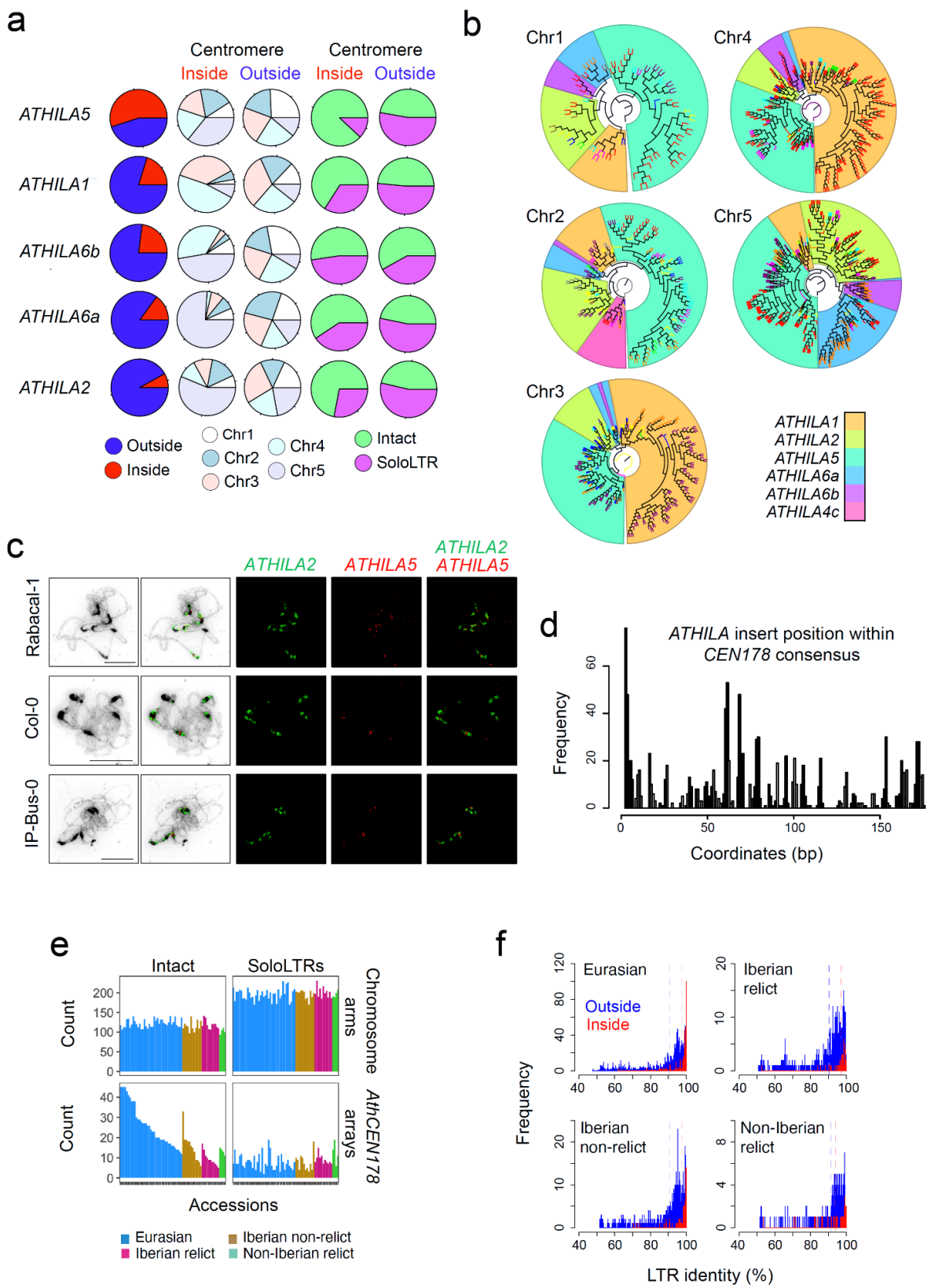
all accessions. **b,** The copy number of each *AthCEN178* repeat was calculated within each chromosome individually. For each chromosome, all *AthCEN178* repeats were divided into 100 bins with an equal number of repeats in each bin. The counts of *AthCEN178* with copy numbers of 1, 2, 3, 4, 5 and >5 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*.


a
Col-0 Chr1
Ler-0 Chr1
Col-0 Chr2
Ler-0 Chr2
Cvi-0 Chr2
Col-0 Chr5
CENH3
DNA methylation
AthCEN178 HOR score
CG
CHG
CHH
AthCEN178+
AthCEN178-
Coordinates (Mb)
b
Tanz-1 Chr2
Col-0 Chr3
Cvi-0 Chr4
Ler-0 Chr5
Coordinates (Mb)
c
Cvi-0 Chr1
Tanz-1 Chr1
Ler-0 Chr3
Cvi-0 Chr3
Tanz-1 Chr3
Col-0 Chr4
Ler-0 Chr4
Tanz-1 Chr4
Cvi-0 Chr5
Tanz-1 Chr5
Coordinates (Mb)

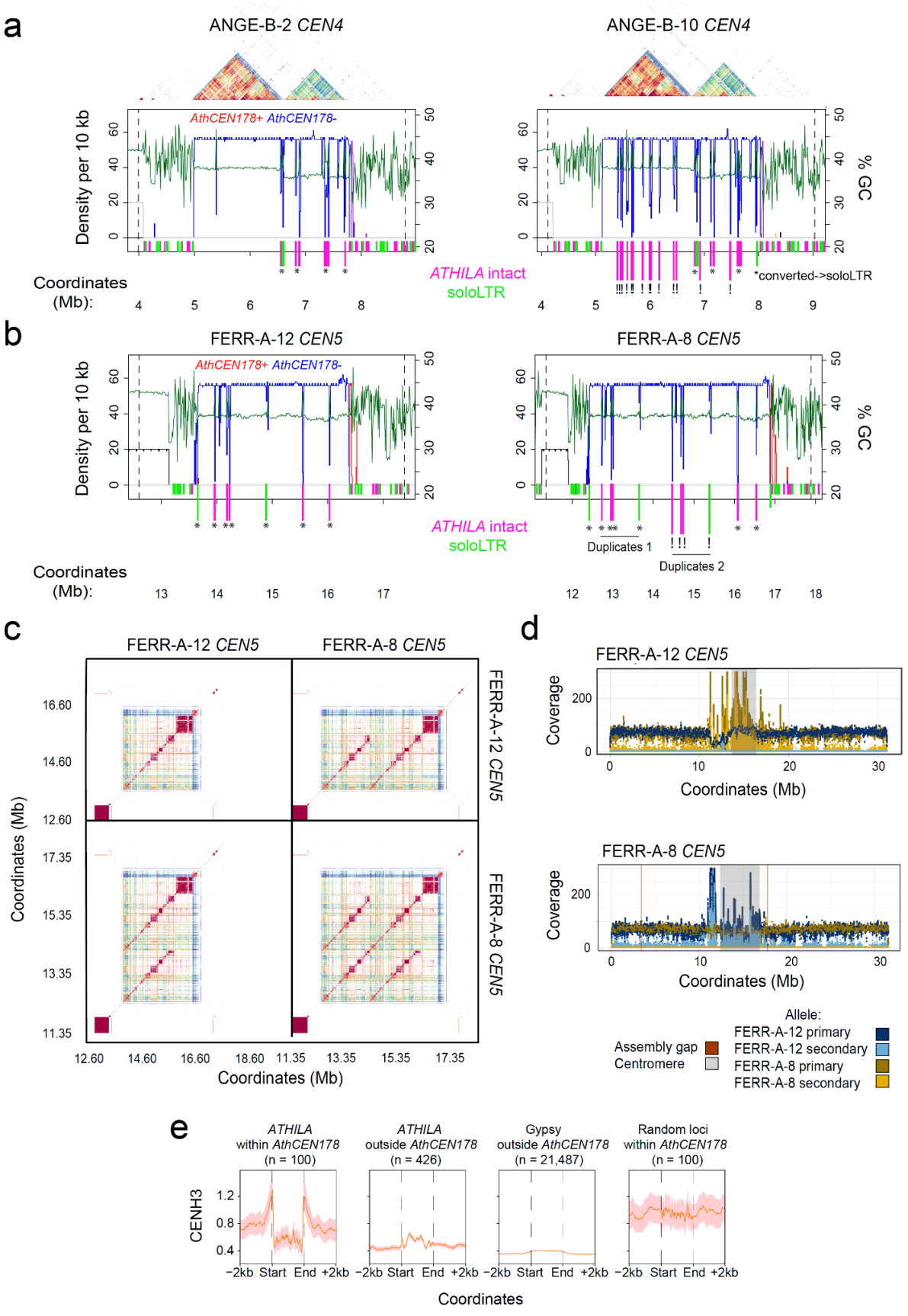
Extended Data Figure 5. DNA methylation, CENH3 ChIP-seq enrichment, and *AthCEN178* higher-order repeat (HOR) structure within the centromere regions of Col-0, Ler-0, Cvi-0 and Tanz-1. a, CENH3 ChIP-seq enrichment (log_2[ChIP/input], black) compared with *AthCEN178* density in 10 kb windows on forward (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 whole 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[59]. 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.



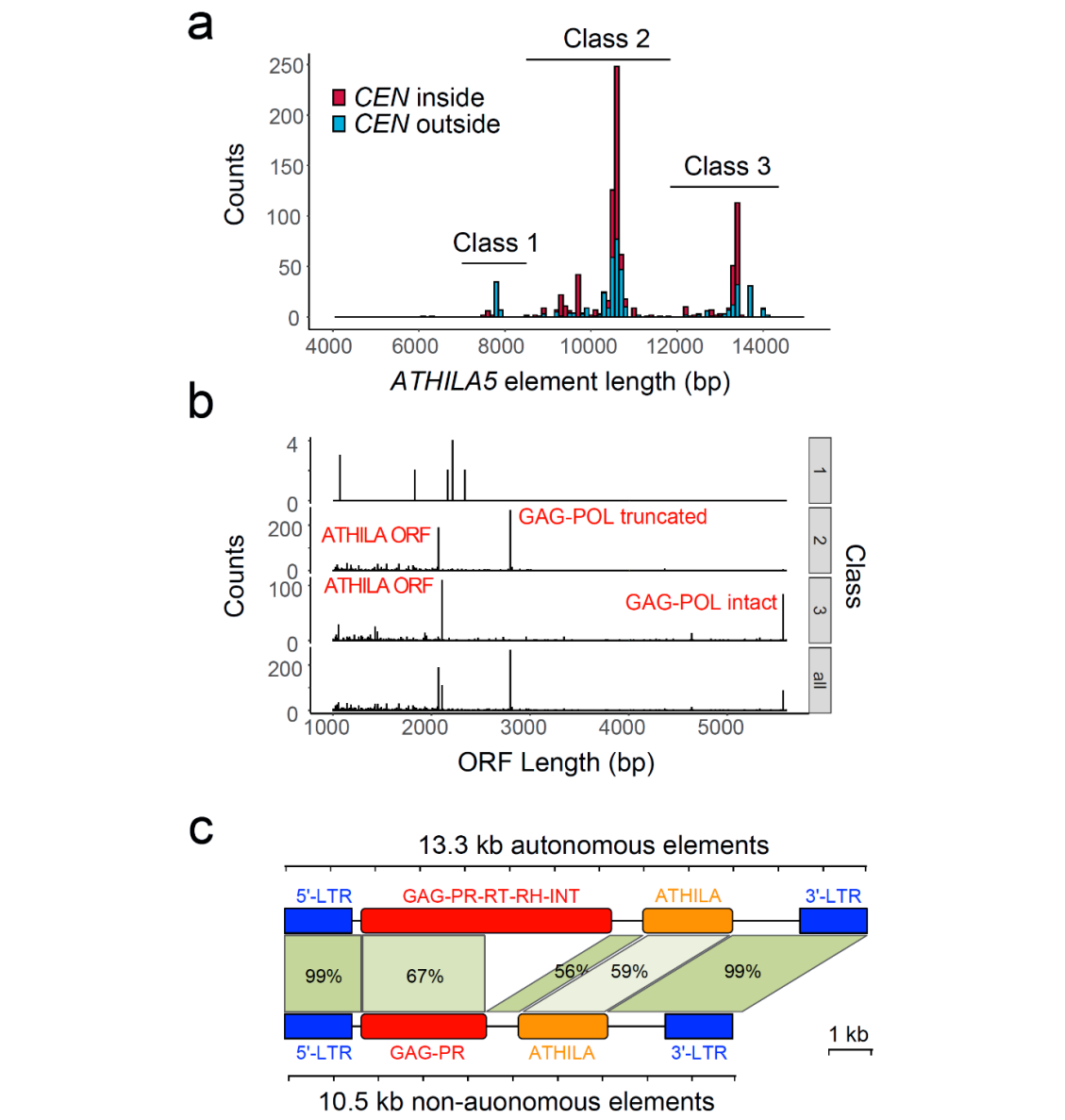
Extended Data Figure 6. Variation in *CENH3* coding sequence in relation to centromere *AthCEN178* similarity groups. A phylogenetic tree based on *CENH3* nucleotide sequences is shown for the 66 *A. thaliana* accessions (left). To the right of the tree is a coloured key indicating the *AthCEN178* similarity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


a
Centromere
Inside Outside
Centromere
Inside Outside
ATHILA5
ATHILA1
ATHILA6b
ATHILA6a
ATHILA2
Outside
Inside
Chr1
Chr2
Chr3
Chr4
Chr5
Intact
SoloLTR
b
Chr1
Chr4
Chr2
Chr5
Chr3
ATHILA1
ATHILA2
ATHILA5
ATHILA6a
ATHILA6b
ATHILA4c
c
ATHILA2
ATHILA5
ATHILA2
ATHILA5
Rabacal-1
Col-0
IP-Bus-0
d
ATHILA insert position within
CEN178 consensus
Frequency
0
20
40
60
0
50
100
150
Coordinates (bp)
e
Intact
SoloLTRs
Count
200
150
100
50
0
Chromosome arms
Count
40
30
20
10
0
AthCEN178 arrays
Accessions
Eurasian
Iberian relict
Iberian non-relict
Non-Iberian relict
f
Eurasian
Outside
Inside
Iberian relict
Iberian non-relict
Non-Iberian relict
Frequency
40
60
80
100
120
0
20
10
8
4
LTR identity (%)

Extended Data Figure 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.

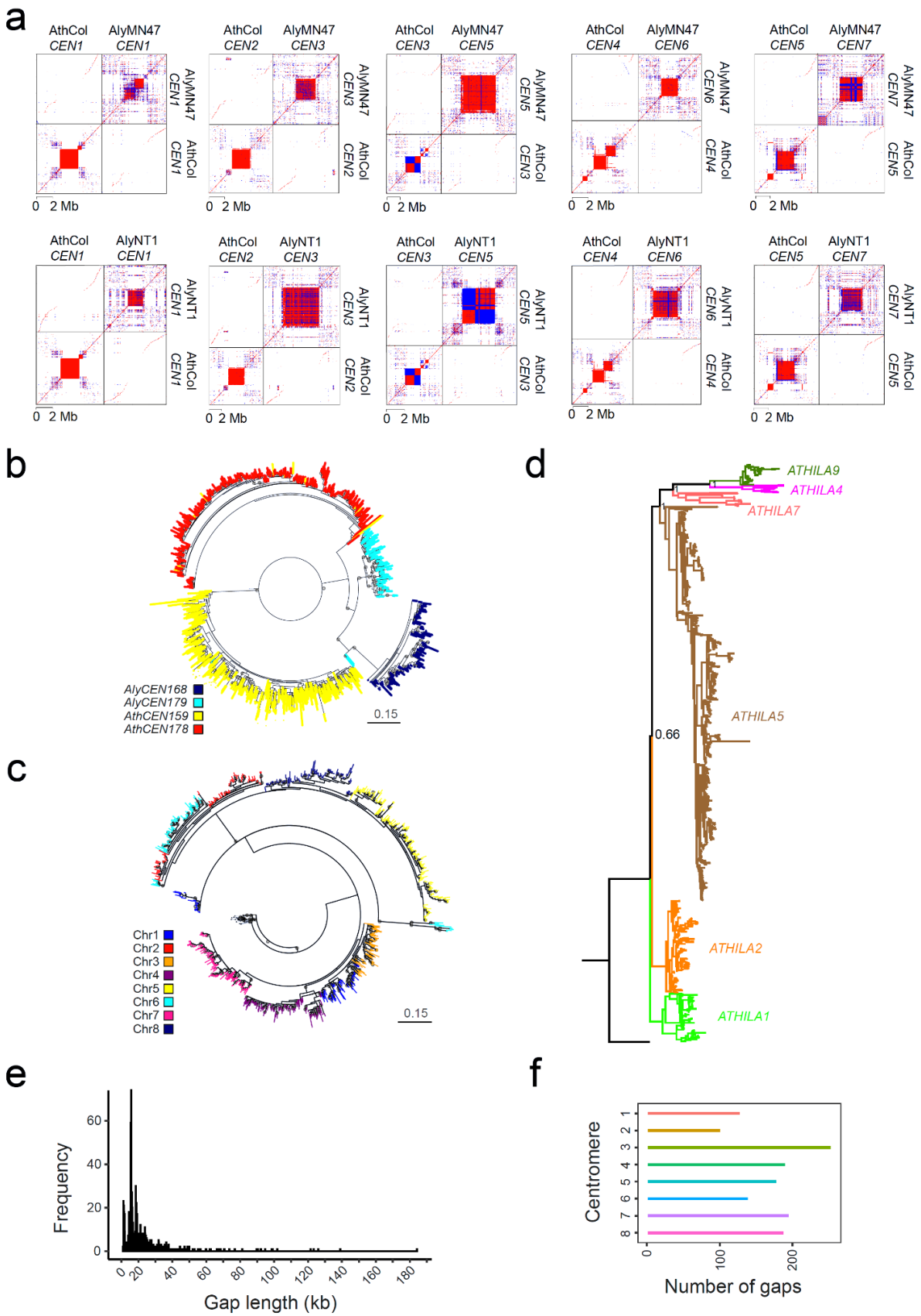

a
ANGE-B-2 CEN4
ANGE-B-10 CEN4
AthCEN178+ AthCEN178-
Density per 10 kb
% GC
ATHILA intact
soloLTR
*converted->soloLTR
Coordinates (Mb):
b
FERR-A-12 CEN5
FERR-A-8 CEN5
Duplicates 1
Duplicates 2
c
FERR-A-12 CEN5
FERR-A-8 CEN5
Coordinates (Mb)
d
Coverage
Allele:
Assembly gap
Centromere
FERR-A-12 primary
FERR-A-12 secondary
FERR-A-8 primary
FERR-A-8 secondary
e
ATHILA within AthCEN178 (n = 100)
ATHILA outside AthCEN178 (n = 426)
Gypsy outside AthCEN178 (n = 21,487)
Random loci within AthCEN178 (n = 100)
CENH3
−2kb Start End +2kb
Coordinates

Extended Data Figure 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. 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 Figure 9. Autonomous and non-autonomous *ATHILA* elements in the *A. thaliana* pancentromere. 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.


a
AthCol CEN1
AlyMN47 CEN1
AthCol CEN2
AlyMN47 CEN3
AthCol CEN3
AlyMN47 CEN5
AthCol CEN4
AlyMN47 CEN6
AthCol CEN5
AlyMN47 CEN7
AlyNT1 CEN1
AlyNT1 CEN3
AlyNT1 CEN5
AlyNT1 CEN6
AlyNT1 CEN7
0 2 Mb
b
AlyCEN168
AlyCEN179
AthCEN159
AthCEN178
0.15
c
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
0.15
d
ATHILA9
ATHILA4
ATHILA7
ATHILA5
0.66
ATHILA2
ATHILA1
e
Frequency
Gap length (kb)
f
Centromere
Number of gaps

Extended Data Figure 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.

Supplementary Table 1. Genome assembly parameters and metrics for 66 *A. thaliana* accessions. The table provides, for each of the 66 *A. thaliana* accessions analysed, the accession name, PCA group membership based on chromosome arm SNPs, ecotype id, country of origin, latitude and longitude of collection, whether DNA was extracted from multiple or single individuals, DNA extraction and shearing method, version of sequencing binding kit used, barcode name and sequence when multiplexed, Hifiasm version used for primary assembly, additional scaffolding information, information regarding data submission to ENA (Project and Sample IDs), HiFi read metrics (N50, total sum, number of reads and longest read), assembly metrics on scaffolded contigs (number of contigs that constituted the scaffolded chromosomes, remaining assembly gaps, largest and shortest contig, contig N50, total length, read depth mean and standard deviation), assembly quality metrics specific to centromeres (centromeres with assembly gaps, total sum of collapsed and expanded regions in Mb, and sum of heterozygosity stretches in total and per centromere). The first column includes a numerical code that can be matched with the first column of Supplementary Tables 2 and 3.

Supplementary Table 2. Key to centromere analysis of 66 *A. thaliana* accessions. This table provides, for each of the 66 *A. thaliana* accessions analysed, the name of the associated fasta file, chromosome arm SNP-PCA genetic group membership, accession code where available, country of origin, latitude and longitude of origin, the total number of *AthCEN178* repeats and their number per chromosome, the *AthCEN178* similarity group for each chromosome, the total number of *AthCEN159* tandem repeats, the number of intact *ATHILA* and soloLTRs in total and per chromosome, the total number of *AthCEN178* HORs (HORs), and *AthCEN178* per chromosome, and *AthCEN178* HOR score in total and per chromosome. Note that the following three pairs of accessions were collected at single sites, and had identical or nearly-identical genomes based on chromosome arm SNPs: (i) CAMA-C-2 and CAMA-C-9, (ii) BARC-A-12 & BARC-A-17, and (iii) BELC-C-10 & BELC-C-12. The first column includes a numerical code that can be matched with the first column of Supplementary Tables 1 and 3.

Supplementary Table 3. Centromere *AthCEN178* array coordinates in *A. thaliana* and *A. lyrata*. For each *A. thaliana* and *A. lyrata* accession, and each chromosome, the start and end coordinates of contiguous centromere satellite arrays are listed, in addition to the widths of each array. The first column includes a numerical code that can be matched with the first column of Supplementary Tables 1 and 2.

Supplementary Table 4. *AthCEN178* HOR parameters per accession. Supplementary Table 4 reports the average number of *AthCEN178* copies, higher-order repeats (HORs), average HOR scores, average HOR length (bp) and the average distance between HOR pairs (bp), for each PCA group defined by chromosome arm SNPs.

Supplementary Table 5. *ATHILA* annotation across 66 *A. thaliana* and two *A. lyrata* genomes. For each *ATHILA* family, we provide the number and proportion of intact elements and soloLTRs in the *AthCEN178* arrays versus the chromosome arms, the mean percentage of LTR identity, the number of elements with identical LTRs, the distribution across chromosomes, the genic coding capacity based on the presence of Pfam Hidden Markov Models, and the intact to soloLTR ratio. Centrophilic families that have invaded the *A. thaliana AthCEN178* satellite arrays are highlighted in red. The mean of LTR identity and intact to soloLTR ratios are shown only for families with >10 intact elements.

Supplementary Table 6. Matching *ATHILA* insertions between pairs of *A. thaliana* accessions and estimation of divergence based on sequence comparisons. Intact and soloLTR *ATHILA* insertions are shown for the centromeres of the following pairs of accessions: BARC-A12 & BARC-A-17, BELC-C-10 & BELC-C-12, CAMA-C-2 & CAMA-C-9, SALE-A-10 & SALE-A-17, ANGE-B-2 & ANGE-B-10, the duplicated region of FERR-A-8 *CEN5*, and the shared element in *CEN1* of BARC-A-17 & IP-Ini-0. Listed for each *ATHILA*, in columns B-S, are their start and stop coordinates, strand, quality (intact or soloLTR), total length (bp), LTR length (bp), target site duplication (TSD; when a TSD was not identified, the pentamer in column L shows the upstream 5

nucleotides of the *ATHILA*), LTR identity (%), *ATHILA* family assignment, and a code that indicates whether elements are matched between the accessions, are unmatched (i.e., ambiguous to classify), or represent a new insertion event in a specific accession. Columns T-AA contain information on the alignment of every matching *ATHILA* pair (substitutions, indels and alignment length), which were used to calculate divergence in generations using the formula: $T = K/2*\mu$, where K is the divergence calculated as: (substitutions+indels)/global_alignment_length (indels of any length count as 1 event), and μ is the estimated mutation rate of 7.0×10^{-9} per site per generation[76].

Supplementary Movie 1. StainedGlass gallery of 330 *A. thaliana* centromeres. An animation of StainedGlass sequence identity heat maps[59], generated using a 10 kb window size, for *CEN1, CEN2, CEN3, CEN4* and *CEN5*. For each chromosome, the centromeres are shown in order of similarity, as inferred from the *AthCEN178* sequence sharing heat maps in **Fig. 1D**. Beneath each heat map is a histogram of sequence identity percentages, showing colour assignments used in the sequence identity heat map.