

The mutational dynamics of the *Arabidopsis* centromeres

<https://doi.org/10.1038/s41586-026-11046-w>

Received: 2 June 2025

Accepted: 14 August 2026

Published online: 23 September 2026

Open access

 Check for updates

Xiao Dong (董笑)¹, Wen-Biao Jiao (焦文标)^{1,10,11}, Lara Goldkuhle², Fernando Rabanal³, Samija Amar¹, Matthew T. Parker¹, José A. Campoy^{1,12}, Yueqi Tao (陶玥琪)³, Bruno Huettel⁴, Jurriaan Ton⁵, Lisa M. Smith⁵, Holger Puchta², Detlef Weigel^{3,6} & Korbinian Schneeberger^{1,7,8,9}✉

Centromeres are essential for faithful chromosome segregation during cell division. Yet despite their conserved function, many centromeres contain highly variable but internally remarkably homogenized tandem-repeat arrays^{1–5} whose evolutionary dynamics remain poorly understood. Here, using replicated genome assemblies of mutation accumulation lines, we define the centromere-specific mutation spectrum in *Arabidopsis thaliana*. We find that kilobase-sized insertion–deletion mutations (indels) occur frequently and consistently preserve tandem-repeat arrays by adding or removing only complete repeat units. Point mutations accumulate at an almost tenfold higher rate than elsewhere in the genome, probably driven by non-allelic gene conversion between closely linked repeat units. These findings suggest a central role for homology-directed DNA repair in centromere evolution, further supported by the accumulation of more frequent and longer tandem-repeat-preserving indels in *Arabidopsis* lines that are deficient in the anti-recombinase helicase RTEL1. Forward-in-time simulations parameterized with the observed mutation spectrum show that kilobase-sized indels and point mutations alone are sufficient to generate the megabase-sized homogenized repeat blocks characteristic of natural centromeres. Together, our results show that centromere evolution is driven by a distinct mutational spectrum shaped by homology-directed DNA repair, providing a quantitative framework for understanding how mutational processes generate and maintain the large-scale architecture of centromeric DNA.

Centromeres are essential for faithful chromosome segregation during cell division. Despite their conserved function, centromeric DNA often consists of highly homogenized, megabase-scale tandem-repeat arrays that vary markedly within and between species^{1–5}. The mutational processes driving this rapid sequence turnover remain poorly understood.

The individual units of centromeric repeat arrays are typically short^{6,7} (100–200 bp) and highly similar, but not identical, in sequence. Often, specific combinations of different repeat units are repeated multiple times, forming higher-order repeat (HOR) regions. Closely linked and highly similar HOR regions can further expand into homogenized blocks that span several megabases, producing the characteristic large-scale patterns of sequence similarity that are commonly observed when visualizing sequence similarities across entire centromeres^{1,5,6}.

The functional core of a centromere typically consists of such homogenized blocks, whereas the peripheries tend to be more heterogeneous and contain more transposable element (TE) insertions and structural rearrangements^{1,5}. Although this general organization is usually conserved, the number, size and arrangement of the homogenized

blocks vary greatly even between individuals of the same species. This variation has led to the hypothesis that megabase-scale mutations or long-range recombination events give rise to the characteristic structures of centromeric satellite arrays⁷.

In *A. thaliana*, the five centromeres are composed of tandem-repeat arrays derived from a 178-bp repeat monomer (*CEN178*)⁸. Long-read assemblies have reconstructed the centromeric tandem arrays in both the reference accession Col-0 and large populations of diverse *Arabidopsis* accessions^{5,8–10}. The extreme sequence divergence observed in natural centromeres supports the hypothesis that these regions evolve through distinct mutational dynamics based on homology-directed DNA repair^{1,11}, with gene conversion and unequal crossovers long proposed to have dominant roles in tandem-repeat evolution and homogenization^{2,12–18}.

To better understand these mutational processes, we analysed *A. thaliana* mutation accumulation (MA) lines, which were propagated through repeated single-seed descent over multiple generations to allow mutations to accumulate. Although long-read sequencing now

¹Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany. ²Department of Molecular Biology, Joseph Gottlieb Kölreuter Institute for Plant Sciences, Karlsruhe Institute of Technology, Karlsruhe, Germany. ³Department of Molecular Biology, Max Planck Institute for Biology, Tübingen, Germany. ⁴Max Planck Genome-Centre Cologne, Max Planck Institute for Plant Breeding Research, Cologne, Germany. ⁵School of Biosciences, Plants, Photosynthesis and Soil Research Cluster, University of Sheffield, Sheffield, UK. ⁶Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany. ⁷Faculty of Biology, LMU Munich, Martinsried, Germany. ⁸Institute for Crop Biology, Faculty of Mathematics and Natural Sciences, Heinrich-Heine University, Düsseldorf, Germany. ⁹Cluster of Excellence on Plant Sciences, Heinrich-Heine University, Düsseldorf, Germany. ¹⁰Present address: National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops, Huazhong Agricultural University, Wuhan, China. ¹¹Present address: College of Informatics, Huazhong Agricultural University, Wuhan, China. ¹²Present address: Department of Pomology, Estación Experimental de Aula Dei (EEAD), CSIC, Saragossa, Spain. ✉e-mail: korbinian.schneeberger@hhu.de

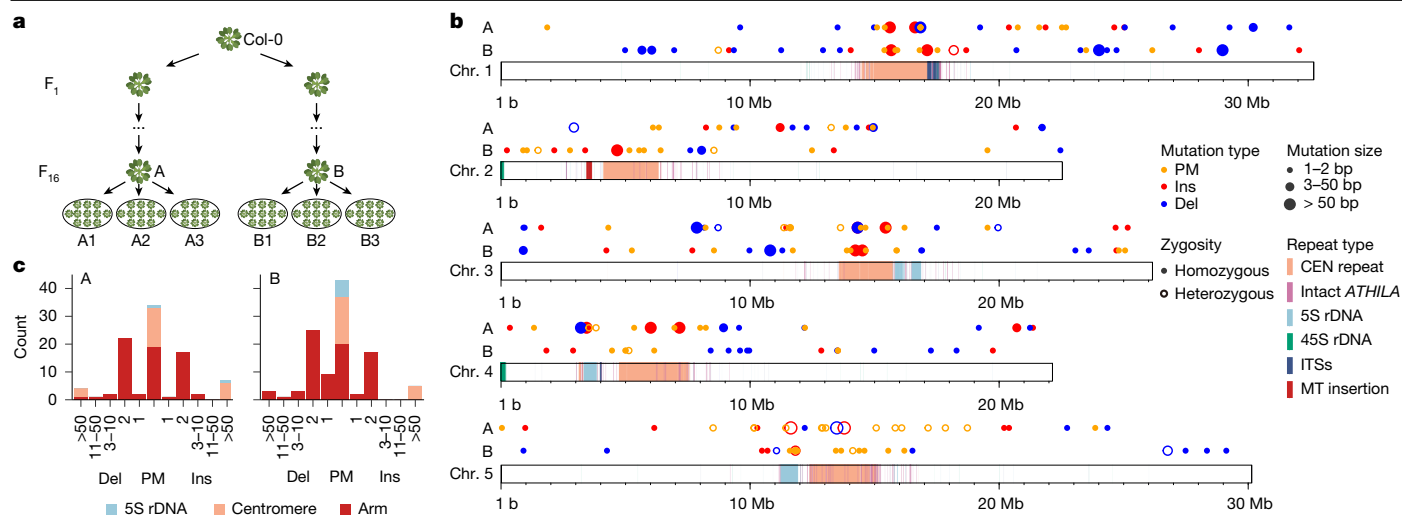


Fig. 1 | Mutation accumulation in *Arabidopsis* revealed by virtually error-free genome assemblies. **a**, Pools of sister plants (approximately 15 individuals per pool) were sequenced, with 3 independent pools generated from the progeny of each mother plant (designated A or B). Pooling sister genomes dilutes somatic mutations and reconstructs the genotype of the maternal plant (F_{16}). Sequence differences between assemblies of sister pools therefore represent assembly errors. **b**, Distribution of true mutations in samples A and B. Two rows of circles depict mutation positions, colour-coded by type; circle size reflects mutation

size; solid circles mark homozygous mutations and open circles indicate heterozygous mutations; and chromosomes are shown as rectangles with repetitive regions highlighted and colour-coded as indicated. CEN, centromere; Del, deletions; Ins, insertions; ITS, interstitial telomeric sequences; MT, mitochondrial; PM, point mutations. **c**, Bar plots showing mutation counts in samples A and B. The central bar denotes point mutations, bars to the right indicate insertions, and bars to the left indicate deletions. Mutation size increases from the centre outward. Colours indicate genomic context.

enables the complete assembly of centromeres^{5,8–10}, identifying rare mutations in these repetitive regions remains challenging because residual assembly errors can still obscure genuine mutational events.

To overcome this challenge, we developed a replicated genome assembly strategy that distinguishes genuine centromeric mutations from assembly artefacts. Applying this approach to MA lines revealed a unique mutational spectrum in centromeres. Frequent tandem-repeat-preserving indels spanning several kilobases added or removed complete repeat units while maintaining the integrity of the repeat array. In addition, we found an almost tenfold enrichment in point mutations, including clusters of point mutations that were probably introduced by non-allelic gene conversion (NAGC). Using MA lines derived from an *RTEL1* mutant, we implicate *RTEL1* in maintaining centromeric repeat array stability and highlight homology-directed DNA repair as a key driver of centromere evolution. In line with these findings, forward-in-time simulations demonstrated that kilobase-scale indels and point mutations are sufficient to generate megabase-scale homogenized blocks, although rare megabase-scale deletions are required to maintain centromere size. Together, our findings uncover the mutational dynamics that shape centromere architecture and drive its evolutionary turnover.

Replication resolves assembly errors

Despite considerable progress in recent years, de novo genome assemblies still contain errors, making it difficult to confidently identify rare mutations in highly repetitive regions. To overcome this limitation, we generated replicated genome assemblies from independent DNA samples of individual plants to first identify and correct assembly errors before detecting genuine mutations.

We generated three independent genome assemblies for each of two *A. thaliana* Col-0 plants using DNA extracted from pooled progeny (Fig. 1a). Both plants were derived from the same Col-0 founder but were independently propagated by self-pollination and single-seed descent for 16 generations as part of a controlled mutation accumulation experiment. We refer to these two MA16 plants as A and B, and to their replicated assemblies as A1–A3 and B1–B3, where numbers

denote the respective genome assembly (Fig. 1a). Differences among the replicated assemblies of the same plant are expected to arise solely from assembly errors, whereas the differences between assemblies of different plants also include mutations that accumulated after their lineages separated.

Two assemblies of each plant were generated using PacBio HiFi (replicates 1 and 2), and a third was generated with Oxford Nanopore Technologies (ONT) sequencing (replicate 3) (Supplementary Table 1). All assemblies were highly contiguous, with the ONT assemblies reconstructing four of the five chromosomes as single contigs (Supplementary Table 2). As expected, none of the assemblies resolved the highly repetitive nucleolus organizer regions on chromosomes 2 and 4, which remain challenging for automated assembly approaches¹⁹. To identify assembly errors, we compared all replicated assemblies to each other and identified 266–577 errors in the PacBio assemblies and 10,801–10,974 errors in the ONT assemblies (Extended Data Fig. 1 and Methods). Despite the large number of assembly differences, we manually inspected each difference by visualizing the underlying read alignments.

The error profiles differed substantially between the two sequencing technologies (Supplementary Results). PacBio assemblies mainly contained small indels associated with simple sequence repeats, low-complexity regions and occasional scaffolding artefacts, whereas ONT assemblies were dominated by indel errors in homopolymers. Conventional assembly polishing had contrasting effects on the assemblies: it substantially improved the ONT assemblies, but overcorrected the PacBio assemblies by introducing additional errors. In both cases, however, residual errors remained (Supplementary Figs. 1–11 and Supplementary Tables 3–10).

Across the six assemblies, 26 of the 30 centromeres were assembled as single contigs. Unexpectedly, centromeric sequences were among the most accurately assembled regions of the genome. Although they comprise 8.3% of the genome, they accounted for less than 1% of all assembly errors in both ONT and PacBio assemblies. In total, we detected only 15 assembly errors within centromeres, including 9 large errors (>50 bp) and 6 small errors (1–2 bp). The large errors primarily resulted from mis-scaffolding of the PacBio assembly contigs, while the small errors were associated with simple sequence repeats.

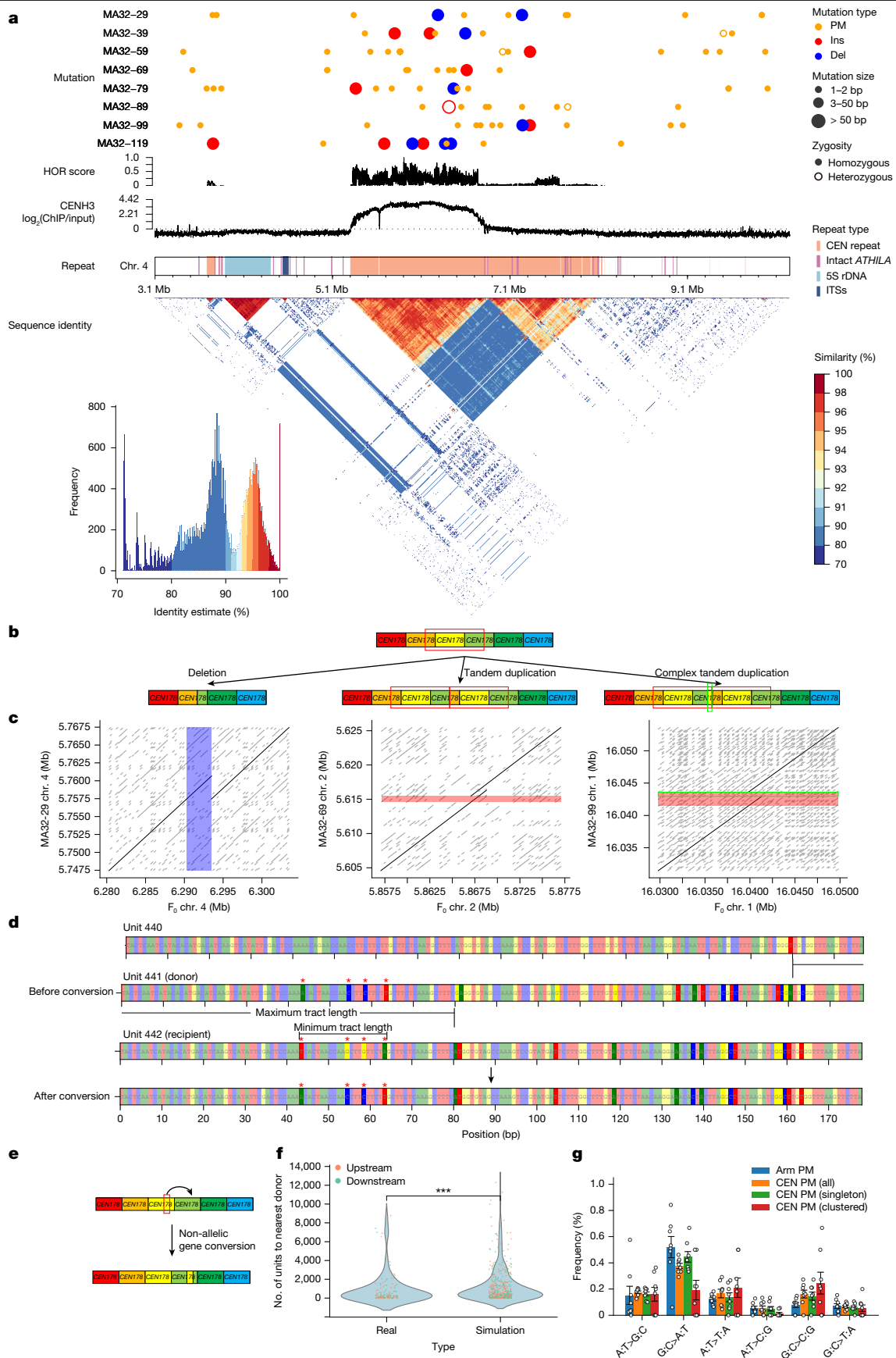


Fig. 2 | See next page for caption.

Fig. 2 | Mutations within centromeres. **a**, Top, eight rows of circles display mutation patterns in the chromosome 4 centromere of MA32 samples. Rows of circles depict mutation positions, colour-coded by type; circle size reflects mutation size; solid circles mark homozygous mutations and open circles indicate heterozygous mutations; and chromosomes are shown as rectangles with repetitive regions highlighted and colour-coded as indicated. The two line graphs show HOR scores and $\log_2$ -transformed CENH3 chromatin immunoprecipitation with sequencing (ChIP-seq) enrichment⁵⁶. Bottom, heat map showing pairwise sequence identity between non-overlapping 10 kb regions, with a histogram summarizing identity values in the lower left corner. **b**, Schematics illustrate deletion, tandem duplication and complex tandem duplication; red boxes highlight the mutated patterns. **c**, Dot plots show

sequence identity between wild-type and mutated sequences. The longest identical sequences are shown in black and others are depicted in grey. Blue and red indicate deleted and inserted regions, respectively; the green box marks a mosaic region. **d**, A schematic example shows the four point mutations generated by a single NAGC event. **e**, Schematic representation of NAGC in centromeres. **f**, Violin plots compare the distance to the donor unit in observed point mutations and 500 random centromeric positions. Donors are colour-coded as downstream or upstream (two-sided Mann–Whitney *U* test, $P = 2.9 \times 10^{-5}$). **g**, Bar plots show the mutation spectra across chromosome arms, all centromeric point mutations, singleton centromeric point mutations and clustered centromeric point mutations. Bars represent mean $\pm$ s.e.m.; dots indicate individual biologically independent MA lines ($n = 8$ per group).

This high accuracy is likely to be due to the marked depletion of simple sequence repeats within centromeres, which were the main source of assembly errors elsewhere in the genome (Extended Data Fig. 1a and Supplementary Results).

Together, these results demonstrate that *Arabidopsis* centromeres can be assembled with extremely high accuracy using either PacBio or ONT sequencing. Moreover, replicated genome assemblies provide a robust framework for distinguishing genuine centromeric mutations from assembly errors.

Mutation identification with whole-genome assemblies

By aligning the error-corrected assemblies from different samples with one another, mutations of any type can be identified throughout the assembled genomes. As heterozygous mutations may also be represented in the assemblies, all candidate mutations were validated against raw-read alignments, and only homozygous mutations were retained for further analysis.

Between the genomes of the two MA16 lines, we identified 200 homozygous mutations, comprising 77 point mutations and 123 indels ranging from 1 to 11,570 bp (Fig. 1b,c and Supplementary Table 11). Using the Col-0 reference genome, we assigned 92 mutations to plant A and 108 to plant B.

To assess the completeness of the mutation calls, we compared the observed mutation rates with previous estimates for the chromosome arms of the *Arabidopsis* genome^{20,21}. Outside the centromeres and 5S rDNA clusters, the number of observed point mutations (19 in A and 20 in B) was consistent with expectations based on previous estimates. Likewise, the observed numbers of mutations within TEs were consistent with previously reported mutation rates in these regions^{20,21}. Together, these comparisons indicate that our mutation detection approach provides highly accurate and complete mutation calls.

The most common mutations in unique genomic regions were 2-bp indels, all of which occurred in dinucleotide repeats in the chromosome arms. This is again consistent with previous reports showing substantially elevated mutation rates in dinucleotide repeats^{21,22}. All other small indels (1–50 bp) also occurred within simple sequence repeats, including three cases involving repeat units of up to 25 bp.

Unexpectedly, 38 of the 77 homozygous point mutations (49%) were located within just 13.8 Mb (9.8%) of the genome, corresponding to the highly repetitive 5S rDNA clusters and centromeres. This suggests an elevated point mutation rate in these regions compared with unique regions and indicates that these regions undergo distinct mutational dynamics as previously described for pericentromeric regions²¹. These findings add to the growing evidence that mutation rates are not uniformly distributed across the genome^{21,23}.

Likewise, the majority of the large indels (>50 bp; 15 out of 19) occurred in highly repetitive regions such as the centromeres and 5S rDNA clusters (Fig. 1c). Only four large deletions were observed outside these regions, three of which shared significant sequence similarity

with closely linked regions, implying that local sequence similarity is a major driver of large indel formation across the genome.

Together, our approach enables the confident identification of all mutation types across the assembled genome, including within the highly repetitive centromeres.

Mutation spectrum in the centromere

To enable accurate estimation of mutation spectra and frequencies while avoiding potential biases arising from using the same genomes for method development and mutation detection, we generated high-quality genome assemblies for 8 additional MA lines propagated for 32 generations. As before, assemblies were produced using the replicated approach with both PacBio and ONT sequencing. In total, these data represent 256 generations of accumulated mutations.

Across the 40 centromeres of the eight MA32 lines, we identified 270 homozygous mutations, including 208 point mutations, 56 large indels (177–5,507 bp), and 6 small indels (1–9 bp) (Fig. 2, Supplementary Figs. 12–15 and Supplementary Table 12). These mutations were consistently distributed across the eight lines and all five chromosomes (Extended Data Fig. 2).

The 56 large indels showed a striking pattern: all preserved the *CEN178* repeat periodicity and exclusively added or removed complete repeat units (Fig. 2b,c). Consequently, the overall structure of the satellite array remained intact over generations. Insertions, however, occurred more frequently than deletions (paired Wilcoxon signed-rank test, $P = 0.062$) and were significantly larger (two-sided Mann–Whitney *U* test, $P = 0.0024$): the 21 deletions removed 1 to 18 repeat units (mean size: 923 bp), whereas the 35 insertions added 1 to 31 units (mean size: 1,870 bp) (Extended Data Fig. 3). In all cases, inserted sequences were either identical to adjacent repeat units (tandem duplications) or were mosaics of adjacent repeat units (complex tandem duplications) (Fig. 2b,c). In addition, most indels generated novel repeat unit variants at their breakpoints by combining segments of two different units (Fig. 2b,c).

The 208 point mutations were distributed across the entire length of the centromeres, with 95% occurring in *CEN178* repeat units and the remaining 5% occurring in centromere-associated TEs (*ATHILA* elements). However, 59 (28%) clustered into 16 mutation groups (Methods). For example, 4 point mutations were clustered within a 20-bp region at the centre of a single *CEN178* repeat unit (Fig. 2d). The adjacent upstream repeat unit contained the exact same sequence variants as those introduced by the four mutations (Fig. 2d,e). Putative donor sequences were identified for 15 out of the 16 clusters, all located on the same chromosome as the mutated cluster, at a median distance of 16 repeat units.

These patterns are most parsimoniously explained by NAGC events that convert sequence variants between different *CEN178* repeat units. The average length of the conversion tract was 58.5 bp (Supplementary Table 12), slightly longer than the previously estimated tract lengths for non-crossover-associated meiotic gene conversions (25–50 bp)²⁴.

Table 1 | Different mutational dynamics in centromeres and chromosome arms

	Predominant mutation types	Point mutation rate per site per generation
Centromeric repeat arrays	Point mutations (NAGCs and spontaneous mutations) Indels (approximately 180 bp to 10 kb)	6.5×10^{-8}
Chromosome arms	Point mutations (spontaneous) Small indels in simple sequence repeats (1–30 bp)	7.0×10^{-9} (ref. 21)

We next searched for putative donor sequences for the 149 singleton point mutations, which may also have arisen through NAGC. Putative donor sequences were identified for 129 mutations (87%). However, donor-like matches were also detected for 86% of simulated spontaneous point mutations, indicating that sequence similarity alone is insufficient to infer NAGC. By contrast, the distances between mutations and their putative donors were significantly shorter in real data than in simulations (Fig. 2f), implying that at least a subset of the singleton point mutations also arose from NAGC.

The presence of NAGC events was further supported by differences in the mutation spectrum. Whereas GC→AT transitions account for the majority of point mutations in chromosome arms (50–60%)^{20,21}, their frequency among point mutations in the centromere was significantly reduced, to around 40%. This difference could not be explained by the slightly higher GC content of centromeres (37.4%) than that of chromosome arms (36.2%). Among clustered mutations, which are most likely to result from NAGC events, the frequency of GC→AT transitions was further reduced to about 20% (Fig. 2g). GC→AT transitions accounted for about 45% of singleton point mutations, intermediate between chromosome-arm and clustered mutations, again suggesting some singleton mutations might also be introduced by NAGC events. Notably, this transition rate reflects the existing sequence differences between repeat units and does not require any GC-biased gene conversion to explain the observed mutation spectrum (Extended Data Fig. 4).

Together, centromeres exhibit a markedly distinct mutation spectrum compared with that of unique regions in the chromosome arms. Centromere mutations are characterized by tandem-repeat-preserving indels spanning several kilobases that add or remove complete repeat units allowing repeat arrays to evolve without disrupting their periodicity. In addition, NAGC events, which transfer short sequence tracts between adjacent repeat units, contribute to the markedly elevated point mutation rate in centromeres.

Mutation rate in the centromere

We estimated the point mutation rate in centromeric repeat arrays (including both spontaneous mutations and point mutations arising from NAGC events) to be 6.5×10^{-8} (95% confidence interval: $5.6\text{--}7.5 \times 10^{-8}$) per site per generation. This rate is almost tenfold higher than recent estimates for chromosome arms in *A. thaliana*²¹ (Table 1). Only methylated cytosines within TEs have been reported to exhibit similar mutation rates, whereas all other genomic regions mutate at significantly lower rates²¹.

To disentangle the contributions of the distinct mutational processes, we decomposed the point mutations according to differences in the transition/transversion (Ti/Tv) ratios associated with each mutational process (Methods). This analysis indicated that approximately 30% of point mutations in centromeres arise from spontaneous mutation, whereas 70% of the point mutations are attributable to NAGC. These proportions correspond to a spontaneous point mutation rate of 2.0×10^{-8} per site per generation, which is still approximately threefold higher than that in chromosome arms, and a NAGC-associated point mutation rate of 4.5×10^{-8} per site per generation.

Together, these results indicate that centromeric point mutations are markedly elevated and are driven by both frequent NAGC events and spontaneous mutations.

Centromeric TE movement is rare

TEs are common components in centromeric repeat arrays. In *Arabidopsis*, *ATHILA* retrotransposons are enriched in centromeric and pericentromeric regions, and previous studies have suggested that these elements contribute substantially to the evolution and divergence of centromeres^{1,5}. Across the 8 MA32 lines propagated for a total of 256 generations, we did not detect a single TE insertion or deletion event in the centromeres. In addition, mutations in the TEs were underrepresented compared with the rest of the centromere. Only 11 (5%) of the centromeric point mutations occurred within *ATHILA* elements, even though these elements comprise 12% of the centromeric sequence in the Col-0 genome.

To search for footprints of recent TE activity in centromeres, we analysed the genomes of five *Arabidopsis* individuals from the HPG1 group, a homogeneous lineage of *Arabidopsis* lines that recently colonized North America. This group is likely to have originated from a single ancestor that was introduced from Eurasia approximately 400 years ago^{25–27}. The five HPG1 individuals were collected from geographically distinct locations and represent a subset of this lineage, which includes non-admixed, non-recombined and quasi-identical individuals that differ only by mutations accumulated since their separation²⁵ (Fig. 3a and Supplementary Table 1). Although natural selection and genetic drift may have influenced the mutations retained in these genomes, they provide a ‘natural experiment’ for identifying mutations accumulated over around 1,500 generations (assuming an average generation time of 1.3 years)²⁵.

We assembled the genomes of all five HPG1 accessions using PacBio long-read sequencing. As in the MA lines, however, we found no evidence of TE movement in centromeres of the HPG1 genomes: all 152 intact *ATHILA* insertion sites within the centromeres were conserved across the five individuals (Supplementary Fig. 16). In fact, no TE insertions, deletions, or transposition events were detected anywhere in the HPG1 genomes. This contrasts with previous population-scale analyses of TE diversity in 216 *Arabidopsis* accessions, which estimated fewer than one TE insertion per 60 generations²⁸, and with reports of individual accessions harbouring dozens of unique TE insertions¹. The absence of TE insertions or deletions in the HPG1 accessions, which have accumulated mutations over around 1,500 generations, suggests that TE activity may not occur at a constant rate but instead in occasional bursts²⁹, such that no events might be observed over extended time intervals.

Despite the absence of TE movement, we identified substantial sequence variation within the centromeric arrays of HPG1 accessions. In total, we detected 1,415 homozygous mutations, including 1,113 point mutations (28 located within *ATHILA* elements), 275 large, kilobase-scale indels (116–13,695 bp) and 27 small indels (1–4 bp) (Supplementary Table 13). Consistent with the patterns observed in the MA lines, kilobase-scale insertions were more frequent than deletions (189 insertions versus 86 deletions) and almost all of these indels preserved the repeat periodicity, with a single exception: a 116-bp deletion disrupted the tandem-repeat structures and introduced a unique 240-bp fragment into the repeat array (Fig. 3b,c and Extended Data Fig. 3).

The observed number of point mutations in HPG1 centromeres is in close agreement with that expected from the empirically estimated mutation rate (Fig. 3d). In total, 367 of the 1,113 point mutations formed 81 clusters within centromeric regions. Of those, 61 clusters could be linked to putative donor sequences, consistent with NAGC events. The remaining clusters lacked identifiable donors, possibly owing to subsequent mutations that have obscured sequence similarity between the donor and mutation cluster.

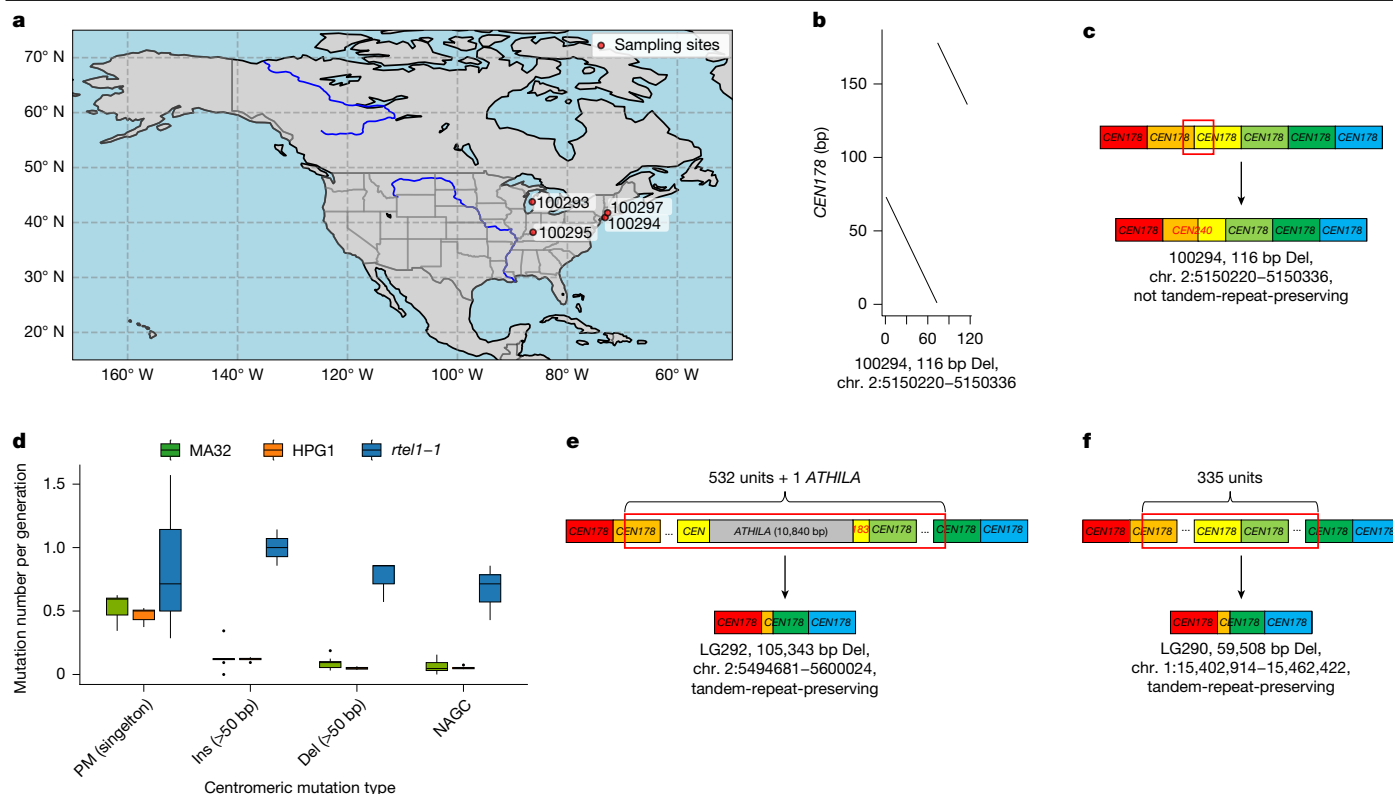


Fig. 3 | Accumulation of centromeric mutations in HPG1 lines and *rtel1-1* mutants. **a**, Geographic distribution of sampling locations for five HPG1 samples; no coordinate information is available for sample 100298. Map Data from Natural Earth, accessed through Cartopy, CC0 1.0. **b**, **c**, A 116-bp deletion in the HPG1 lineage that does not preserve repeat periodicity. **b**, Dot plot showing alignment of the 116-bp deletion to the CEN178 consensus sequence. **c**, Schematic representation of the 116-bp deletion. **d**, Box plots showing the number of mutations per generation across three groups. MA32 ($n = 8$), HPG1 ($n = 5$) and *rtel1-1* ($n = 3$) represent independent lines or accessions. Centre lines indicate the median; box limits indicate the top and bottom quartiles; whiskers extend to the most extreme values within $1.5 \times$ the interquartile range; and dots indicate outliers. Clustered point mutations are classified as NAGC events, whereas isolated point mutations are classified as singleton point mutations. Two-sided Mann–Whitney U tests comparing MA32 and HPG1 show no significant differences for insertions ($U = 20$, $P = 1.000$), deletions ($U = 29$, $P = 0.210$) or isolated point mutations ($U = 29$, $P = 0.210$). **e**, **f**, Schematics of the two largest deletions identified in *rtel1-1*: 105,343 bp (**e**) and 59,508 bp (**f**). The former includes a complete *ATHILA* element. Despite their large sizes, both deletions preserve repeat periodicity. Breakpoint positions are shown to scale in the schematics.

Homology-directed repair drives centromeric mutations

A common feature of tandem-repeat-preserving indels and NAGC events is their reliance on sequence homology within the centromeric repeat array. This suggests that homology-directed DNA repair contributes both to repeat array stability and, through its intrinsic mutagenic potential³⁰, to centromeric evolution.

To test whether altered homology-directed repair affects centromeric mutational dynamics, we analysed MA lines deficient in REGULATOR OF TELOMERE ELONGATION 1 (RTEL1), a DNA helicase that dismantles recombination intermediates such as D-loops and thereby limits inappropriate or prolonged homologous recombination³¹. In *Arabidopsis*, RTEL1 deficiency causes hyperrecombination³² and a massive reduction of 45S rDNA copy number³³.

We generated replicated PacBio and ONT genome assemblies for three *rtel1-1* MA lines propagated for seven generations. Overall, we identified 127 homozygous centromeric mutations, including 75 point mutations, 49 large indels (29 insertions, 20 deletions) and 3 small indels (Supplementary Table 14). As in wild-type centromeres, point mutations frequently formed local clusters, with 57 of the 75 mutations occurring in 13 distinct clusters. Despite accumulating mutations for only 21 generations, the 3 *rtel1-1* MA lines carried almost as many centromeric indels as the 8 wild-type MA32 lines, which together accumulated mutations for 256 generations (1.8 versus 0.2 large singleton indels and 0.7 versus 0.1 NAGC clusters per generation) (Fig. 3d).

whiskers extend to the most extreme values within $1.5 \times$ the interquartile range; and dots indicate outliers. Clustered point mutations are classified as NAGC events, whereas isolated point mutations are classified as singleton point mutations. Two-sided Mann–Whitney U tests comparing MA32 and HPG1 show no significant differences for insertions ($U = 20$, $P = 1.000$), deletions ($U = 29$, $P = 0.210$) or isolated point mutations ($U = 29$, $P = 0.210$). **e**, **f**, Schematics of the two largest deletions identified in *rtel1-1*: 105,343 bp (**e**) and 59,508 bp (**f**). The former includes a complete *ATHILA* element. Despite their large sizes, both deletions preserve repeat periodicity. Breakpoint positions are shown to scale in the schematics.

Again, all large indels preserved the tandem-repeat periodicity and the size distribution of centromeric indels in *rtel1-1* lines was similar to that in wild-type MA lines (Extended Data Fig. 3). However, two deletions in *rtel1-1* MA lines were markedly longer (59,508 bp and 105,343 bp), removing 335 repeat units and 532 repeat units, respectively. Despite their size, both deletions preserved the tandem-repeat periodicity intact (Fig. 3e,f), consistent with the role of RTEL1 in dismantling recombination intermediates, which otherwise can lead to large mutations.

Together, the *rtel1-1* mutation causes a massive increase in the centromeric mutation rate while largely preserving the characteristic tandem-repeat-preserving mutation spectrum observed in wild-type lines. This is consistent with RTEL1 limiting excessive homology-directed DNA repair and provides genetic evidence that homology-directed repair is a major driver of centromeric mutational processes in wild-type plants.

Emergence of homogenized blocks

In both animals and plants, centromeric repeat arrays often contain large homogenized blocks spanning hundreds of kilobases to megabases^{1,11}. These blocks are composed of highly similar repeat units, contain many HORs, and are typically distinct from the surrounding repeat units (Fig. 4a). Functional centromeres are often concentrated in highly homogenized blocks. Yet despite this conserved function, the homogenized blocks show little sequence conservation across

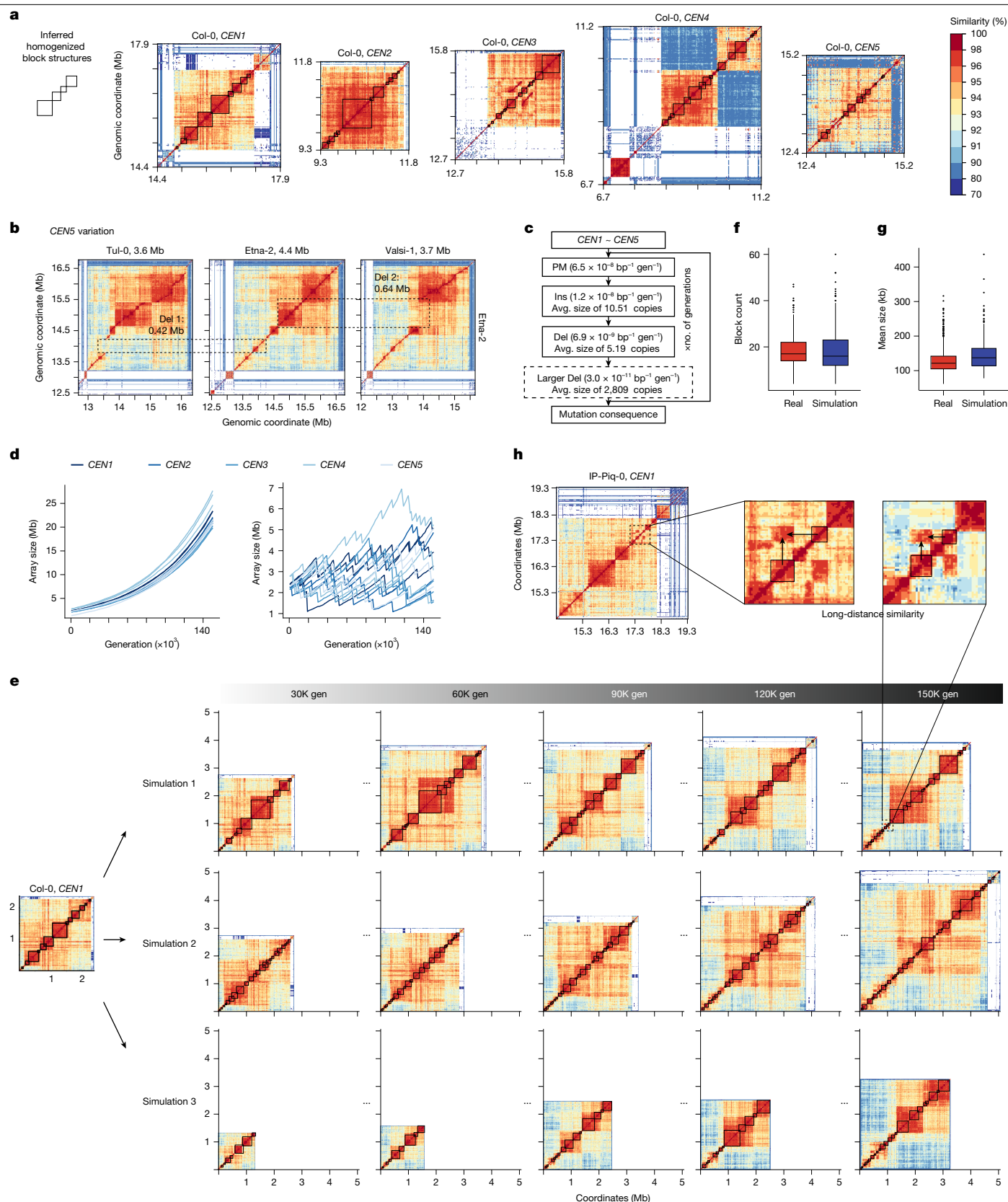


Fig. 4 | See next page for caption.

the global population of *Arabidopsis* accessions (examples shown in Extended Data Fig. 5a). Although several mechanisms have been proposed to explain their origin and turnover (for example, layered expansion, tandem duplication or unequal crossover^{1,11}), how such homogenized blocks arise remains unclear.

To investigate how centromeric mutations contribute to the formation of homogenized blocks, we performed forward-in-time simulations using the empirically estimated mutation spectrum to introduce mutations into the centromeric sequences of all five *Arabidopsis* chromosomes. Because the insertion rate was approximately twofold

Fig. 4 | Simulations generate block structures resembling those of natural centromeres. **a**, Heat map showing self-sequence identity across the centromeres of all five chromosomes in Col-0. Homogenized block structures are highlighted by black boxes (see Methods for definition). **b**, Pairwise heat maps comparing chromosome 5 of Etna-2 with Tul-0 and Valsi-1, revealing large-scale structural differences, including two putative deletions highlighted by dashed boxes. **c**, Schematic overview of the simulation pipeline. Mutation frequencies were estimated from the eight MA32 lines and combined with megabase-scale deletion events (dashed box). Gen, generation. **d**, Changes in centromere size during simulations without (left) or with (right) the incorporation of Mb-scale deletions. **e**, Heat maps showing self-sequence identity of simulated centromeres over 150,000 generations. Simulations were initiated from the Col-0 centromeric

sequences and incorporated mutation patterns observed in the MA32 lines together with large deletions (mean size 500 kb). Homogenized blocks are highlighted by black boxes. **f,g**, Box plots comparing the number (**f**) and mean length (**g**) of homogenized blocks between 570 real centromeres and 500 simulated centromeres (100 replicates per chromosome) at generation 150,000. Centre lines indicate the median; box limits indicate the top and bottom quartiles; whiskers extend to the most extreme values within 1.5× the interquartile range; and dots indicate outliers. **h**, Heat maps showing examples of long-distance similarity in a real centromere (chromosome 1 of accession IP-Piq-0) and in a simulated sequence (simulation 1), with zoomed-in views of the highlighted regions. Long-distance similarity is indicated by solid boxes and arrows. All heat maps use a consistent colour scale representing sequence identity.

higher than the deletion rate, the simulated repeat arrays expanded continuously, rapidly reaching unrealistically large sizes (approximately 20 Mb after 150,000 generations) (Fig. 4d).

This suggested that the observed mutation spectrum lacks balancing events that constrain repeat array expansion, such as the very large deletions observed in the *rtell-1* mutant. Such events would probably be too rare to be observed in MA lines, and would therefore probably need to span hundreds of kilobases to have a substantial effect on array size. To identify such events, we searched for their footprints in 114 previously published genome assemblies from a global set of *Arabidopsis* accessions^{1,34}. As most centromere haplotypes are too divergent to reveal individual mutation events, we focused on accessions with similar centromere haplotypes, which in turn supported the detection of recent large-scale mutations. For example, an approximately 0.42-Mb region on chromosome 5 (13.80–14.22 Mb) in Etna-2 was absent from Tul-0, while the surrounding regions were structurally conserved, suggesting a large deletion in Tul-0 (Fig. 4b). Similarly, another approximately 0.64-Mb region (14.67–15.31 Mb) on the same chromosome was present in Etna-2 but absent in Valsi-1 (Fig. 4b and Extended Data Fig. 5b; see Extended Data Fig. 6 for additional examples). HiFi read alignments confirmed that both regions were truly absent from the respective genomes (Supplementary Fig. 17).

We combined the empirically estimated mutation spectrum with rare large-scale deletions, whose sizes were sampled from the patterns observed in natural centromeres. These deletions were introduced at a rate sufficient to stabilize centromere size over time (3.0×10^{-11} per bp per generation; mean size: 0.5 Mb; Methods), corresponding to approximately one event every 2,800 generations. The simulations were initialized with the five centromeric sequences of the *Arabidopsis* reference genome, and mutations were sampled in each generation (Fig. 4c). Each simulation was run for 150,000 generations, consistent with the estimated divergence time of the major *Arabidopsis* populations (120,000–90,000 years)³⁵.

Within the simulated centromeres, existing homogenized blocks expanded and contracted, with some disappearing and new ones emerging (Fig. 4e and Supplementary Videos). Some newly formed blocks eventually reached megabase sizes comparable to those observed in natural centromeres. To enable a quantitative comparison, we defined homogenized blocks on the basis of the patterns observed in natural centromeres. Using this definition, the simulated centromeres closely recapitulated the number and size distributions of homogenized blocks observed in natural centromeres (Fig. 4f,g and Methods).

These results show that kilobase-scale indels are sufficient to generate megabase-scale homogenized blocks. To maintain overall centromere sizes, additional megabase-scale deletions are required and introduce severe changes to the global structures of the simulated centromeres.

Unexpectedly, the simulations also recapitulated long-distance similarities between distant homogenized blocks that can also be observed in some natural centromeres^{5,36} (Fig. 4h and Extended Data Fig. 7). These patterns arose when repeat units with distinct sequence variation expanded within an existing homogenized block, thereby

splitting this block into two distinct parts. These two separated parts then shared greater similarity with each other than with the newly formed block between them. The simulated patterns emerged exclusively through kilobase-scale indels and point mutations and did not require long-distance recombination as previously proposed^{5,36}.

Together, the simulations showed how centromeric mutations drive the turnover and formation of homogenized blocks, whereas rare large-scale deletions are required to balance the unequal frequencies of insertions and deletions, thereby limiting the overall size of centromeres. The interplay of these processes is sufficient to generate centromeric repeat arrays that recapitulate key features of the homogenized blocks in natural centromeres, including block size and number.

Discussion

Here we generated virtually error-free genome assemblies of *A. thaliana* MA lines to investigate the mutational dynamics of centromeric tandem-repeat arrays. We found that centromeres exhibit an almost tenfold higher point mutation rate than chromosome arms, consistent with previous observations in human centromeres and pericentromeric regions in *Arabidopsis*^{11,21}.

The spectrum of centromeric mutations provided insights into the mechanism underlying rapid centromere sequence turnover. The clustering of point mutations strongly suggested that a substantial fraction of these mutations arose from NAGC events, which reshuffle sequence variants among neighbouring repeat units. Likewise, the frequent kilobase-scale indels almost exclusively preserved the tandem-repeat periodicity, further supporting a homology-dependent mechanism of centromeric mutagenesis. These mutational processes are consistent with homology-directed DNA repair, which may be promoted by the absence of meiotic crossovers^{37,38}, and provide a mechanistic basis for the concerted evolution of tandem-repeat arrays^{2,12–18}. The rare non-tandem-repeat-preserving deletion identified in the HPG1 lines, which accumulated mutations over around 400 years, mirrors the patterns observed in natural centromeres, where most repeats remain largely intact, apart from some rare truncated or partial repeats¹.

To further investigate the mechanisms underlying centromeric mutagenesis, we analysed mutation accumulation in *rtell-1* lines. Although the characteristic mutation spectrum remained unchanged, both NAGC events and tandem-repeat-preserving indels occurred at substantially higher frequencies than in wild-type plants. Because *RTTEL1* suppresses inappropriate or prolonged recombination^{39,40}, these findings provide genetic evidence that homology-directed repair has a dual role in centromeric repeat arrays by maintaining structural integrity while simultaneously generating sequence diversity.

These findings also provide insight into the mechanisms underlying centromere mutagenesis. Several features of the observed centromeric mutations are consistent with break-induced replication (BIR), a homology-directed repair pathway proposed to operate in centromeric regions^{41–44}. BIR readily explains the preservation of tandem-repeat periodicity, the excess of insertions over deletions, and the template switching that gives rise to complex tandem

duplications. Deletions might also arise through single-strand annealing between neighbouring repeat units⁴⁵, a pathway that is enhanced in the absence of RTEL1 (ref. 46).

Using forward-in-time simulations, we show that megabase-scale homogenized blocks can emerge through the accumulation of kilobase-scale mutations alone. This provides an example of emergence, whereby local mutational processes generate complex higher-order structures⁴⁷. Although repeat homogenization can arise from the observed mutation spectrum alone, this does not exclude the possibility that CENH3 binding or kinetochore formation influences where mutations occur^{1,48,49}. More broadly, these dynamics resemble concerted evolution of rDNA and other satellite DNA families^{38,49–52}, and are consistent with proposed feedback models of satellite repeat evolution⁵³, thereby providing a mechanistic basis for the emergence of such patterns.

The simulations further suggested that new centromeric repeat consensus sequences can evolve gradually through cumulative sequence turnover rather than through de novo formation of entire centromeric arrays^{2,16,17}. Because centromere function does not depend on a specific repeat sequence, these regions are likely to be subject to relatively weak sequence constraints. Together with suppressed meiotic recombination, this may facilitate the diversification and propagation of tandem-repeat arrays, ultimately giving rise to new centromeric consensus sequences.

Our conclusions are nevertheless constrained by the timescale of mutation accumulation experiments, which primarily capture frequent mutational events. Consequently, although comparisons among natural centromeres suggest that megabase-scale deletions contribute to centromere evolution, such rare events are unlikely to be captured in MA lines^{20,21,54}.

Our findings also leave important questions unresolved. Although local centromere-specific mutations shape centromeric repeat arrays, the chromosome-specific repeat consensus sequences remain highly similar (approximately 95% sequence identity) across the five *Arabidopsis* centromeres. This suggests occasional inter-chromosomal exchange of repeat units, potentially mediated by centrophilic TEs, inter-chromosomal NAGC events or population-level processes⁵⁵. However, despite analysing mutations accumulated over approximately 1,500 generations, we detected neither TE-associated mutations nor evidence of inter-chromosomal NAGC events.

In sum, our study defines the spectrum and frequency of centromeric mutations in *Arabidopsis* and shows how point mutations and tandem-repeat-preserving indels drive centromere evolution. Our findings provide a mechanistic framework for how local mutational processes generate, maintain and remodel centromeric repeat arrays, offering new insights into long-standing questions about centromere evolution.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-026-11046-w>.

- Włodzimierz, P. et al. Cycles of satellite and transposon evolution in *Arabidopsis* centromeres. *Nature* **618**, 557–565 (2023).
- Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
- Talbert, P. B. & Henikoff, S. What makes a centromere? *Exp. Cell. Res.* **389**, 111895 (2020).
- Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
- Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, eabi7489 (2021).
- 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).

- Naish, M. & Henderson, I. R. The structure, function, and evolution of plant centromeres. *Genome Res.* **34**, 161–178 (2024).
- Wang, B. et al. High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. *Genomics Proteomics Bioinformatics* **20**, 4–13 (2022).
- 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).
- Hou, X., Wang, D., Cheng, Z., Wang, Y. & Jiao, Y. A near-complete assembly of an *Arabidopsis thaliana* genome. *Mol. Plant* **15**, 1247–1250 (2022).
- Logsdon, G. A. et al. The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
- Dover, G. M. Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–117 (1982).
- Smith, G. P. Unequal crossover and the evolution of multigene families. In *Cold Spring Harbor Symposia on Quantitative Biology* Vol. 38 507–513 (Cold Spring Harbor Laboratory Press, 1974).
- Dvořák, J., Jue, D. & Lassner, M. Homogenization of tandemly repeated nucleotide sequences by distance-dependent nucleotide sequence conversion. *Genetics* **116**, 487–498 (1987).
- Gangloff, S., Zou, H. & Rothstein, R. Gene conversion plays the major role in controlling the stability of large tandem repeats in yeast. *EMBO J.* **15**, 1715–1725 (1996).
- Hall, S. E., Kettler, G. & Preuss, D. Centromere satellites from *Arabidopsis* populations: maintenance of conserved and variable domains. *Genome Res.* **13**, 195–205 (2003).
- 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).
- Ugarković, Đ. & Plohl, M. Variation in satellite DNA profiles—causes and effects. *EMBO J.* **21**, 5955–5959 (2002).
- Fultz, D., McKinlay, A., Enganti, R. & Pikaard, C. S. Sequence and epigenetic landscapes of active and silent nucleolus organizer regions in *Arabidopsis*. *Sci. Adv.* **9**, ead4509 (2023).
- Ossowski, S. et al. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
- Weng, M.-L. et al. Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics* **211**, 703–714 (2019).
- Marriage, T. N. et al. Direct estimation of the mutation rate at dinucleotide microsatellite loci in *Arabidopsis thaliana* (Brassicaceae). *Heredity* **103**, 310–317 (2009).
- Monroe, J. G. et al. Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* **602**, 101–105 (2022).
- Wijner, E. et al. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *eLife* **2**, e01426 (2013).
- Exposito-Alonso, M. et al. The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.* **14**, e1007155 (2018).
- Shirsek, G. et al. Multiple sources of introduction of North American *Arabidopsis thaliana* from across Eurasia. *Mol. Biol. Evol.* **38**, 5328–5344 (2021).
- Platt, A. et al. The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000843 (2010).
- Stuart, T. et al. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife* **5**, e20777 (2016).
- Tsukahara, S. et al. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* **461**, 423–426 (2009).
- Rodgers, K. & McVey, M. Error-prone repair of DNA double-strand breaks. *J. Cell. Physiol.* **231**, 15–24 (2016).
- Hourvitz, N., Awad, A. & Tzfati, Y. The many faces of the helicase RTEL1 at telomeres and beyond. *Trends Cell Biol.* **34**, 109–121 (2024).
- Recker, J., Knoll, A. & Puchta, H. The *Arabidopsis thaliana* homolog of the helicase RTEL1 plays multiple roles in preserving genome stability. *Plant Cell* **26**, 4889–4902 (2014).
- Röhrig, S., Schröpfer, S., Knoll, A. & Puchta, H. The RTR complex partner RMI2 and the DNA helicase RTEL1 are both independently involved in preserving the stability of 45S rDNA repeats in *Arabidopsis thaliana*. *PLoS Genet.* **12**, e1006394 (2016).
- Lian, Q. et al. A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. *Nat. Genet.* **56**, 982–991 (2024).
- Durvasula, A. et al. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **114**, 5213–5218 (2017).
- Wang, L. et al. A telomere-to-telomere gap-free assembly of soybean genome. *Mol. Plant* **16**, 1711–1714 (2023).
- Talbert, P. B. & Henikoff, S. Centromeres convert but don't cross. *PLoS Biol.* **8**, e1000326 (2010).
- Shi, J. et al. Widespread gene conversion in centromere cores. *PLoS Biol.* **8**, e1000327 (2010).
- Villeneuve, A. M. Ensuring an exit strategy: RTEL1 restricts rogue recombination. *Cell* **135**, 213–215 (2008).
- Goldkühle, L., Lechner, L., Capdeville, N., Dorn, A. & Puchta, H. The DNA helicase RTEL1 is involved in the repair of replicative DNA damage independently of the alternative end joining and the DNA–protein cross-link repair pathways in *Arabidopsis*. *Plant J.* **126**, e70903 (2026).
- Showman, S., Talbert, P. B., Xu, Y., Adeyemi, R. O. & Henikoff, S. Expansion of human centromeric arrays in cells undergoing break-induced replication. *Cell Rep.* **43**, 113851 (2024).
- Llorente, B., Smith, C. E. & Symington, L. S. Break-induced replication: what is it and what is it for? *Cell Cycle* **7**, 859–864 (2008).
- Smith, C. E., Llorente, B. & Symington, L. S. Template switching during break-induced replication. *Nature* **447**, 102–105 (2007).
- Rice, W. Why do centromeres evolve so fast: BIR replication, hypermutation, transposition, and molecular-drive. Preprint at <https://doi.org/10.20944/preprints202012.0669.v1> (2020).

45. Siebert, R. & Puchta, H. Efficient repair of genomic double-strand breaks by homologous recombination between directly repeated sequences in the plant genome. *Plant Cell* **14**, 1121–1131 (2002).
46. Dvořák Tomašíková, E. et al. The interplay of homology-directed repair pathways in the repair of zebularine-induced DNA–protein crosslinks in *Arabidopsis*. *Plant J.* **119**, 1418–1432 (2024).
47. Goldstein, J. Emergence as a construct: history and issues. *Emergence* **1**, 49–72 (1999).
48. 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).
49. Rudd, M. K., Wray, G. A. & Willard, H. F. The evolutionary dynamics of α -satellite. *Genome Res.* **16**, 88–96 (2006).
50. Wolfgruber, T. K. et al. High quality maize centromere 10 sequence reveals evidence of frequent recombination events. *Front. Plant Sci.* **7**, 308 (2016).
51. Durfy, S. J. & Willard, H. F. Concerted evolution of primate alpha satellite DNA. *J. Mol. Biol.* **216**, 555–566 (1990).
52. Chatterjee, B. & Lo, C. W. Chromosomal recombination and breakage associated with instability in mouse centromeric satellite DNA. *J. Mol. Biol.* **210**, 303–312 (1989).
53. Nijman, I. J. & Lenstra, J. A. Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeats. *J. Mol. Evol.* **52**, 361–371 (2001).
54. Lu, Z. et al. Genome-wide DNA mutations in *Arabidopsis* plants after multigenerational exposure to high temperatures. *Genome Biol.* **22**, 160 (2021).
55. Birchler, J. A. & Presting, G. G. Retrotransposon insertion targeting: a mechanism for homogenization of centromere sequences on nonhomologous chromosomes. *Genes Dev.* **26**, 638–640 (2012).
56. Maheshwari, S., Ishii, T., Brown, C. T., Houben, A. & Comai, L. Centromere location in *Arabidopsis* is unaltered by extreme divergence in CENH3 protein sequence. *Genome Res.* **27**, 471–478 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

Methods

Plant material

Four sets of *A. thaliana* materials were analysed in this study, including MA16 and MA32 lines, HPG1 accessions, and *rtell-1* mutants (Supplementary Table 1). The MA16 lines (two lines, samples A and B) were derived from a trans-generational mutation accumulation experiment, originating from a single Col-0 mother plant (Nottingham *Arabidopsis* Stock Centre ID N1092) and propagated independently for 16 generations by self-pollination and single-seed descent (SSD). The MA32 dataset comprised eight lines from a previous mutation accumulation experiment⁵⁷ that were propagated for 32 generations under a similar SSD scheme. The HPG1 dataset included five natural accessions representing natural populations that diverged approximately 400 years ago^{25–27} and were used to investigate recent centromeric TE activity. The *rtell-1* mutant (three samples; SALK_113285) was propagated for seven generations by SSD to assess the role of homology-directed repair in centromere evolution.

DNA extraction and PacBio sequencing

For MA16 samples A and B, HMW DNA was extracted from 1.5 g of pooled vegetative tissue using the NucleoBond HMW DNA kit (Macherey-Nagel). DNA quality was assessed using a FEMTOpulse system (Agilent), and concentration was measured with a Quantus fluorometer (Promega). HiFi SMRTbell libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (PacBio), including fragmentation with g-TUBEs (Covaris) and size selection using SageELF (Sage Science). Libraries were sequenced on the PacBio Sequel II platform at the Max Planck Genome Centre (MP-GC), Cologne, Germany. In addition, PCR-free Illumina paired-end libraries were prepared from independently extracted DNA (Macherey-Nagel DNA Maxi kit) and sequenced by Novogene.

For the eight MA32 lines, HMW DNA was extracted at the Max Planck Institute for Biology Tübingen using a modified protocol⁹, including β -mercaptoethanol during lysis and a phenol purification step. DNA was further purified using two rounds of bead cleanup (SeraMag Speed-Beads and AMPure PB beads). Libraries were prepared using the SMRTbell prep kit 3.0. HiFi sequencing was performed on the PacBio Revio device at MP-GC.

For the five HPG1 accessions, HMW DNA extraction and library preparation were performed at the Max Planck Institute for Biology Tübingen using the same protocol as for MA32. Libraries were prepared with the HiFi SMRTbell Express Template Prep Kit 2.0 (PacBio). Libraries were size-selected using the BluePippin system (Sage Science) and sequenced on a Sequel II system with Binding Kit 2.2 at the Max Planck Institute for Biology Tübingen.

For the three *rtell-1* samples, HMW DNA was extracted using a kit-based protocol at KIT (Karlsruhe Institute of Technology), followed by HiFi library preparation using the SMRTbell prep kit 3.0 and libraries sized with BluePippin (Sage Science). Sequencing was performed on the PacBio Revio device at MP-GC. In addition, PCR-free Illumina paired-end sequencing was carried out at the Institute of Clinical Molecular Biology (IKMB), Kiel.

Nanopore sequencing and basecalling

For the two MA16 lines, the eight additional MA32 lines, and the three *rtell-1* samples, library preparation was performed using the Ligation Sequencing gDNA–Native Barcoding Kit 24 V14 (SQK-NBD114.24, Oxford Nanopore Technologies). The resulting libraries were loaded onto FLO-PRO114M flow cells, and sequencing was conducted on a PromethION 2 Solo platform.

ONT sequencing data were basecalled with Dorado v1.1.1 (<https://github.com/nanoporetech/dorado/>) using the sup model for high accuracy basecalling, including modified base detection (5mC and 5hmC) and move table output. Barcode demultiplexing was guided by

the SQK-NBD114-24 kit and a sample sheet. Resulting BAM files were split by barcode using samtools⁵⁸ split v1.17 for downstream analysis.

Genome assembly and scaffolding

We tested multiple assemblers, including Hifiasm⁵⁹ v0.16.0 (r369), Hicanu⁶⁰ v2.2, ipa v1.0.5 (<https://github.com/PacificBiosciences/pbipa>), peregrine v1.6.3 + 3.g008082a.dirty (<https://github.com/cschin/peregrine>) to assemble the A1 genome and found that Hifiasm produced the most contiguous assembly. We used Hifiasm with the parameter “-l0” for all four generation-16 samples (A1, A2, B1 and B2), eight generation-32 MA lines and the *rtell-1* mutants. Organellar contigs were then identified on the basis of sequence alignment to TAIR10 (ref. 61) mitochondria and chloroplast reference sequences (GCF_000001735.4), retaining those with $\geq 80\%$ identity and coverage. Non-organellar contigs were scaffolded into pseudo-chromosomes using RagTag⁶² v1.0.1, on the basis of alignment to the Col-CEN⁵ reference genome. We further evaluated polishing strategies for HiFi-based assemblies using the MA16 lines. Polishing introduced over-corrections and led to an increase in assembly errors (Supplementary Results). Therefore, no polishing was applied to HiFi-based assemblies in subsequent analyses. Additionally, genome assemblies for five HPG1 samples were generated using Hifiasm v0.16.1-r375 and scaffolded with RagTag⁶² v2.0.1 (scaffold-q60-f30000-l0.5-remove-small), excluding contigs <100 kb.

To complement the HiFi-based assemblies, ONT long-read assemblies were also generated using Hifiasm⁶³ v0.25.0 with the parameters –ont -l0-r1-cut 10000–sc-cut 15, which restricts assembly to high-quality reads ≥ 10 kb with estimated quality value (QV) ≥ 15 . The ONT contigs were filtered to remove organellar sequences and further polished. ONT reads were first aligned to the contig-level assemblies using the Dorado aligner, followed by sorting and indexing with samtools⁵⁸ v1.19.2. Coverage profiles were computed using mosdepth⁶⁴ v0.3.1, and regions with abnormal coverage were filtered prior to polishing using a custom script. To evaluate polishing strategies, two approaches were tested for the MA16 lines: polishing with move table information (‘with moves’) and without move table information. Comparative assessment showed that polishing with move table information resulted in fewer assembly errors. Therefore, all subsequent polishing of ONT-based assemblies was performed using the move-aware Dorado polishing mode with region filtering and GPU acceleration.

The polished contigs were then scaffolded with RagTag⁶² using the Col-CEN⁵ reference, following the same strategy as for the HiFi assemblies.

Assembly evaluation

We computed Benchmarking Universal Single-Copy Ortholog (BUSCO) scores using BUSCO⁶⁵ v5.2.2 with the parameters “-l embryo-phyte_odb10 -m genome.” Additionally, we assessed consensus quality and completeness using Merquy⁶⁶ v1.3 by comparing *k*-mers in the de novo assemblies with those from Illumina short reads. *k*-mer databases (*k* = 18) were generated for each Illumina paired-end read set using Meryl⁶⁶ v1.3 and then merged with Meryl’s union-sum function. Merquy⁶⁶ was subsequently applied to each assembly to obtain genome-wide consensus quality values and completeness scores.

Repeat annotation

Following the approach of Rabanal et al.⁹, we used RepeatMasker v4.0.9 (<http://www.repeatmasker.org>) with a custom library (-lib rDNA_Naish-CEN_telomeres.fa-nolow-gff-xsmall-cutoff 200) to annotate 5S rDNA, 45S rDNA, and telomere sequences. Mitochondrial insertions on chromosome 2 were identified by aligning to the TAIR10 (ref. 61) mitochondrial sequence using minimap2 (ref. 67 v2.24-r1122). Centromeres were annotated using TRASH⁶⁸ v1.2 (–seqt CEN178.csv–horclass CEN178–par 5), and the HOR score of each centromeric repeat unit was calculated. For samples A and B, we further annotated simple sequence repeats with a custom Python script to identify mono-, di-, tri- and heptanucleotide

repeats. Bedtools⁶⁹ v2.29.0 intersect was used to compare assembly errors and mutations across different repeat types.

We further identified the CENH3 enrichment regions. Raw paired-end CENH3 ChIP-seq (SRR4430537) and corresponding input reads (SRR4430555)⁵⁶ were first adapter-trimmed and quality-filtered using Cutadapt⁷⁰ v5.1 and aligned to the reference genome using Bowtie2 (ref. 71) v2.5.1 with sensitive parameters “–very-sensitive–no-mixed–no-discordant”. Alignments were sorted and indexed using Samtools⁵⁸ v1.19.2, and genome-wide coverage tracks were generated using deepTools⁷² v3.5.6 with BPM normalization. Enrichment of ChIP signal over input was calculated as log₂ ratios using bigwigCompare.

Identification of assembly errors

To identify assembly discrepancies between replicate genome assemblies, we performed pairwise alignments of the six assemblies (A1–A3, B1–B3) using minimap2 (ref. 67) v2.24-r1122 (-ax asm5–eqx). Assembly differences were identified with SYRI⁷³ v1.0. To determine which replicate contained the assembly error, we mapped both HiFi and Illumina reads to the reference and their corresponding assemblies using minimap2 (ref. 67) v2.24-r1122 (-ax map–hifi) and BWA-MEM⁷⁴ v0.7.17-r1188, respectively. Alignments were sorted and converted to BAM files with Samtools⁵⁸ v1.19.2. Additionally, HiFi reads from each replicate were aligned to the opposing replicate’s assembly for cross-comparison. For regions flagged by SYRI⁷³, we examined alignments in IGV⁷⁵ v2.13.0. If HiFi and Illumina data aligned cleanly to one assembly but showed mismatches in the other, the error was attributed to the mismatching assembly, and the error-free one was considered correct. This approach allowed us to systematically identify and verify which sample contained assembly errors. It is worth noting that regions near GA repeats exhibited incomplete assemblies due to reduced HiFi read coverage⁹. These were classified as incomplete rather than erroneous assemblies, as the low-depth pattern was consistent across samples.

Identification and validation of mutations

To identify mutations for each of the four groups (MA16, MA32, HPG1 and *rtel1-1*), we first selected the most contiguous assembly and then aligned each of the other assemblies to it with minimap2 (ref. 67) v2.24-r1122. Alignments were sorted with Samtools⁵⁸, and sequence differences were detected using SYRI⁷³. To validate candidate mutations and determine their zygosity, HiFi, ONT, and Illumina reads were aligned to both the corresponding sample assembly and the reference assembly (long reads using minimap2 v2.24-r1122 and short reads using BWA-MEM⁷⁴ v0.7.17-r1188). Mutations were visually assessed in IGV⁷⁵ to confirm their presence and to assess whether they were homozygous or heterozygous on the basis of read alignment. To distinguish mutated from wild-type alleles, we referenced the TAIR10 (ref. 61) and the Col-CEN⁵ assemblies, assigning the allele matching the reference as wild-type and the differing one as a mutation. Variants shared by at least two individuals within a group were considered likely derived from segregation of pre-existing heterozygous variants in the parental material and were excluded from downstream analysis.

For MA16 samples, mutations were validated genome-wide, whereas for the MA32 lines, HPG1 samples, and *rtel1-1* mutants, analyses focused primarily on centromeric regions (defined as the interval spanning the outermost *CEN178* repeats and *ATHILA* elements), as well as single-nucleotide variants located on chromosome arms. Clusters of mutations were defined as groups of variants with consistent zygosity located within 1 kb of each other; variants separated by ≤1 kb were considered part of the same mutation cluster.

To ensure accurate coordinate mapping and facilitate identification of potential NAGC donor sequences, ancestral (F0) reference assemblies were reconstructed. Specifically, positions identified as mutations relative to the reference but shared across other samples were reverted to the wild-type allele using the consensus function implemented in bcftools⁵⁸ v1.16. All alignments and mutation validations

were subsequently repeated against these reconstructed reference assemblies.

Curation of mutation calls

Due to the highly repetitive nature of centromeric sequences, alignment errors can fragment a single large variant into multiple smaller mutation calls. To further correct for alignment artefacts and refine mutation calls, we developed a word-based alignment approach. For each candidate mutation, the mutated sequence and the corresponding reference sequence, each including 10 kb flanking regions, were extracted using the subseq function in seqtk v1.4-r122 (<https://github.com/lh3/seqtk>). Exact matches were identified using a custom Python script based on *k*-mer indexing (*k* = 150, no mismatches), using long exact matches as anchors between sequences in both forward and reverse-complement orientations. This approach identifies long, uninterrupted matching segments as high-confidence anchors, avoiding mismatches and gaps that can confound conventional alignment algorithms in repetitive regions. On the basis of these anchors, mutations were redefined using a left-alignment strategy to obtain the most parsimonious representation of each variant. For precise characterization of large indels, the word-based alignments were further visualized using dot plots generated by a custom Python script. Compared to standard alignment methods, this mismatch-free approach enables more accurate delineation of mutation boundaries and reduces false-positive fragmentation of variants in centromeric regions.

Assessment of repeat periodicity in centromeric indels

To assess whether large centromeric indel mutations preserve the periodic structure of centromeric repeat units, indel sequences were aligned to the *CEN178* consensus sequence using MUMmer⁷⁶ v4.0.0beta2 (nucmer, –maxmatch -l 10 -c 20). Alignments were converted from .delta to .coords format using show-coords (-c), and dot plots were generated using the R script dotPlotly (<https://github.com/tpoorten/dotPlotly>, 2023-11-13) (-m 10 -q 10 -k 5 -l -x).

Statistical assessment of mutation distribution

The curated mutation counts for each mutation type across samples and across chromosomes were analysed using chi-square goodness-of-fit tests to assess deviations from uniformity. For chromosome-level analyses, expected counts were scaled by centromeric sequence size to account for differences in mutation opportunity. *P* values were adjusted using the Benjamini–Hochberg false discovery rate method. Due to low counts and violation of test assumptions, insertions and deletions were combined into a single ‘indel’ category to improve statistical power. Mutation distributions along chromosomes were visualized using the R package karyoploteR⁷⁷ v1.22.0.

Identification of candidate donor sequences for NAGC

To identify potential donor sequences underlying NAGC events, we performed targeted searches within centromeric repeat arrays using mutation-containing sequences as queries.

For mutation clusters, the full mutated sequence spanning each cluster was used as the query to identify exact matches within centromeric repeats. Candidate donor sites were required to map to the same relative position within the repeat consensus as the mutation cluster. For singleton point mutations, all possible 20-bp sequences spanning each mutation site were extracted and used as queries, applying the same positional constraint within the repeat consensus.

When multiple candidate donors were identified, the repeat unit with the shortest distance to the mutation site (measured in repeat units) was selected as the most likely donor. In cases in which only a single candidate donor was identified, it was always located on the same chromosome as the mutation. No instances were observed in which candidate donors were exclusively located on different chromosomes, suggesting that inter-chromosomal NAGC events are rare.

Identification of mutations associated with novel TE insertions

TEs and *ATHILA* elements were annotated in both the reference and query genomes using EDTA⁷⁸ v2.2.2 and ATHILAFinder⁷⁹ v1.0, respectively. We used a TE presence/absence-based strategy to identify structural variants potentially associated with novel TE insertions.

Mutations of at least 50 bp between the reference and query genomes were identified using SyRI⁷³ as described above. The coordinates of these structural variants in both genomes were intersected with the corresponding intact TE annotations using bedtools⁶⁹ intersect. A structural variant was considered to be associated with a novel TE insertion only if it overlapped a TE annotation in the query genome but showed no overlap with any TE annotation at the corresponding location in the reference genome.

To further validate the integrity and conservation of the identified *ATHILA* elements and to exclude the possibility of misidentification, we performed a phylogenetic reconstruction. For each identified *ATHILA* locus, sequences from five HPG1 samples were extracted and aligned using MAFFT⁸⁰ v7.407 with the `-auto` parameter. The resulting multiple sequence alignments were used to infer maximum-likelihood phylogenetic trees using FastTree⁸¹ v2.1.11 with the `-nt` (nucleotide) model. The resulting phylogenetic trees were visualized, annotated, and refined using iTOL⁸² (Interactive Tree Of Life, v7). The robust clustering pattern demonstrated that these *ATHILA* elements are orthologous across the HPG1 samples and have been maintained with high sequence integrity. These results provided compelling evidence that these elements were inherited from the ancestral genome and that no novel *ATHILA* insertion events or associated large-scale structural mutations have occurred at these loci across the sampled accessions.

Mutation rate calculations

Mutation rates were calculated as mutations per nucleotide per generation using the formula:

$$\mu = \frac{m}{N \times g}$$

where m is the total observed mutations, N is genome size (in nucleotides) and g is number of generations. Mutations from all eight MA32 lines were pooled to calculate the overall mutation rate. Total mutations were divided by the cumulative nucleotide-generations ($\sum[N \times g] = 3.032 \times 10^9$), yielding:

$$\mu = 6.50 \times 10^{-8} \text{ (95\% confidence interval: } 5.62 \times 10^{-8} \text{ to } 7.47 \times 10^{-8})$$

Confidence intervals (95%) were derived from Poisson statistics using the chi-square approximation:

$$\lambda_{\text{lower/upper}} = \frac{1}{2} \chi_{\alpha/2, 2m}^2 \text{ with } \alpha = 0.05.$$

To distinguish spontaneous point mutations from those arising through NAGC, we decomposed the observed Ti/Tv ratios. Clustered mutations, which are characteristic of NAGC, exhibited a Ti/Tv ratio of 0.78, whereas mutations on chromosome arms (representing spontaneous mutations) showed a Ti/Tv ratio of 2.24. The overall Ti/Tv ratio of centromeric point mutations was 1.22. On the basis of previous studies suggesting that the Ti/Tv ratio of spontaneous mutations is similar between centromeres and chromosome arms²¹, we decomposed the observed overall Ti/Tv (1.22) into a linear combination of spontaneous mutations and NAGC-derived mutations. We estimated that 30.3% of point mutations arose from spontaneous mutations and 69.7% from NAGC. This yielded a spontaneous point mutation rate of 2.0×10^{-8} per bp per generation and a NAGC-associated point mutation rate of 4.5×10^{-8} per site per generation.

Analysis of natural centromeres

To analyse the sequence features and structural patterns of natural centromeres, we used 114 publicly available high-quality HiFi-based genome assemblies of *A. thaliana*, including 66 from Wlodzimierz et al.¹ and 48 from Lian et al.³⁴. Centromeric sequences were identified using both CentroAnno⁸³ v1.0.2 and TRASH⁶⁸ v1.2. Downstream analyses were based mainly on CentroAnno annotations, owing to its higher computational efficiency and scalability for large datasets. TRASH⁶⁸ was additionally applied to ensure consistency with analyses performed on MA lines and to validate the robustness of centromere annotations across methods, which showed overall concordant results. Using the subseq function in seqtk v1.4-r122 (<https://github.com/lh3/seqtk>), we extracted the genomic regions spanning the two outermost centromeric repeats for each chromosome. A total of 570 centromeric sequences (114 accessions $\times$ 5 chromosomes) were analysed. Sequence identity was quantified using ModDotPlot⁸⁴ v0.9.8 for both self-comparisons and pairwise comparisons (restricted to the same chromosome across accessions), with parameters `“-w10000 -d0”`. It partitioned genomic sequences into non-overlapping windows of 10 kb and calculated similarity between all window pairs. Finally, sequence identity matrices were visualized as heat maps using custom Python scripts.

To validate putative large deletions identified from pairwise comparisons between natural centromeres, we first inspected the corresponding regions in the genome assemblies and confirmed that they did not overlap assembly gaps. HiFi sequencing data for the relevant accessions were obtained from the NCBI Sequence Read Archive (SRA) and converted to FASTQ format using faster-dump from SRA Toolkit v3.2.1 (<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>). Reads were aligned to both their native assemblies and the assemblies of related accessions using minimap2 (ref. 67) v2.24-r1122 with the `“-ax map-hifi”` option. Alignments were inspected in IGV⁷⁵ v2.13.0. Reads with mapping quality of <2 were excluded from visualization. Coverage profiles and read alignments were examined for both native and cross-accession mappings, including the putatively deleted regions.

Formal definition of homogenized blocks

To formally describe the appearance of homogenized blocks, we developed a definition based on the pairwise similarity matrix from ModDotPlot⁸⁴ v0.9.8. Homogenized blocks were defined as continuous genomic regions spanning at least five windows (≥ 50 kb), with a minimum average pairwise similarity of 97%. Self-comparisons and immediately adjacent windows were excluded from the calculation to avoid inflated similarity. For each region, the average similarity was computed across all valid window pairs. Nested blocks (overlapping or fully contained regions representing redundant detection) were subsequently filtered to retain only the largest non-redundant regions. Blocks were further filtered on the basis of their overlap with centromeric repeat annotations, requiring at least 90% of each block to be annotated as centromeric repeat sequence, while blocks not meeting this criterion were excluded. For visualization, similarity matrices were plotted as heat maps using custom colour palettes, and detected blocks were overlaid as rectangular annotations.

Forward-in-time simulation of centromere evolution

We estimated mutation rates on the basis of 270 homozygous mutations identified across eight MA32 samples, including 208 point mutations, 56 large indels, and 6 small indels. This corresponded to mutation rates of 6.5×10^{-8} per bp per generation for point mutations, 1.2×10^{-8} for large insertions (mean size 1,870 bp; 10.51 repeat units) and 6.9×10^{-9} for large deletions (mean size 923.1 bp; 5.19 repeat units).

To assess whether the observed mutation spectrum of clustered centromeric point mutations can be explained by NAGC and to exclude the contribution of GC-biased processes, we performed NAGC-only

Article

simulations under the same parameter framework. Five centromeric sequences were independently simulated, each with three replicates. In each replicate, NAGC events were iteratively introduced, and mutation spectra were calculated from the resulting sequences and compared to the observed clustered mutation spectrum. In addition, simulations were performed with varying recipient unit distances (adjacent, and separated by 1, 9 or 99 intervening repeat units) to evaluate the effect of spatial separation on mutation patterns. This analysis allowed us to determine whether NAGC alone can recapitulate the observed mutation spectrum without invoking additional mutational biases such as GC-biased gene conversion.

To estimate the frequency of NAGC events, we simulated gene conversion on the reference genome. Parameter exploration indicated that tract length primarily determines the number of introduced variants, whereas donor distance has minimal effect (Extended Data Fig. 4 and Supplementary Fig. 18). On the basis of observed mutation clusters, NAGC tracts were modelled using a geometric distribution (mean = 0.05), with donors restricted to adjacent repeat units. Across 1,000 simulations per chromosome (five centromeres total), 6,684 point mutations were introduced. As each NAGC event introduced an average of 1.34 point mutations, the estimated NAGC-associated point mutation rate of 4.5×10^{-8} per bp per generation corresponds to an estimated NAGC rate of 3.4×10^{-8} per bp per generation.

Using these estimated rates, we simulated centromeric repeat evolution on the five Col-0 centromeric repeats, incorporating spontaneous point mutations, NAGC events, and large insertions and deletions (schematic simulation pipeline in Supplementary Fig. 19). Each centromere was simulated independently for 150,000 generations with 20 replicates. These simulations revealed exponential array expansion due to a bias toward more frequent and larger insertions (Fig. 4d), resulting in unrealistically large arrays (>20 Mb).

To balance array size, we introduced rare but large deletions, motivated by patterns observed across natural centromeres. Large deletions were modelled with an average size of 500 kb (2,809 repeat units) and a rate of 3.04×10^{-11} per bp per generation (to balance the array size). With this parameterization, each centromere was simulated for 100 replicates over 150,000 generations.

To visualize array dynamics, simulation outputs were plotted every 1,000 generations and compiled into videos using FFmpeg⁸⁵ v7.0.1 (10 frames per second, libx264 encoding). To compare simulated and natural centromeres, self- and pairwise similarity analyses were performed using ModDotPlot⁸⁴ v0.9.8, and homogenized blocks were identified from self-similarity matrices as described above.

Reporting summary

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

Data availability

PacBio HiFi reads for A1, previously used for the first assembly of the *Arabidopsis* centromeres, are available through the European Nucleotide Archive (ENA) under BioProject PRJEB46164. CENH3 ChIP-seq data (SRR4430537) and corresponding input data (SRR4430555) were retrieved from the NCBI Sequence Read Archive. The TAIR10 mitochondrial and chloroplast reference sequences were obtained from the NCBI RefSeq assembly GCF_000001735.4. Previously published genome assemblies^{1,34} analysed in this study are publicly available through the ENA under BioProjects PRJEB55353, PRJEB55632, PRJEB50694 and PRJEB51511, and through NCBI under BioProject PRJNA1033522. To ensure comprehensive data access and support further analysis, we have released the complete set of HiFi, ONT and Illumina reads used in this study. Data related to the replicated MA16 genome assemblies are available through NCBI under BioProject PRJNA1259971; the data for the eight MA32 lines are available under PRJEB112076; and the data for

three *rtel1-1* lines are available under PRJNA1458610. Genome assemblies and sequencing data for the five HPG1 accessions are available through ENA under BioProject PRJEB75768. This BioProject includes the three HPG1 accessions analysed previously⁸⁶, as well as the two additional accessions analysed in this study. The corresponding replicated HiFi and ONT genome assemblies for the MA16, MA32 and *rtel1-1* samples are available on Figshare (<https://doi.org/10.6084/m9.figshare.32135497> (ref. 87); see Supplementary Table 1 for accession numbers).

Code availability

Scripts used in this article are available at <https://github.com/schneebergerlab/replicated-assemblies-centromere-study>.

57. Shaw, R. G., Byers, D. L. & Darnø, E. Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. *Genetics* **155**, 369–378 (2000).
58. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
59. 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).
60. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
61. Lamesch, P. et al. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
62. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
63. Cheng, H. et al. Efficient near-telomere-to-telomere assembly of nanopore simplex reads. *Nature* <https://doi.org/10.1038/s41586-026-10105-6> (2026).
64. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
65. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
66. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
67. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
68. Włodzimierz, P., Hong, M. & Henderson, I. R. TRASH: Tandem Repeat Annotation and Structural Hierarchy. *Bioinformatics* **39**, btad308 (2023).
69. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
70. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
71. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
72. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160 (2016).
73. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyR1: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
74. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://doi.org/10.48550/arxiv.1303.3997> (2013).
75. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
76. Marçais, G. et al. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
77. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
78. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
79. Bousios, A. & Primetis, E. ATHILAfinder: a tool to detect ATHILA LTR retrotransposons in plant genomes. Preprint at *bioRxiv* <https://doi.org/10.64898/2026.03.20.713144> (2026).
80. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
81. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
82. 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).
83. Qi, J. et al. De novo annotation of centromere with centroAnno. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.02.19.639205> (2025).
84. Sweeten, A. P., Schatz, M. C. & Phillippy, A. M. ModDotPlot—rapid and interactive visualization of tandem repeats. *Bioinformatics* **40**, btac493 (2024).
85. Tomar, S. Converting video formats with FFmpeg. *Linux J.* **2006**, 10 (2006).
86. Tao, Y. et al. Atlas of telomeric repeat diversity in *Arabidopsis thaliana*. *Genome Biol.* **25**, 244 (2024).
87. Dong, X. & Schneeberger, K. The mutational dynamics of the *Arabidopsis* centromeres. *Figshare* <https://doi.org/10.6084/m9.figshare.32135497> (2026).

Acknowledgements We thank C. Dent and L. Rauschnig for helpful discussions, K. Fritsch for technical help and P. Ehrenstein for introducing us to the concept of emergence.

Author contributions X.D., H.P., D.W. and K.S. developed and supervised the project. L.G., F.R., S.A., J.A.C., Y.T., J.T., L.M.S. and B.H. performed plant work and/or generated data. X.D., F.R. and Y.T. assembled the genomes. X.D. performed the data analysis with help from W.-B.J., F.R. and M.T.P. X.D. and K.S. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Funding This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC 2048/1–390686111 (K.S.), the Transregio Collaborative Research Center TRR356/1 2023 'Genetic diversity shaping biotic interactions of plants' (491090170) (D.W. and K.S.), and the DFG project 'Assessing the effects of DNA repair and homologous recombination pathways on genome integrity in plants' (458717077) (H.P. and K.S.). In addition, the work was funded by the Max Planck Society (D.W.), the European Research Council

(ERC) grants BYTE2BITE (101124694; K.S.) and Prime-A-Plant (309944) (J.T.) and the BBSRC (grant BB/W015250/1) (J.T. and L.M.S.). Open access funding provided by Max Planck Society.

Competing interests The authors declare no competing interests.

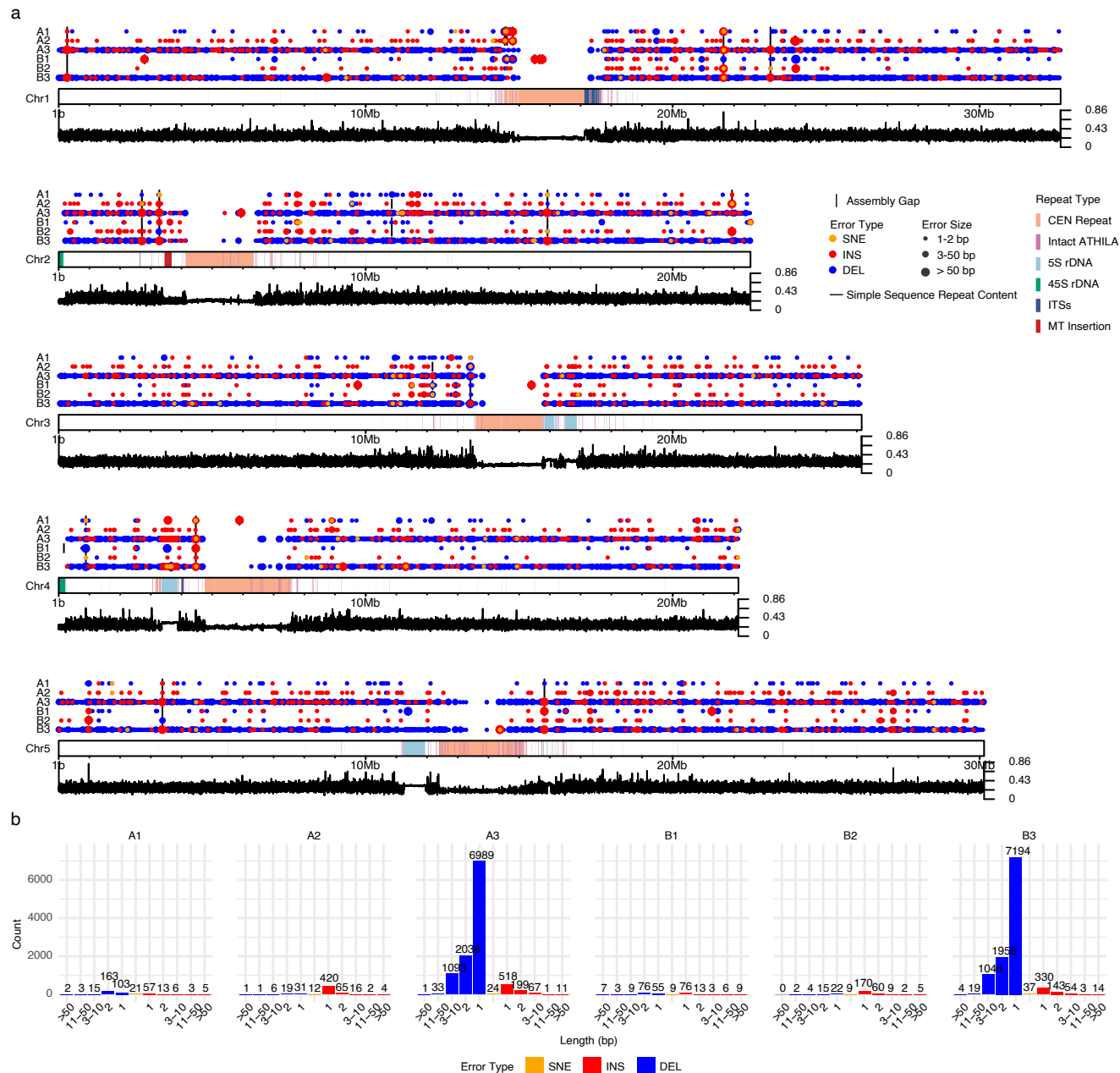
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-026-11046-w>.

Correspondence and requests for materials should be addressed to Korbinian Schneeberger.

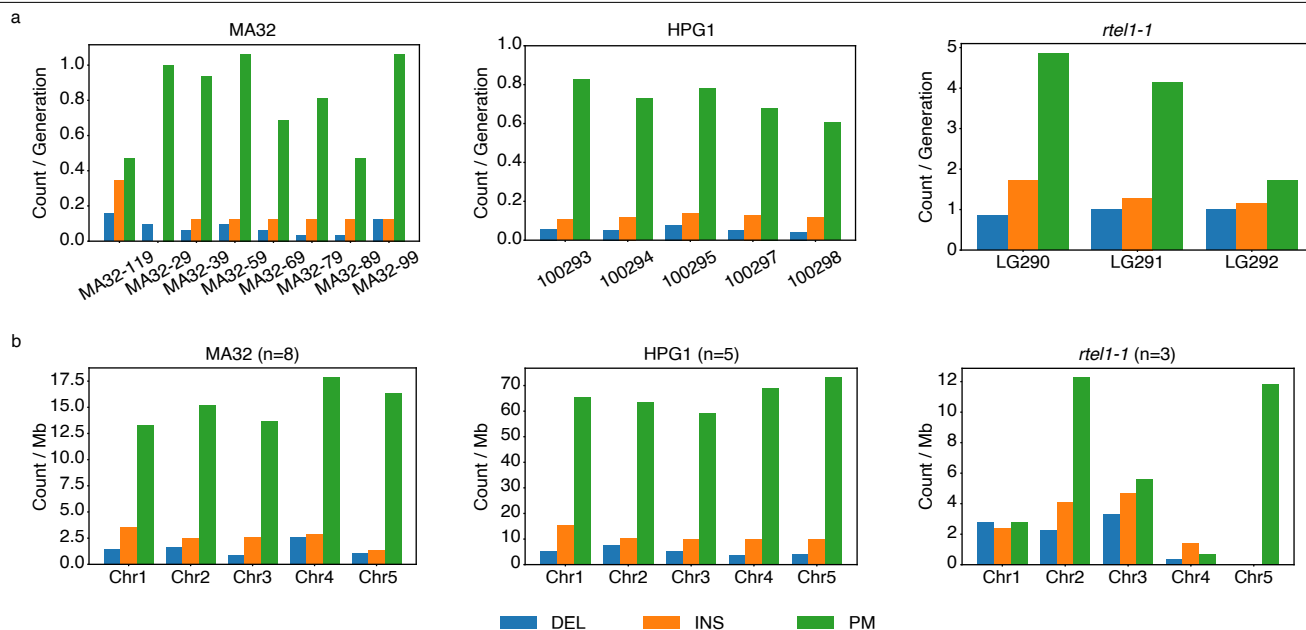
Peer review information *Nature* thanks Ian Henderson and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



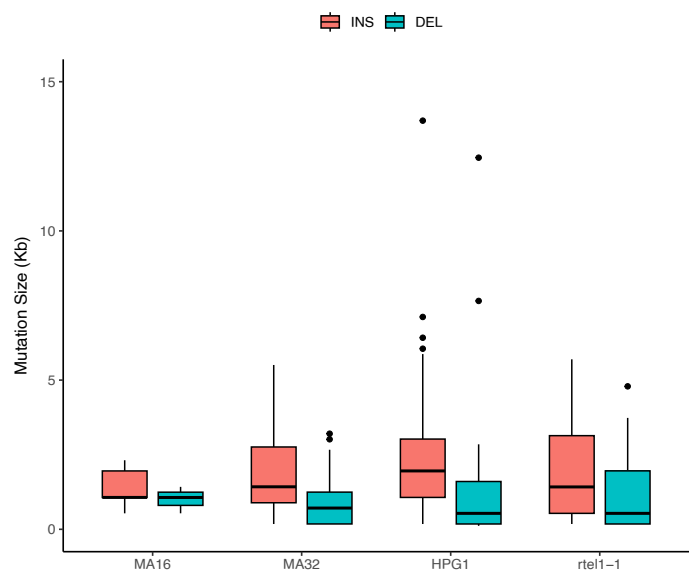
Extended Data Fig. 1 | Assembly errors identified from long-read-based genome assemblies. a. Distribution of assembly errors across the five chromosomes. From top to bottom, tracks correspond to three independent sequencing and assembly replicates for samples A and B. A1–A2 and B1–B2 were generated using PacBio HiFi reads, whereas A3 and B3 were generated using Oxford Nanopore (ONT) reads. Each circle represents an assembly error, colour-coded by type: single-nucleotide errors (yellow), insertions (red), and deletions (blue). Circle size reflects error size: small (1–2 bp), medium (3–50 bp), and large (>50 bp). Vertical black lines indicate assembly gaps. Rectangles

highlight repetitive regions: centromeres (peach), intact ATHILA elements (magenta), 5S rDNA (light blue), 45S rDNA (green), interstitial telomeric sequences (dark blue), and mitochondrial insertions (red). The bottom track shows the content of simple sequence repeats within 10 kb windows. **b.** Bar plots showing the distribution of assembly errors of different sizes across the six genome assemblies. CEN, centromere; ITSs, interstitial telomeric sequences; MT, mitochondrial sequences; SNE, single-nucleotide error; INS, insertion; DEL, deletion.

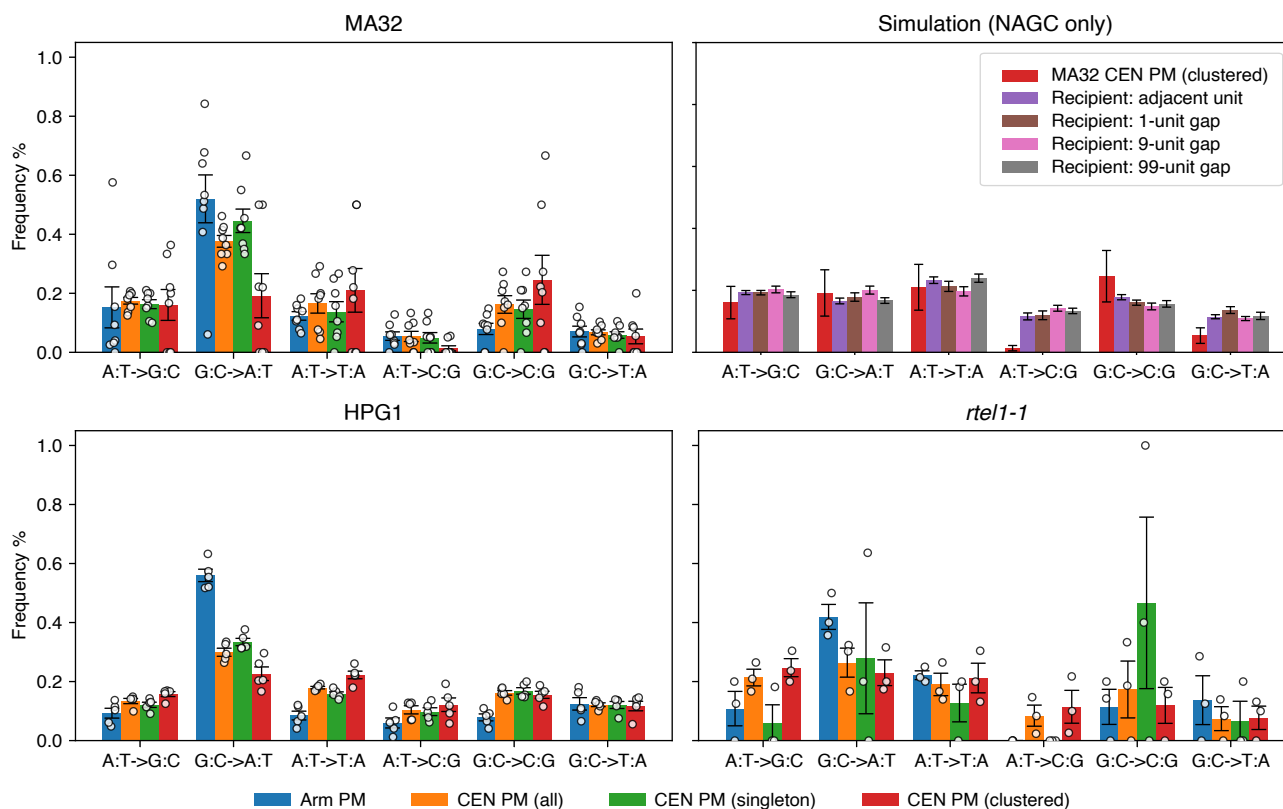


Extended Data Fig. 2 | Distribution of true homozygous centromeric mutations across three groups. Mutation counts are shown at the level of individual samples (a) and individual chromosomes (b). For comparability, mutation counts per sample were normalized by the number of selfing generations, whereas mutation counts per chromosome were normalized by centromere size. Mutations are colour-coded by type: point mutations (green), insertions (orange), and deletions (blue). Chi-square goodness-of-fit tests

revealed no significant differences in mutation counts among samples (MA32: adjusted $P = 0.0515$ for both point mutations and indels; HPG1: adjusted $P = 0.0578$ and 0.610 for point mutations and indels, respectively) or among chromosomes (MA32: adjusted $P = 0.797$ for both mutation types; HPG1: adjusted $P = 0.580$ for both mutation types), indicating a relatively uniform distribution of centromeric mutations. Statistical power was limited for *rtel1-1* due to the small sample size. PM, point mutation; INS, insertion; DEL, deletion.

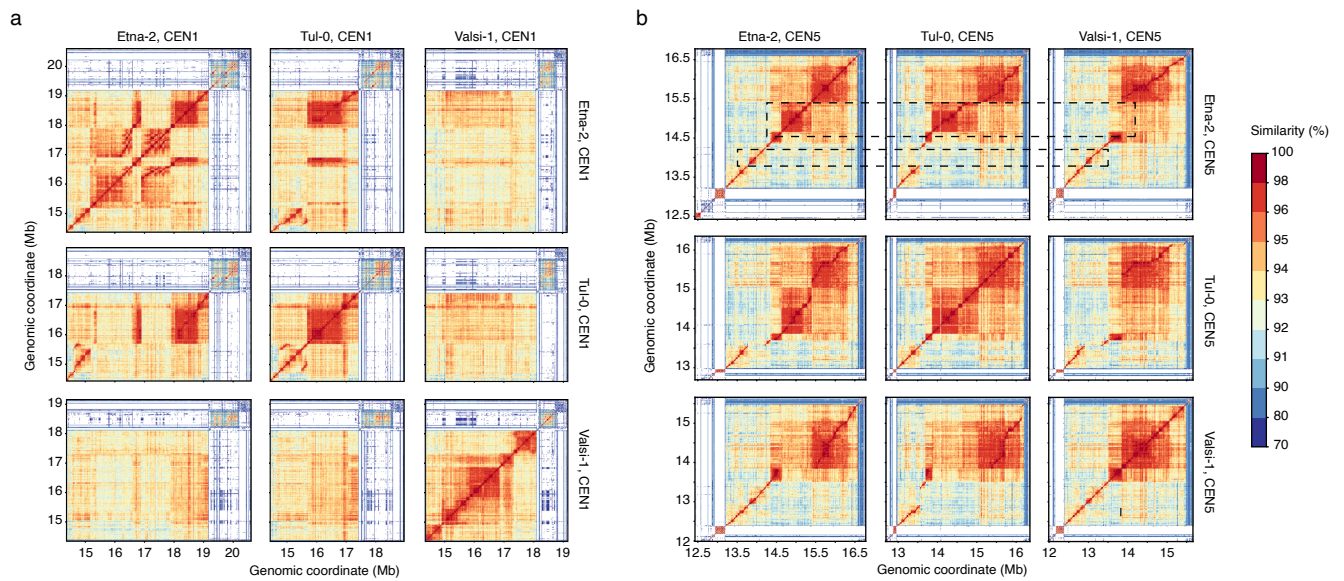


Extended Data Fig. 3 | Size distribution of homozygous centromeric indels across four groups. The boxplot illustrates the length of identified indels in MA16 (11 insertions, 3 deletions), MA32 (35 insertions, 21 deletions), HPG1 (189 insertions, 86 deletions), and *rtel1-1* (29 insertions, 20 deletions). Centre lines indicate the median; box limits indicate the upper and lower quartiles; whiskers extend to the most extreme values within $1.5 \times$ the interquartile range; points indicate outliers. Two exceptionally large deletions in *rtel1-1* (59,508 bp and 105,343 bp) are excluded from the plot for clarity.



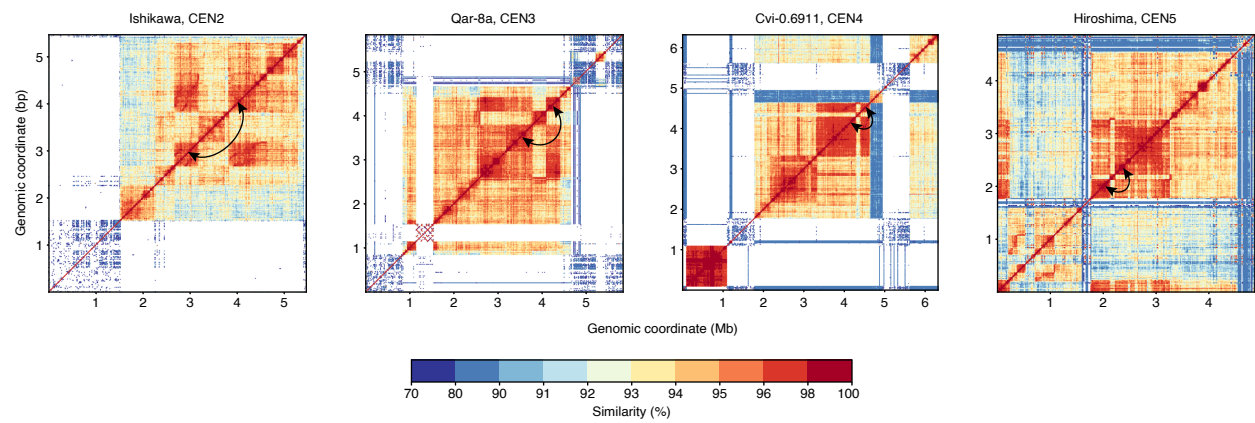
Extended Data Fig. 4 | Mutation spectra of homozygous centromeric point mutations across different groups and comparison with NAGC-only simulations. Bar plots show the frequencies of homozygous centromeric point mutations across four datasets (MA32, HPG1, *rtel1-1*, and simulation). Bars represent mean $\pm$ SEM. Dots indicate individual biologically independent lines for the experimental groups (MA32, $n = 8$; HPG1, $n = 5$; *rtel1-1*, $n = 3$). Colours indicate genomic regions or mutation categories: chromosome arms (blue), all centromeric point mutations (orange), singleton centromeric point mutations (green), and clustered centromeric point mutations (red). The upper-right

panel shows simulations including NAGC only. The simulation panel shows mean $\pm$ SEM calculated from $n = 15$ independent simulation replicates (five chromosomes, each simulated three times) without individual data points. Simulated mutation spectra are stratified by recipient unit position relative to the donor, including adjacent units, and units separated by 1, 9, or 99 intervening repeat units. This comparison allows evaluation of whether the mutation patterns generated by NAGC recapitulate the observed clustered mutation spectra.



Extended Data Fig. 5 | Examples of self- and pairwise centromeric sequence identity. a. Heatmaps showing self- and pairwise sequence identity for the centromeres of chromosome 1 in Etna-2, Tul-0, and Valsi-1. Most centromeric sequences show limited conservation between accessions, similar to the

relationships between Valsi-1 and Etna-2 or Tul-0. **b.** In contrast, accessions with similar centromeric haplotypes (as shown in centromeres of chromosome 5 in Etna-2, Tul-0, and Valsi-1) facilitate the detection of large-scale mutations, such as the putative deletion highlighted by the dashed box.



Extended Data Fig. 7 | Examples of long-distance similarities in centromeric sequences. Heatmaps show self-sequence identity for the centromeres of chromosome 2 in Ishikawa, chromosome 3 in Qar-8a, chromosome 4 in Cvi-0.6911, and chromosome 4 in Hiroshima. Sequence identity was calculated using ModDotPlot with a 10 kb window size. Regions exhibiting long-distance similarity are indicated by arrows.

Reporting Summary

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

Statistics

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

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No code used for data collection.
Data analysis	Custom code and pipelines used in this study are publicly available via the following GitHub repositories: https://github.com/schneebergerlab/replicated-assemblies-centromere-study. Publicly available software used in this study include: Hifiasm (v0.16.0-r369, v0.16.1-r375, v0.25.0), Hicanu (v2.2), ipa (v1.0.5), peregrine (v1.6.3+3.g008082a.dirty), Dorado (v1.1.1), RagTag (v1.0.1, v2.0.1), BUSCO (v5.2.2), Merquy (v1.3), RepeatMasker (v4.0.9), TRASH (v1.2), CentroAnno (v1.0.2), Bedtools (v2.29.0), minimap2 (v2.24-r1122), SYRI (v1.0), BWA-MEM (v0.7.17-r1188), Bowtie2 (v2.5.1), SAMtools (v1.17, v1.19.2), IGV (v2.13.0), BCFtools (v1.16), deepTools (v3.5.6), MUMmer (v4.0.0beta2), Pilon (v1.24), pbsv (v1.16), Sniffles (v2.0.7), DeepVariant (v1.4.0), ModDotPlot (v0.9.8), ffmpeg (v7.0.1), Cutadapt (v5.1), ATHILAFinder (v1.0), EDTA (v2.2.2), MAFFT (v7.407), FastTree (v2.1.11), iTOL (v7), Python (v3.8.8), mosdepth (v0.3.1), seqtk (v1.4-r122), SRA Toolkit (v3.2.1) We also used R (v4.1), the R script dotPlotly and the following R packages: karyoploteR (v1.22.0), maps (v3.4.3) and ggplot2 (v3.5.1).

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

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

PacBio HiFi reads for A1, previously used for the first assembly of the Arabidopsis centromeres, are available through the European Nucleotide Archive (ENA) under BioProject PRJEB46164. CENH3 ChIP-seq data (SRR4430537) and corresponding input data (SRR4430555) were retrieved from the NCBI Sequence Read Archive. The TAIR10 mitochondrial and chloroplast reference sequences were obtained from the NCBI RefSeq assembly GCF_000001735.4. Previously published genome assemblies analysed in this study are publicly available through the ENA under BioProjects PRJEB55353, PRJEB55632, PRJEB50694 and PRJEB51511, and through NCBI under BioProject PRJNA1033522. Data for the replicated MA16 genome assemblies are available through NCBI under BioProject PRJNA1259971; data for the eight MA32 lines are available through ENA under BioProject PRJEB112076; and data for the three rtl1-1 lines are available through NCBI under BioProject PRJNA1458610. Genome assemblies and sequencing data for the five HPG1 accessions are available through ENA under BioProject PRJEB75768. This BioProject includes the three HPG1 accessions analysed previously as well as the two additional accessions analysed in this study. The corresponding replicated HiFi and ONT genome assemblies for the MA16, MA32 and rtl1-1 samples are available on Figshare (<https://doi.org/10.6084/m9.figshare.32135497>).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\)](#), [and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	This study did not involve human participants, their data, or biological material. Only plant samples (<i>Arabidopsis thaliana</i>) were used.
Reporting on race, ethnicity, or other socially relevant groupings	This study did not involve human participants, their data, or biological material. Only plant samples (<i>Arabidopsis thaliana</i>) were used.
Population characteristics	This study did not involve human participants, their data, or biological material. Only plant samples (<i>Arabidopsis thaliana</i>) were used.
Recruitment	This study did not involve human participants, their data, or biological material. Only plant samples (<i>Arabidopsis thaliana</i>) were used.
Ethics oversight	This study did not involve human participants, their data, or biological material. Only plant samples (<i>Arabidopsis thaliana</i>) were used.

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

Field-specific reporting

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

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

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

Ecological, evolutionary & environmental sciences study design

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

Study description	We investigated mutation dynamics within centromeric satellite arrays using replicated genome assemblies from Arabidopsis mutation accumulation (MA) lines. Our results reveal a markedly elevated rate of centromeric point mutations, largely driven by nonallelic gene conversion between repeat units, along with frequent kilobase-scale indels that preserve tandem repeat periodicity. Analysis of rtl1-1 mutant lines supports a role for homology-directed DNA repair in shaping this mutational spectrum. Simulations based on these findings recapitulate the emergence of homogenized repeat blocks characteristic of natural centromeres, demonstrating how local mutational processes can generate and maintain higher-order centromeric structure without requiring large-scale rearrangements.
Research sample	This study analysed Arabidopsis thaliana using four datasets. The MA16 dataset comprised six samples (A1–A3 and B1–B3), representing two biological lines with three technical replicates each. The MA32 dataset included eight lines from a previously published mutation accumulation (MA) experiment. In addition, five natural HPG1 accessions and five rtl1-1 mutant lines (generation 7) were analysed.

Sampling strategy	For the generation-16 MA samples, approximately 15 sister plants were pooled per sample to generate three technical replicates (A1-A3, B1-B3), which were sequenced using both PacBio HiFi and ONT platforms. For the eight generation-32 MA samples, DNA was extracted from individual plants and sequenced using both HiFi and ONT. The five HPG1 accessions were processed following the same DNA extraction strategy as the MA32 samples and sequenced using HiFi. In addition the five rtl1-1 mutant samples followed the MA16 DNA extraction strategy (pooled sister plants) and were sequenced using HiFi and ONT.
Data collection	Genomic DNA was extracted from these samples and used for long-read and short-read sequencing and genome assembly.
Timing and spatial scale	Not applicable, as the study was based on laboratory-grown Arabidopsis thaliana lines under controlled conditions, without spatial or temporal environmental variation.
Data exclusions	No data were excluded.
Reproducibility	All samples were sequenced with ultra-high coverage PacBio HiFi reads (average 105x), ensuring robust assembly quality. For both generation-16 and generation-32 MA lines, complementary HiFi and ONT sequencing data were generated, allowing cross-platform validation and reducing biases associated with any single sequencing technology. The MA line background limits differences to mutation sites, facilitating accurate assembly validation. Detected mutations were visually confirmed in IGV and manually curated, yielding a high-confidence mutation dataset. Importantly, the mutational patterns identified were highly consistent across all four datasets, supporting the robustness and reproducibility of our findings.
Randomization	Not applicable, as samples represent predefined mutation accumulation lines propagated through single-seed descent without selection.
Blinding	Not applicable, as data analysis and mutation validation were conducted without prior knowledge.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Plants

Seed stocks	Generation-16 MA samples originated from a single Arabidopsis thaliana Col-0 mother plant (stock ID N1092), while generation-32 MA samples, also in the Col-0 background, were obtained from the Shaw experiment. The HPG1 dataset comprised five natural accessions representing populations that diverged approximately 400 years ago. The rtl1-1 samples (three lines; SALK_113285, Col-0 background) were propagated for several generations by selfing.
Novel plant genotypes	Mutations accumulated through self-pollinating and single seed descent under controlled growth conditions.
Authentication	All samples used in this study are derived from Arabidopsis thaliana. The MA and rtl1-1 lines are in the Col-0 background, whereas the HPG1 accessions represent natural populations. Sample identity and genetic background were verified based on genome-wide sequence similarity to the reference genome, except for de novo mutations accumulated during propagation.