---

**Preprint:**
Dong, X. orcid.org/0000-0001-6120-1330, Jiao, W.-B. orcid.org/0000-0001-8355-2959, Campoy, J.A. orcid.org/0000-0002-6018-5698 et al. (5 more authors) (Submitted: 2025) The mutational dynamics of the Arabidopsis centromeres. [Preprint - bioRxiv] (Submitted)

https://doi.org/10.1101/2025.06.02.657473

---

# The mutational dynamics of the Arabidopsis centromeres

Xiao Dong[1], Wen-Biao Jiao[1,#], José A Campoy[1,#], Fernando Rabanal[2], Jurriaan Ton[3], Lisa M Smith[3], Detlef Weigel[2,4] and Korbinian Schneeberger[1,5,6*]

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

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

[3] School of Biosciences and Institute for Sustainable Food, University of Sheffield, Sheffield, UK

[4] Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany

[5] Faculty of Biology, LMU Munich, Planegg-Martinsried, Germany

[6] Cluster of Excellence on Plant Sciences, Heinrich-Heine University, Düsseldorf, Germany

[#] *Present addresses*: National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops, Huazhong Agricultural University, Wuhan, China; College of Informatics, Huazhong Agricultural University, Wuhan, China (WBJ); Department of Pomology, Estación Experimental de Aula Dei (EEAD), CSIC, Saragossa, Spain (JAC).

* Correspondence: Korbinian Schneeberger (k.schneeberger@lmu.de)

**Key words:** centromere, tandem repeats, mutations, *de novo* genome assembly, *Arabidopsis thaliana*

## Abstract

**Centromeres are specialized chromosome regions essential for sister chromatid cohesion and spindle attachment during mitosis. Many centromeres comprise highly variable, megabase-scale satellite DNA arrays, yet the mutation spectrum driving this variability remains poorly understood. Using replicated genome assemblies of six Arabidopsis mutation accumulation lines, we identified centromeric mutations consisting almost exclusively of point mutations and structure-preserving, in-frame indels spanning a few kilobases. Centromeric point mutations occurred at a ninefold higher rate ($6.1 \times 10^{-8}$/bp/gen) than in chromosome arms, frequently introduced by non-allelic gene conversions from closely linked repeat units. Forward-in-time simulations based on the observed mutation spectrum recapitulated the emergence of megabase-scale higher-order repeat (HOR) structures, including long-range sequence similarities, without requiring large-scale rearrangements, closely mirroring centromeric divergence among natural genomes. Our results show that centromere evolution is driven by a unique mutational spectrum, providing a quantitative framework to understand how small-scale mutations shape and maintain the large-scale architecture of centromeric DNA.**

## Introduction

Centromeres are essential for accurate chromosome segregation during cell division. But despite their conserved function, the centromeric sequences, which often consist of megabase-scale satellite repeat arrays, show remarkable variation both within and between species[1-5]. However, the mutational dynamics driving this rapid sequence turnover remain poorly understood.

The individual repeat units of the satellite repeat arrays are not identical but instead show sequence differences that give rise to distinct patterns, such as tandem duplications or triplications, which can be further extended into higher-order repeat (HORs) structures[1,5,6]. The cores of centromeres often contain megabase-scale clusters of highly homogenized HORs, flanked by more heterogeneous peripheries characterized by more variable HORs, structural rearrangements, and transposable element insertions[1,5]. This specific organization has led to the hypothesis that

megabase-scale mutations or long-range recombination events might be required to form centromeric sequences[7].

In *Arabidopsis thaliana*, centromeres consist of megabase-scale tandem arrays of 178 bp repeat monomers (CEN178)[5]. Long-read assemblies have successfully reconstructed these arrays across all five chromosomes within the reference line and across entire populations[5,8-10]. The extreme sequence divergence observed between the centromeres of different *A. thaliana* genotypes supports the hypothesis that these regions are shaped by fundamentally different mutational dynamics compared to the rest of the genome[1,11].

Although long-read sequencing now allows the reconstruction of centromeric sequences[5,8-10], identifying rare centromeric mutations remains challenging due to the highly repetitive nature of the centromeric DNA that impacts assembly quality. To overcome this, we applied a novel replicated genome sequencing strategy to eliminate genome assembly errors, allowing us to distinguish genuine centromere mutations from assembly artifacts. Applying this approach to *A. thaliana* mutation accumulation (MA) lines revealed a unique mutational spectrum in the centromeres. This was marked by frequent in-frame indels of several kb in size that added or deleted entire repeat units, while keeping the repeat array structure intact. In addition, we found highly frequency point mutations that were frequently introduced by non-allelic gene conversions. However, we did not find any large-scale mutations that could explain the megabase-scale differences between closely-related centromeres. Interestingly, forward-in-time simulations over millions of generations demonstrated that small-scale indels, in combination with non-allelic gene conversions, are sufficient to drive the formation of megabase-scale HOR structures and ongoing centromere turnover. The simulated centromere sequences recapitulated the divergence observed in natural genomes including the formation of variable HOR structures, the occurrence of long-distance similarities and the extreme size variation or the entire array. Together, these unexpected findings uncover how centromeres evolve through a distinctive mutational regime that underly their remarkable diversification.

# Results

## Replicated genome assemblies eliminate errors

Despite significant progress in recent years, *de novo* genome assemblies still contain hundreds to thousands of errors, making it difficult to confidently identify mutations in highly repetitive regions that distinguish closely related individuals. Hence, identifying mutations in the centromeres to then simulate their dynamics requires error-free genome assemblies. To achieve this, we generated replicated genome assemblies (*i.e.*, two independent assemblies from independent samples of the same individual) to first identify errors and then use error-corrected assemblies to identify true mutations.

We first reconstructed the genomes of two *A. thaliana* Col-0 plants using two independent DNA samples extracted from pooled progeny (Fig. 1a). Both mother plants (referred to as A and B) were originally derived from the same founder plant, but were kept separate for 16 generations based on self-pollination (as part of a controlled trans-generational mutation accumulation experiment). We refer to the four DNA samples as “A1”, “A2”, “B1” and “B2” to distinguish the origin of the samples (“A” or “B” plant) and their replication (“1” and “2”) (Fig. 1a). In general, Col-0 is a diploid, inbred and selfing plant with a homozygous genome, which is identical between different Col-0 plants, except for the mutations that occurred after individual lineages were split.

We sequenced the four samples with PacBio’s HiFi technology with very high coverage ranging from 96x to 142x (Supplementary Table 1) and generated highly contiguous genome assemblies with NG50 values of 9.4 to 14.3 Mb. We scaffolded 19 to 31 contigs of each assembly into five pseudo-chromosomes using homology to the reference sequence[12] (Supplementary Table 2).

When comparing the replicated genome assemblies (A1 against A2 and B1 against B2), we identified 391, 577, 266, and 298 errors in the four assemblies, where the errors were validated and assigned to individual replicates using alignments of additional short reads (Supplementary Fig. 1a; Supplementary Table 3-6). The low error rate was confirmed by Merqury[13], which estimated quality values (QVs) ranging from 58 to 62 (Supplementary Table 2). Closer examination of the assembly errors revealed three distinct reasons (Supplementary Results, Supplementary Table 3-6).

The most common source, accounting for approximately ~80% of all errors, was simple sequence repeats, where HiFi reads failed to accurately capture the correct repeat unit count. These errors were almost exclusively small indels, typically 1-2 bp in size. Around 17% of errors were associated with random sequencing errors near GA(A) repeats, where low coverage hindered error correction through read consensus. The remaining ~3% arose from assembly artifacts or errors introduced during reference-based scaffolding. A detailed discussion of these error types is provided in the Supplementary Results.

## HiFi assemblies feature almost perfectly assembled centromeres

Centromeric regions, which have long posed major challenges for genome assembly due to their repetitive nature, were exceptionally well resolved. Across all four genomes, 17 of the 20 centromeres were fully assembled into a single contig. Unexpectedly, centromeres were among the most accurately assembled regions: although they comprise 8.3% of the genome, they accounted for less than 1% of all assembly errors.

In total we detected only 12 assembly errors, including seven large (>50 bp) and five small (1-2 bp) errors. The large errors were primarily due to mis-scaffolding, while the small ones were associated with simple sequence repeats. This high level of accuracy is likely due to a marked depletion of simple sequence repeats within centromeres, which were the main source of assembly errors elsewhere in the genome (Supplementary results; Supplementary Fig. 1a). Notably, visual inspection of raw PacBio read alignments enabled reliable distinction between true mutations and assembly errors, even without replicated genome assemblies.

## Identifying centromeric mutations with error-free assemblies

Comparing the error-free assemblies of samples A and B, we identified 200 homozygous and 29 heterozygous mutations, comprising 97 point mutations (PMs) and 132 indels ranging in size from 1 to 11,570 bp (Fig. 1b,c; Supplementary Table 7). Using allele information from the Col-0 reference genome, we assigned 115 mutations to sample A and 114 to sample B.

Previous studies have estimated a point mutation rate in *A. thaliana* of $6.95 \times 10^{-9}$ per site and generation[14 15]. The 77 homozygous point mutations (34 in A and 43 in B) are

more than twice as many as expected given this mutation rate. It is important to note, however, that the original estimate was based on "unique" regions of the genome (*i.e.,* regions accessible with short-read alignments) and did not include repetitive regions of the genome. In fact, when excluding the centromeres and 5S rDNA clusters (119.8 Mb), we identified only 39 fixed point mutations (19 in A and 20 in B), leading to a mutation rate similar to earlier estimations when considering the elevated mutation rates in transposable elements[14 15].

The remaining 38 homozygous point mutations (49%) were located within just 13.8 Mb (9.8%) of the genome, corresponding to the 5S rDNA clusters and the centromeres. This suggested an elevated point mutation rate in these regions compared to unique regions and indicates that these regions undergo different mutational dynamics as previously described for pericentromeric regions[15]. These findings contribute to the growing evidence that mutation rates are not uniformly distributed across the genome[15,16].

In the unique regions of the genome, the most common type of mutation was not point mutations but 2 bp indels, which accounted for 55.1% (n=81) of all mutations detected in these regions. All of these small indels were exclusively found in dinucleotide repeats within the chromosome arms, consistent with previous reports showing that mutation rates in dinucleotide repeats are orders of magnitude higher than those of point mutations[17]. All other small indels (1-50 bp) also occurred within simple sequence repeats, including three cases with slightly larger repeat units of up to 25 bp. In contrast, large indels (59 to 11,570 bp) were not associated with simple sequence repeats. While most small indels were confined to unique regions, large indels showed a distinct pattern in their genomic distribution, with the majority of them (15 out of 19) occurring in highly repetitive regions such as the centromeres or the 5S rDNA clusters (Fig. 1c). Only four of the large deletions occurred outside of these regions, of those three shared significant sequence homology with closely linked regions, implying that local homology is a major driver of large indel formation.

## Mutation spectrum in the centromere

In the centromeres, we identified 31 point mutations and 14 indels (11 insertions and three deletions) ranging from 535 to 2,314 bp, but no other type of mutation. This was in stark contrast to the chromosome arms, where small indels in simple sequence

repeats were the predominant mutation type. The mutations were distributed over the entire span of centromeres, including very homogeneous HORs as well as divergent peripheries, without any apparent preference for either of the regions (Fig. 2a,b). Also, no mutation bias for any of the sites within the consensus of the CEN178 repeats could be found (Extended Data Fig. 1).

Unexpectedly, the breakpoints of all indel mutations were in-frame with the repeat units (i.e., the start and end positions of the indel mutations corresponded to adjacent bases in the consensus of the CEN178 repeat) implying that centromeric indels either delete or add complete repeat units only and thereby preserve the integrity of the satellite repeat array (Fig. 2c), which is a prerequisite for highly dynamic repeat clusters, which conserve the repeat structure itself as observed between the centromeres of natural *A. thaliana* accessions. The three deletions removed 3 to 8 repeat units, while the 11 insertion mutations introduced 3 to 13 units (Extended Data Fig. 2). The inserted sequences were either entirely or at least partially identical (in the case of complex tandem duplications) to the repeat units right next to the inserted sites (Fig. 2c, d; Extended Data Fig. 3). Most insertions and deletions led also to the formation of a few variants of the repeat unit, as new recombinant variants arose at their breakpoints (Fig. 2c).

As mentioned above, the centromeres featured an increased point mutation rate as compared to the rest of the genome. While most of the point mutations were isolated and found across the entire stretch of the centromeres, five point mutations were located in the centre of a single CEN178 repeat unit, in a region not larger than 33 bp (Fig. 2f). Intriguingly, the CEN178 unit just upstream of the mutated unit contained the exact same sequence variation as introduced by the five mutations (Fig. 2e, f). The most parsimonious explanation is that the five point mutations were introduced by a non-allelic gene conversion (NAGC) and not by the more complex occurrence of five independent point mutation events. Considering the additional differences between the two units that were not converted, the estimated gene conversion tract size might have been between 33 and 126 bp (Fig. 2f).

Although the remaining point mutations were not clustered, this does not exclude the possibility that they also arose through non-allelic gene conversions. When we searched for putative donor sequences, we found that 25 of the 26 unclustered centromeric point mutations had putative donor sequences on the same centromere.

This suggests that gene conversion events are largely confined to within individual chromosomes, with no evidence for inter-chromosomal exchange. This finding is consistent with prior studies showing that centromeric satellite variants are often chromosome-specific[1,5,7,18,19], indicating that mutations within centromeres rarely propagate between chromosomes.

Some donor sequences were in closely proximity to the mutated sites, including eight cases where the putative donor was within ten adjacent repeat units, and four cases where it was directly adjacent. To test whether such proximity could occur by chance, we performed random simulations of point mutations. A high proportion of random mutations (86%) coincidentally resembled other repeat units, making it difficult to assess the actual contribution of gene conversions to the increased rate of point mutations. However, the distances between mutations and their putative donors were significantly shorter in real data than in the simulations (Fig. 2g, h), supporting the idea that the elevated point mutation rate in centromeres is, at least in part, driven by non-allelic gene conversions within the same chromosome rather than by an increased rate of spontaneous mutations.

## Estimating centromere mutation frequencies

To enable accurate estimation of mutation frequencies, we generated high-quality genome assemblies of four additional trans-generational MA lines, each propagated for 32 generations. In total, this added up to 160 generations of accumulated mutations in our panel. To confidently distinguish true centromeric mutations from the few assembly errors, we manually inspected read alignments at each assembly difference, allowing for the reliable identification of genuine centromeric mutations.

Combined with the two replicated lines, we identified 116 point mutations, 45 large indels, and 4 small indels within the centromeric regions (Extended Data Fig. 4a; Supplementary Table 8). As observed in genomes A and B, all additional large centromeric indels were in-frame, with a notable enrichment of complex tandem duplications (Extended Data Fig. 4b, d; Supplementary Fig. 7-8). The increased number of generations also revealed a broader range of indel sizes, from single-unit events to insertions spanning 31 repeat units and over 5,507 bp in length (Extended Data Fig. 4c). Overall, however, considering the overall sizes of the centromeric repeat arrays (megabase-scale), all observed indel mutation were still rather small (kb-scale).

The point mutation rate within centromeric repeat arrays was estimated at $6.12 \times 10^{-8}$ (95% CI: $5.06–7.34 \times 10^{-8}$) per site per generation, which is approximately nine times higher than a recent estimate for point mutation rates in the chromosome arms of *A. thaliana*[15] (Table 1). Besides methylated cytosines in transposable elements, which also exhibited elevated mutation rates, all other genomic regions in the chromosome arms showed significantly lower rates of point mutation.

**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 (non-allelic gene conversions)<br>• Indels (~180bp-10kb) | $6.12 \times 10^{-8}$ |
| Chromosome arms | • Point mutations (spontaneous)<br>• Small indels in simple sequence repeats (1-30 bp) | $6.95 \times 10^{-9}$ (15) |

As observed in A and B, many of the centromeric point mutations appeared to result from non-allelic gene conversions. We identified four clusters of point mutations: two clusters with two adjacent point mutations each, one with three mutations within a 104 bp window, and a striking case involving 16 point mutations and a 1 bp insertion within just 144 bp. In all cases, putative donor sequences were present on the same chromosome as the point mutation cluster further supporting the rarity of inter-chromosomal gene conversion. Despite the varying distance to their potential donor units (ranging from 10 to 3,570 repeat units), the presence of multiple clusters suggested that the elevated point mutation rate in centromeres is not primarily driven by an elevated rate of spontaneous mutations, but rather by gene conversion events enabled by the high levels of homology within centromeric repeat arrays. This interpretation is further supported by an altered point mutation spectrum in the centromeres: while GC→AT transitions are the most common type of the point mutations in the chromosome arms (~60%)[14,15], their prevalence in the centromeres was notably reduced to ~40%, consistent with a distinct mutational process. Notably, this difference could not be explained by a reduced GC content, as centromeres even have a slightly higher GC content (37.4%) as compared to chromosome arms (36.2%).

Taken together, the chromosome arms and centromeric regions display strikingly distinct mutational dynamics. These differences in mutation spectra stem from the contrasting repeat landscapes of the two regions. In chromosome arms, simple sequence repeats predominantly give rise to small indels, typically only a few base

pairs in length, coupled with spontaneous point mutations in unique regions. By contrast, the extensive local homology among repeat units in centromeric arrays facilitates in-frame indels spanning several kilobases, as well as non-allelic gene conversions that mediate short-tract sequence exchange between adjacent repeat units.

## Small-scale indels and gene conversions are sufficient to shape natural centromere structures

In both animals and plants, many centromeric repeat arrays contain large, homogenized regions composed of highly similar repeat units spanning multiple megabases[1,11]. These higher-order repeat (HORs) regions are often distinct from the rest of the centromeric repeat units. Although their origin remains unclear, mechanisms such as layered expansion, tandem duplication or unequal crossover have been proposed[1,11]. In this study, the largest indels observed in the centromeres were only a few kilobases in size, yet all were precisely in-frame with the repeat unit structure. This raised the question whether these small-scale, but structure-preserving indels could drive the emergence of HOR patterns or whether rarer, larger mutations – not detected in the 160 generations of analysed MA lines here – are also required.

To address this, we performed forward-in-time simulations of the mutational dynamics in the centromere using an initial sequence of 15,000 identical CEN178 repeats. We ran the simulations for six million generations, and in each generation, we introduced random mutations, which were sampled from the frequency distributions and spectra of the mutations found within the MA lines (Fig. 3a).

Within the burn-in phase, in which the initial similarity of the starting sequence was erased, we observed a quick sequence turnover (Fig. 3b, c, f; Extended Data Video 1-5). After approximately one million generations, none of the original repeat units remained unaltered and new similarities between regions, which did not result from the similarities of the starting sequence, could be observed (Fig. 3b). Unexpectedly, the high turnover of the consensus sequence did not lead to a continuously increasing divergence between the individual repeat units, instead, some regions became more similar over time (Fig. 3g, h). Excitingly, this led to the formation of HOR structures resembling those large-scale HOR structures seen in natural genomes, including megabase-scale clusters of highly similar repeat units, which were distinct from their

surroundings. The sizes of these HORs varied dynamically, expanding and contracting over time. This suggested that the random combination of point mutations, non-allelic gene conversions and medium-sized, kilobase-large in-frame indels is sufficient to generate megabase-scale patterns of homogenized repeat arrays – without requiring any large rearrangements or layered expansion (Fig. 3g, h).

We also observed that some of the HOR clusters were split by the emergence of distinct HOR clusters within already existing HOR cluster. This separated the original HOR cluster into to separated parts, which showed long-distance similarities (red box in Fig. 3g). These patterns appeared as they were introduced through long-distance recombination, however, here they were solely introduced by small-scale indels and gene conversions.

Also, the overall centromere length varied dramatically, some centromeres more than doubled in size, while other repeat arrays were entirely lost (Fig. 3d, e). This was unexpected, as the insertion and deletion mutations were rather small (compared to the entire centromere) and were introduced with equal probability, which should theoretically stabilize centromere length. However, across tens of thousands of generations, strong deviations from the original lengths emerged, reflecting the divergent centromeres between natural genomes and even the loss of entire centromeres observed in natural genomes that have been separated for tens to hundreds of thousands of years separated[20] – again all the patterns were shaped only through continuous introduction of small-scale, in-frame indels and gene conversions.

In summary, the simulations evidenced that small-scale mutations can recapitulate the emergence of megabase-scale HOR structures, long-distance similarities and drastic differences between centromere sizes. The HOR structures had a high turn-over rate, which mirrored the strong differences between different accessions in *A. thaliana*, which are usually tens to hundreds of thousands of generations apart. In contrast, the consensus sequences themselves changed only gradually, and there were still significant similarities between the consensus sequences at the end of the simulations, which were on the same order as the similarities between the consensus sequences of different chromosomes in *A. thaliana*.

## Discussion

In this study, we leveraged *A. thaliana* MA lines to explore the mutational dynamics of centromeres. To overcome the inherent challenges of detecting rare mutations in repetitive regions, we generated replicated, error-free genome assemblies. Despite the long-standing difficulty of assembling centromeric repeat arrays, we found that PacBio HiFi assemblies not only fully captured the centromeric sequences, but also contained significantly fewer assembly errors than other genomic regions. Remarkably, the previously "unassemblable" centromeres emerged as some of the most precisely assembled parts of the genome.

Recent analysis of human centromeres and pericentromeric regions in *A. thaliana* have suggested increased point mutation rates relative to chromosome arms[11,15]. Indeed, although centromeric repeat arrays share well-defined consensus sequences, individual repeat units can differ by up to 40% within a single centromere. Our analysis of the MA lines confirmed a nine-fold increase in point mutation rates within centromeres, driven largely by non-allelic gene conversions that shuffle alleles among nearby repeat copies. This elevated gene conversions rate may stem from the absence of meiosis-associated crossovers in centromeres as previously hypothesized[21,22]. Alternatively, it could result from somatic homologous recombination during DNA repair in meristematic tissues. In contrast to earlier studies implicating transposable elements (TEs) in centromere diversification[1], we found no TE-associated mutations, suggesting that TE-driven structural change is rare and likely occurs over longer evolutionary timescales.

In addition to point mutations, we observed frequent in-frame indels in centromeres, with lengths up to 5.5 kb[14]. However, we did not detect any large-scale mutations that could account for the mega-base differences observed between the centromeres of different *A. thaliana* accessions[1].

The perhaps most exciting result of our study was that forward-in-time simulations, parameterized with empirically observed mutation types and frequencies, demonstrated that random, small-scale (*i.e.*, several kb-long) indels and gene conversions alone are sufficient for the formation of megabase-scale HOR structures closely. The simulated repeat arrays closely mirrored those observed in natural centromeres, highlighting that homogenization does not require a specific mechanism linked to CENH3 binding or kinetochore formation, as previously proposed[1,23,24].

Instead, the simulations revealed a self-organizing process of repeat homogenization reminiscent of concerted evolution, which has been described for rDNA and other satellite DNA families[22,24-27]. These self-organizing dynamics fit to a feedback model proposed for the evolution of satellite repeats[28], which can now be explained with specific mutational processes capable of driving such patterns.

Interestingly, our simulations also revealed that new HOR structures can emerge within existing clusters, effectively subdividing them into distinct domains. The two parts of the separated cluster shared more similarity to each other than with the new separating HOR, a pattern previously attributed to long-range recombination[5,29,30]. However, our results demonstrate that such structures can also arise solely through short-range indels and gene conversions.

Furthermore, the simulations showed that changes in repeat unit consensus do not require the genesis of entirely new centromeric arrays but can emerge through gradual, cumulative sequence changes. Since centromere function does not depend on a specific repeat sequence, reduced meiotic recombination in centromeric regions and relaxed selective constraints may allow the spontaneous formation and propagation of tandem repeat arrays ultimately enabling the birth of new centromeric consensus sequences.

Nonetheless, this also means that some questions remain unresolved. Despite the high turnover rates of centromeric repeats, the consensus sequences across different chromosomes remain strikingly similar. This implies some sort of inter-chromosomal homogenization, potentially facilitated by retrotransposons enriched at centromeres[31]. However, we found no direct evidence supporting such exchange in our dataset.

Taken together, our study provides a comprehensive analysis of the spectrum and frequency of centromeric mutations in *A. thaliana*, showing how small-scale changes can have large-scale impacts over evolutionary time. Our findings offer new answers to long-standing questions about centromere evolution and shed light on the mechanisms that change and maintain these highly dynamic regions of the genome.

# Online Methods

## Plant material

Samples A and B were derived from a trans-generational mutation accumulation (MA) experiment carried out here. Both lines originated from the same *A. thaliana* mother plant (Col-0, stock ID N1092) and were propagated independently for 16 generations through self-pollination and single-seed descent under controlled conditions (22°C, 16-hour photoperiod). For genome assembly of generation 16, we sequenced pools of generation-17 sister plants. Whole plants (excluding flower stems) were harvested at 3-4 weeks old, pooling material from approximately 15 sister plants per sample to ensure sufficient DNA yield. Each sample was processed as two independent DNA pools, generating duplicates: A1 and A2, B1 and B2. Additionally, four samples from generation 32 were included from the Shaw experiment[https://academic.oup.com/genetics/article-abstract/155/1/369/6048019]. Seeds were germinated on soil, stratified in darkness at 4°C for 6 days, and transferred to long-day conditions (23°C, 16-hour photoperiod). To minimize starch accumulation, 26-day-old plants were placed in darkness for 24 hours before harvesting. Approximately 300 mg of flash-frozen rosettes from each individual plant was ground in liquid nitrogen.

## DNA extraction and genome sequencing

For samples A and B, high molecular weight (HMW) DNA was extracted from 1.5 grams of pooled vegetative tissue using the NucleoBond HMW DNA kit (Macherey Nagel). DNA quality was assessed with FEMTOpulse (Agilent), and concentration was measured with Quantus fluorometer (Promega). Four HiFi libraries were prepared following the "Procedure & Checklist – Preparing HiFi SMRTbell® Libraries using SMRTbell Express Template Prep Kit 2.0", with DNA fragmentation via g-Tubes (Covaris) and size selection using SageELF (Sage Science). Libraries were sequenced on four SMRT cells using the Sequel II system with Binding Kit 2.0 and Sequel II Sequencing Kit 2.0 for 30 hours at the Max Planck Genome Center, Cologne, Germany. To assess assembly quality, we also sequenced additional pools of A and B sister plants using PCR-free Illumina paired-end short reads. DNA extraction was

performed with the Macherey-Nagel DNA Maxi kit, and sequencing was conducted by Novogene.

For the four additional MA lines from generation 32, HMW-DNA was extracted with a modified protocol[10] with β-mercapto-ethanol added during lysis and a phenol:chloroform purification step. After lysis and protein precipitation, DNA was purified through two rounds of bead cleanups using SeraMag SpeedBeads® and AMPure PB magnetic beads. DNA was fragmented with a gTUBE (Covaris), and 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.

## Genome assembly and scaffolding

We tested multiple assemblers[32-34] to assemble the A1 genome and found that Hifiasm[32] v0.16.0-r3699 produced the most contiguous assembly. We used Hifiasm[32] with the parameter "-l0" for all four generation-16 samples. Organellar contigs were then identified based on sequence alignment to TAIR10[35] mitochondria and chloroplast references (GCF_000001735.4), retaining those with ≥80% identity and coverage. Non-organellar contigs were scaffolded into pseudo-chromosomes using RagTag[12] v1.0.1, based on alignment to the Col-CEN[5] reference genome. Additionally, genome assemblies for four generation-32 MA lines were generated using Hifiasm[32] v0.16.1-r375 and scaffolded with RagTag[12] v2.0.1 (scaffold -q 60 -f 30000 -l 0.5 -remove-small), excluding contigs <100 kb.

## Assembly evaluation

We computed Benchmarking Universal Single-Copy Ortholog (BUSCO) scores using BUSCO[36] v5.2.2 with the parameters “-l embryophyte_odb10 -m genome.” Additionally, we assessed consensus quality and completeness using Merqury[33] 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 using Meryl[37] v1.3 and then merged with Meryl’s union-sum function. Merqury[13] was subsequently applied to each assembly to obtain genome-wide consensus quality values (QV) and completeness scores.

## Repeat annotation

Following the approach of Rabanal and colleagues[10], we used RepeatMasker v4.0.9 (http://www.repeatmasker.org) with a custom library (-lib rDNA_NaishCEN_telomeres.fa -nolow -gff -xsmall -cutoff 200) to annotate 5S rDNA, 45S rDNA, and telomere sequences. Mitochondrial insertions on Chr2 were identified by aligning to the TAIR10[35] mitochondrial sequence using minimap2[38] v2.24-r1122. Centromeres were annotated using TRASH[29] v1.2 (--seqt CEN178.csv --horclass CEN178 --par 5 --horonly), 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-nucleotide and hepta-repeats. Bedtools[39] v2.29.0 intersect was used to compare assembly errors and mutations across different repeat types.

## Identification of assembly errors

To identify assembly discrepancies between replicate genome assemblies, we performed pairwise alignments of the four assemblies (A1, A2, B1, B2) — A1 vs. A2 and B1 vs. B2 — using minimap2[38] v2.24-r1122 (-ax asm5 --eqx). Assembly differences were identified with SYRI[40] v1.0. To determine which replicate contained the assembly error, we mapped both HiFi and Illumina reads to their corresponding assemblies using minimap2[38] (-ax map-hifi) and BWA-MEM[41] v0.7.17-r1188, respectively. Alignments were sorted and converted to BAM files with Samtools[42] v1.19.2. Additionally, HiFi reads from each replicate were aligned to the opposing replicate's assembly for cross-comparison. For regions flagged by SYRI[40], we examined alignments in IGV[43] 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 coverage[10]. These were classified as incomplete rather than erroneous assemblies, as the low-depth pattern was consistent across samples.

## Identification of mutations

Assembly differences between sample A and B, and among the four generation-32 MA lines, were first detected by aligning assemblies with minimap2[38], sorting the alignments with Samtools[42], and calling differences using SYRI[40]. Mutations were manually validated in IGV[43] by visually inspecting whether HiFi reads aligned cleanly to their respective assemblies but showed mismatches when aligned to the counterpart. For generation-32 MA lines, we focused specifically on mutations within centromeric regions, while for samples A and B, mutations were identified across the entire genome. To distinguish mutated from wild-type alleles, we referenced the TAIR10[35] and the Col-CEN[5] assemblies, assigning the allele matching the reference as wild-type and the differing one as a mutation. To ensure accurate coordinate mapping and cross-sample comparison, assemblies representing the ancestral state — prior to any mutations — were constructed for samples A and B, as well as for the four generation-32 MA lines (see Supplementary Methods).

## Mutation rate calculation

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

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

where $m$ = total observed mutations, $N$ = genome size (nucleotides), and $g$ = generations. Mutations from all six MA lines were pooled to calculate the overall mutation rate. Total mutations were divided by the cumulative nucleotide-generations ($\sum[N \cdot g] = 1.895 \times 10^9$), yielding:

$$\mu = 6.12 * 10^{-8}\ (95\%\ CI: 5.06 * 10^{-8}\ -\ 7.34 * 10^{-8})$$

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

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

## Dot plot analysis for centromeric mutations

To investigate the relationship between centromeric indel mutations and the CEN178 centromeric repeat unit, we aligned the indel sequences to the CEN178 consensus

sequence using MUMmer[44] v4.0.0beta2 with nucmer (--maxmatch -l 10 -c 20). Alignments were converted from .delta to .coords format using show-coords (-c). Dot plots were generated with the R package dotPlotly (https://github.com/tpoorten/dotPlotly) (-m 10 -q 10 -k 5 -l -x). For a precise analysis of large indel mutations, we applied a word-based alignment approach. This method identified the exact matching sequences (or words) of at least 50 bp between the reference sequence (comprising 3 kb flanking each side of the mutation site in F0) and the query sequence (the mutation sequence plus 3 kb flanks). Unlike conventional alignment algorithms that allow mismatches and gaps, the word-based approach emphasizes high-confidence anchors by detecting long, uninterrupted matches, which makes it particularly effective for mapping mutations in repetitive regions. A custom Python script was used to generate dot plots from these alignments. Unlike the earlier method, this approach disallowed mismatches, enabling a more detailed examination of centromeric indel mutation patterns.

## Simulation of mutation accumulation in centromeres over time

Large-scale simulations were performed starting with 15,000 copies of the CEN178 consensus sequence (2.67 Mb), simulating spontaneous point mutations, in-frame indels, and NAGC, to investigate the impact of these mutation types on centromere evolution (see Supplementary Methods).

## Data availability

The HiFi reads for A1 used in this study were previously published in the context of the first assembly of the Arabidopsis centromeres[5]. To ensure comprehensive data access and support further analysis, we are releasing the complete set of HiFi and Illumina reads generated for this study. All data supporting the findings are provided within the main text and supplementary information files. Data related to the replicated genome assemblies are available through NCBI under BioProject accession numbers PRJNA1259971; the data for the other four assemblies can be found at PRJEB77512.

## Code availability

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

## Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2048/1–390686111 (KS) and the Transregio Collaborative Research Center TRR356/1 2023 ‘Genetic diversity shaping biotic interactions of plants’ (491090170) (DW, KS), the Max Planck Society (DW) and the European Research Council (ERC) grant “BYTE2BITE” (101124694) (KS).

## Author contributions

XD, DW and KS developed and supervised the project. JAC, FR, JT, LMS performed plant work and generated data. XD, FR assembled the genomes. XD performed the data analysis with help from WBJ and FR. XD and KS wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

# Figures

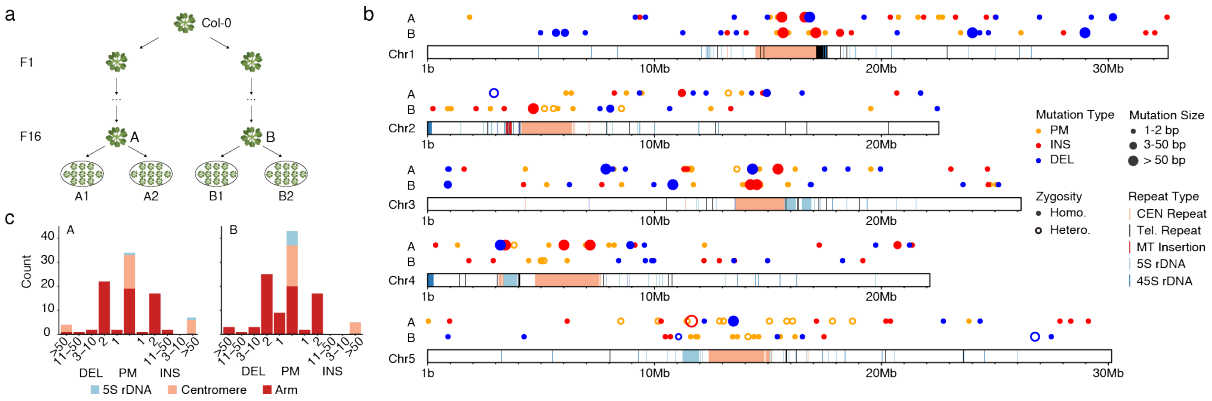



**Fig. 1. Mutation accumulation in closely related *A. thaliana* genomes revealed by error-free genome assemblies. a.** Pools of sister plants were sequenced (sampled from ~15 distinct sister plants per pool), with two independent pools from the progeny of each mother plant (referred to as A or B). Pooling genomes of sister plants dilutes somatic and gametophytic mutations in the mother plant and reconstitutes the genome of the mother plant (F16). All differences between genome assemblies of sister pools should therefore result from assembly errors. **b.** Distribution of true mutations in samples A and B. Two rows of circles represent mutation patterns, colour-coded by type: PMs (orange), insertions (red), deletions (blue), with circle size reflecting mutation size: small (1-2 bp), medium (3-50 bp), and large (>50 bp). Solid circles mark homozygous, hollow circles heterozygous mutations. Chromosomes are shown as rectangles with repetitive regions highlighted: centromere (peach), mitochondrial insertions (red), 5S rDNA (light blue), and 45S rDNA (dark blue). **c.** Bar plots show mutation counts in samples A and B, with the middle bar representing point mutations, right bars showing insertions, and left bars showing deletions. Mutation size increases from the centre outward, with colour coding distinguishing counts in 5S rDNA (light blue), centromere (peach), and remaining regions (red). PM: point mutation; INS: insertion; DEL: deletion.

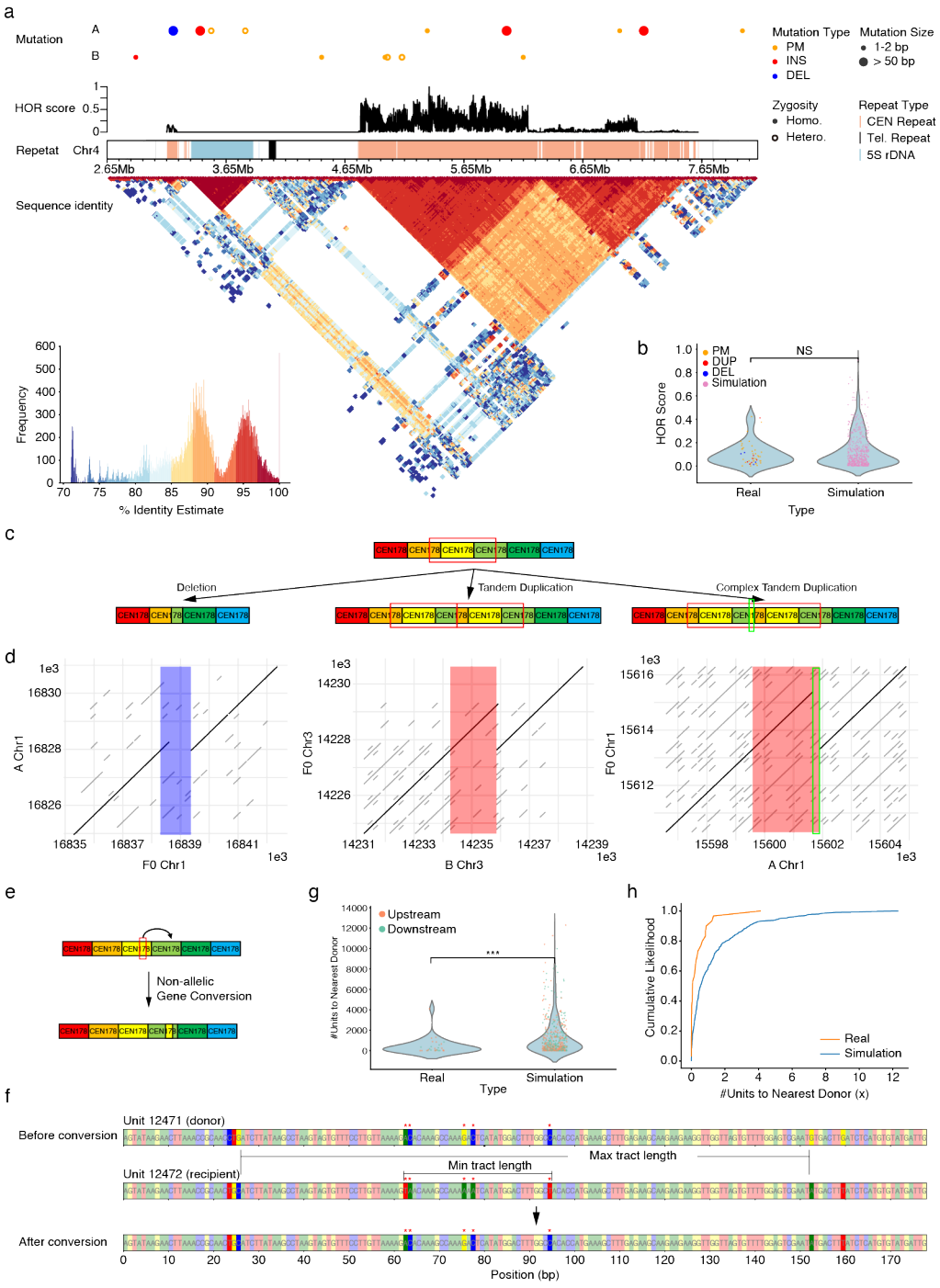



**Fig. 2. Mutations within centromeres. a.** Two rows of circles display mutation patterns in the centromere of chromosome 4, colour-coded as point mutations (PMs; orange), insertions (red), and deletions (blue), with circle size reflecting mutation length: small (1-2 bp), medium (3-50 bp), and large (>50 bp). Solid circles indicate homozygous mutations, while hollow circles indicate heterozygous ones. A line plot shows higher-order repeat (HOR) scores, rectangles highlight repetitive regions, and the heatmap below represents pairwise sequence identity between non-overlapping

10 kb regions, with a histogram summarizing identity values in the lower-left corner. **b.** Violin plots compare HOR scores of observed mutations (colour-coded by type) to 500 random centromere positions (grey), with a Mann-Whitney U test showing no significant difference (P = 0.826). **c.** Schematics illustrate three mutation types: deletion (left), tandem duplication (middle), and complex tandem duplication (right), with red boxes highlighting mutated patterns. **d.** Dot plots show collinearity of three example mutations against the CEN178 consensus sequence, demonstrating in-frame mutation patterns. The plots show sequence identity between wild-type and mutated sequences (A or B), with the longest identical sequences in black and others in grey. Mutated regions are colour-coded: blue (deletion) and red (insertion). **e.** Schematic representation of NAGC in the centromere, with a simplified example where the donor unit is adjacent to the recipient (though in reality, they may be separated). **f.** A schematic example shows five point mutations introduced by a single NAGC event rather than five independent mutations, with colours representing nucleotide bases. **g.** Violin plots compare repeat unit numbers between potential donors of observed point mutations and 500 random centromere positions, with donors being colour-coded: downstream (blue) and upstream (pink), showing a significant difference (P = 3.3e-05, Mann-Whitney U test). **h.** Line plots show the cumulative likelihood of donor distance (in repeat units) for real and simulated mutations, with asterisks indicating five-point mutations located centrally within a single CEN178 repeat unit. DEL: deletion; TD: tandem duplication; CTD: complex tandem duplication.

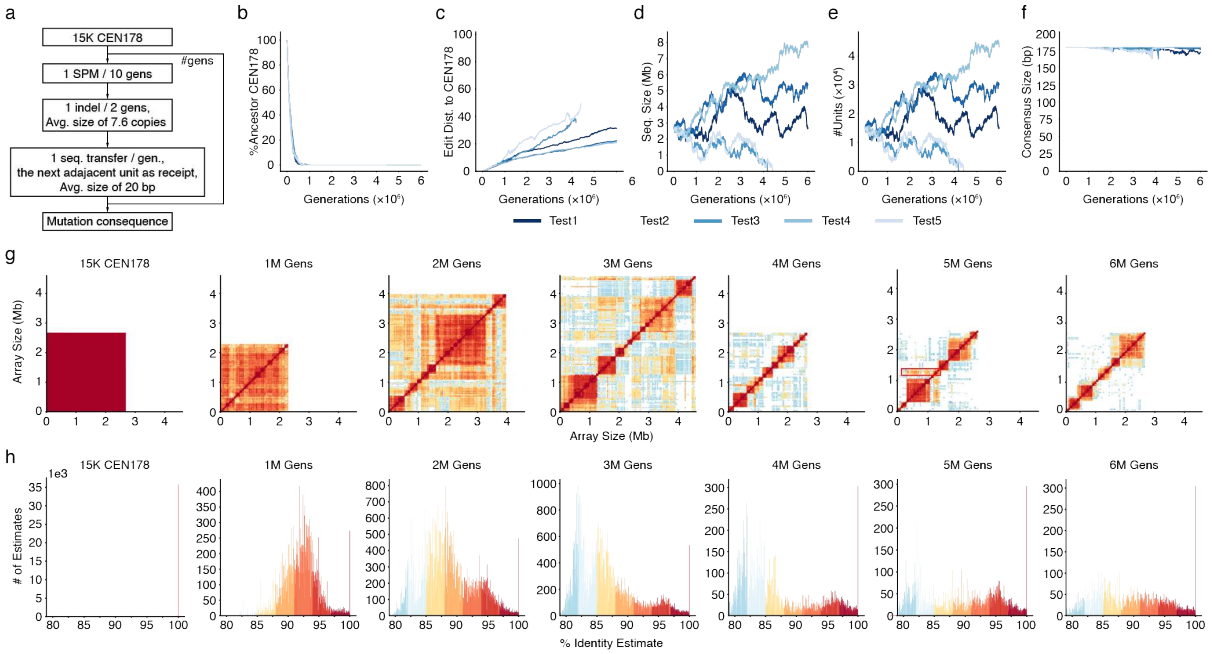
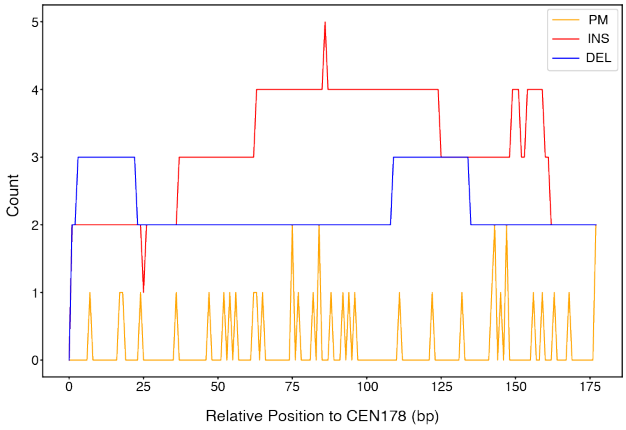
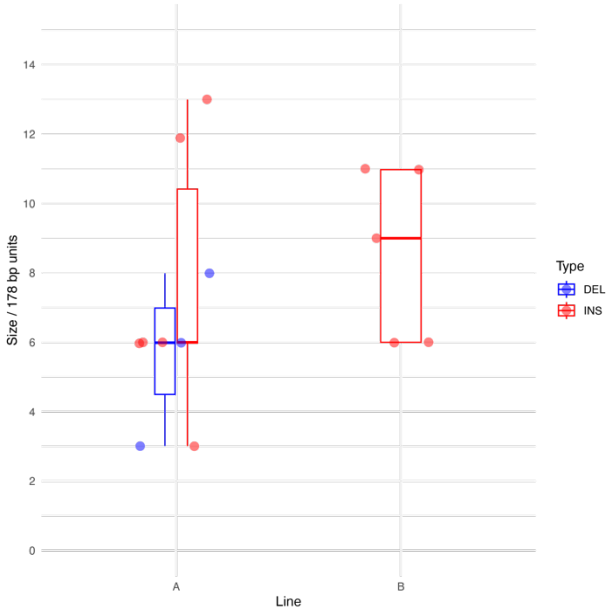


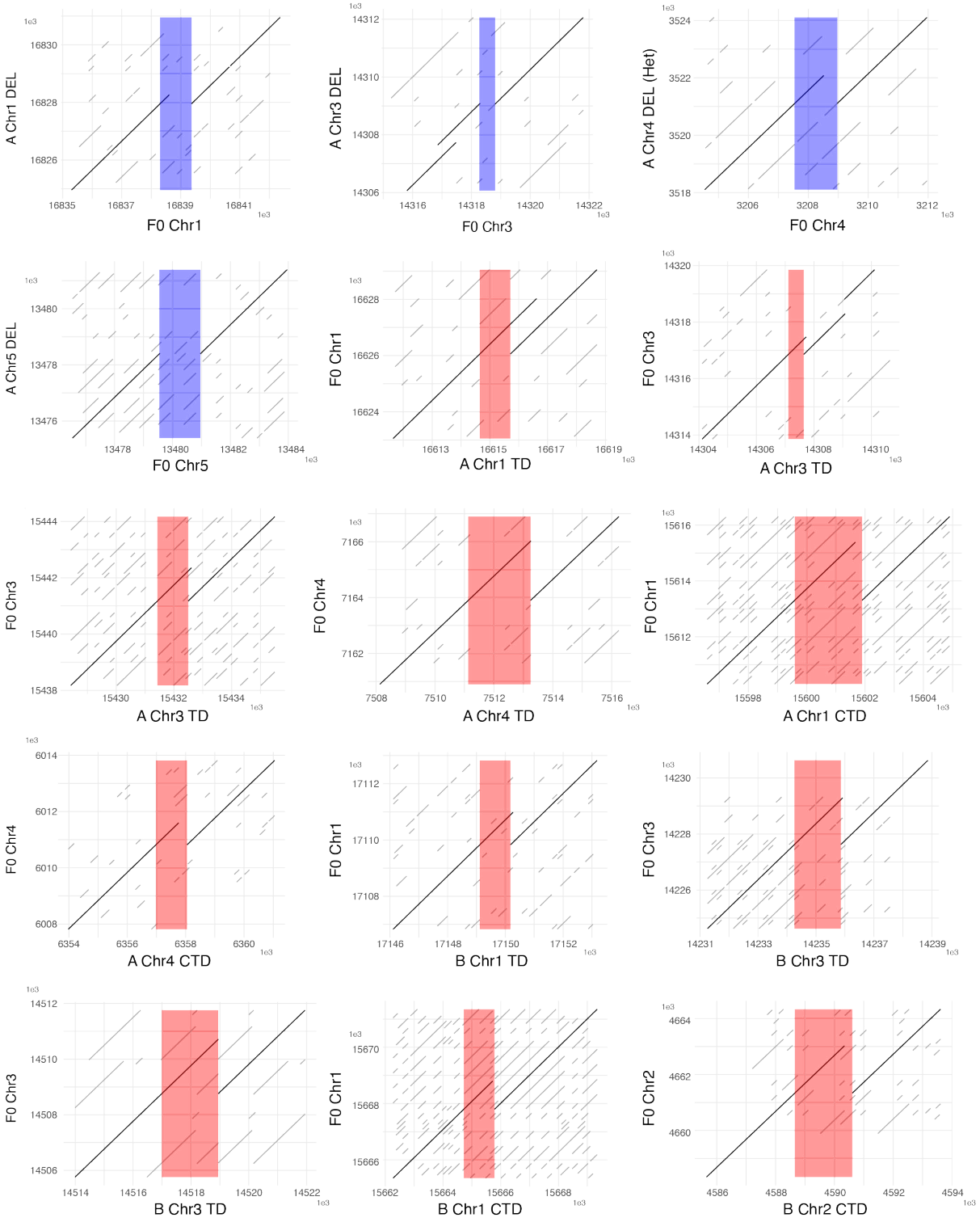

**Fig. 3. Small-scale mutations can drive the emergence of megabase-scale HOR blocks and homogenize the centromeres. a.** Diagram of forward-in-time simulation of centromeric mutation accumulation. **b-f.** Line plots tracking changes in five repeat array characteristics across five tests over six million generations: **b.** Proportion of ancestral CEN178 units. **c.** Average edit distances of repeat units to CEN178. **d.** Total array sizes. **e.** Repeat unit count. **f.** Average repeat unit sizes. **g.** Heatmap showing pairwise sequence identity between non-overlapping 10 kb regions in the simulated repeat array (Test 1) after 1 to 6 million generations, colour-coded by sequence identity values shown in **h.** The red box in the heatmap at 5 million generations highlights a region of long-distance sequence similarity. **h**. Histogram of pairwise sequence identity values from **g** (Test 1). Indel: insertion and deletion mutations.



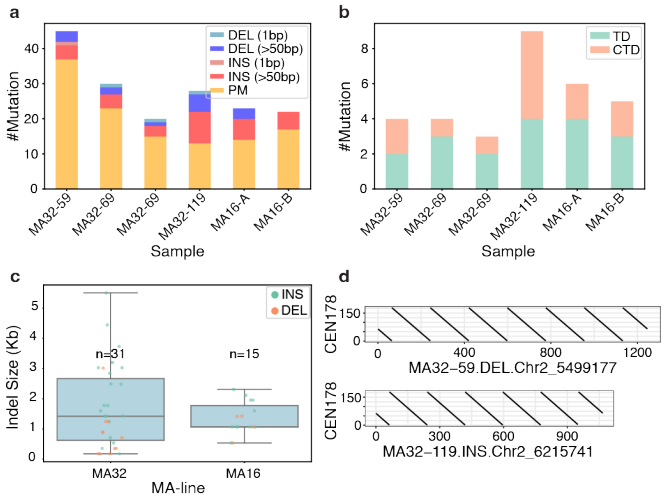

**Extended Data Fig. 1. Mutations occur without preference relative to the CEN178 consensus sequence.** Line plots display the number of mutated positions relative to the CEN178 consensus sequence. Mutation types are color-coded: point mutations (orange), insertions (red), and deletions (blue). Note that the left alignment of indels may shift their apparent position. We evaluated potential indel sites multiple times using word-based alignment to ensure consistent positioning.




**Extended Data Fig. 2. Centromeric indel sizes are preferentially multiples of 178 bp.** Box plots show the sizes of centromeric indels in samples A and B. The data reveal a strong tendency for indel sizes to occur in increments of 178 bp, which is the length of the Arabidopsis centromeric repeat unit.




**Extended Data Fig. 3. Centromeric indels exhibit complex structures.** Dot plots display the 15 centromeric indels from samples A and B, along with 3 kb of flanking sequence aligned to the wild-type sequence. Segments indicate regions with a minimum 50 bp sequence identity based on word-based alignment, revealing complex rearrangement patterns of these mutations.




**Extended Data Fig. 4. Similar centromeric mutation dynamics observed in generation-16 and generation-32 samples. a.** Bar plots show the number of fixed centromeric mutations in six samples (four generation-32 and two generation-16). Mutations are color-coded: point mutations (orange), small insertions (light red), large insertions (dark red), small deletions (light blue), and large deletions (dark blue). **b.** Bar plots show the number of tandem duplications and complex tandem duplications identified across the six samples. **c.** Box plots illustrate the distribution of indel sizes in the two MA lines. **d.** Dot plots provide examples of in-frame indels from the additional generation-32 samples, demonstrating the same in-frame patterns across all samples.

**Extended Data Video 1-5. Dynamic changes of a simulated centromere (Test 1-5).** The videos show the progression of pairwise sequence identity between non-overlapping 10 kb regions in the simulated repeat arrays (Test 1-5) over six million generations. Sequence identity values are color-coded according to the scale shown on the right.

# Reference


1 Wlodzimierz, P. *et al.* Cycles of satellite and transposon evolution in Arabidopsis centromeres. *Nature* **618**, 557-565 (2023).

2 Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098-1102 (2001).

3 Talbert, P. B. & Henikoff, S. What makes a centromere? *Experimental cell research* **389**, 111895 (2020).

4 Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).

5 Naish, M. *et al.* The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* **374**, eabi7489 (2021).

6 Melters, D. P. *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology* **14**, 1-20 (2013).

7 Naish, M. & Henderson, I. R. The structure, function, and evolution of plant centromeres. *Genome Research* **34**, 161-178 (2024).

8 Wang, B. *et al.* High-quality Arabidopsis thaliana genome assembly with nanopore and HiFi long reads. *Genomics, proteomics & bioinformatics* **20**, 4-13 (2022).

9 Hou, X., Wang, D., Cheng, Z., Wang, Y. & Jiao, Y. A near-complete assembly of an Arabidopsis thaliana genome. *Molecular Plant* **15**, 1247-1250 (2022).

10 Rabanal, F. A. *et al.* Pushing the limits of HiFi assemblies reveals centromere diversity between two Arabidopsis thaliana genomes. *Nucleic Acids Research* **50**, 12309-12327 (2022).

11 Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629**, 136-145 (2024).

12 Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome biology* **23**, 1-19 (2022).

13 Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology* **21**, 1-27 (2020).

14 Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *science* **327**, 92-94 (2010).

15 Weng, M.-L. *et al.* Fine-grained analysis of spontaneous mutation spectrum and frequency in Arabidopsis thaliana. *Genetics* **211**, 703-714 (2019).

16 Monroe, J. G. *et al.* Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature* **602**, 101-105 (2022).

17 Marriage, T. N. *et al.* Direct estimation of the mutation rate at dinucleotide microsatellite loci in Arabidopsis thaliana (Brassicaceae). *Heredity* **103**, 310-317 (2009).

18 Song, J.-M. *et al.* Two gap-free reference genomes and a global view of the centromere architecture in rice. *Molecular plant* **14**, 1757-1767 (2021).

19 Wang, T. *et al.* A complete gap-free diploid genome in Saccharum complex and the genomic footprints of evolution in the highly polyploid Saccharum genus. *Nature plants* **9**, 554-571 (2023).

20 Lysak, Martin A. "Live and let die: centromere loss during evolution of plant chromosomes." (2014): 1082-1089.

21 Talbert, P. B. & Henikoff, S. Centromeres convert but don't cross. *PLoS biology* **8**, e1000326 (2010).
22 Shi, J. *et al.* Widespread gene conversion in centromere cores. *PLoS biology* **8**, e1000327 (2010).
23 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 genetics* **5**, e1000641 (2009).
24 Rudd, M. K., Wray, G. A. & Willard, H. F. The evolutionary dynamics of α-satellite. *Genome research* **16**, 88-96 (2006).
25 Durfy, S. J. & Willard, H. F. Concerted evolution of primate alpha satellite DNA: evidence for an ancestral sequence shared by gorilla and human X chromosome alpha satellite. *Journal of molecular biology* **216**, 555-566 (1990).
26 Chatterjee, B. & Lo, C. Chromosomal recombination and breakage associated with instability in mouse centromeric satellite DNA. *Journal of molecular biology* **210**, 303-312 (1989).
27 Wolfgruber, T. K. *et al.* High quality maize centromere 10 sequence reveals evidence of frequent recombination events. *Frontiers in plant science* **7**, 308 (2016).
28 Nijman, I. J. & Lenstra, J. A. Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeats. *Journal of Molecular Evolution* **52**, 361-371 (2001).
29 Wlodzimierz, P., Hong, M. & Henderson, I. R. TRASH: tandem repeat annotation and structural hierarchy. *Bioinformatics* **39**, btad308 (2023).
30 Wang, L. *et al.* A telomere-to-telomere gap-free assembly of soybean genome. *Molecular Plant* **16**, 1711-1714 (2023).
31 Birchler, J. A. & Presting, G. G. Retrotransposon insertion targeting: a mechanism for homogenization of centromere sequences on nonhomologous chromosomes. *Genes & Development* **26**, 638-640 (2012).
32 Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**, 170-175 (2021).
33 Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome research* **30**, 1291-1305 (2020).
34 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* **37**, 540-546 (2019).
35 de, A. G. I. g. t. o. g. g. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *nature* **408**, 796-815 (2000).
36 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).
37 Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818-2824 (2008).
38 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
39 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

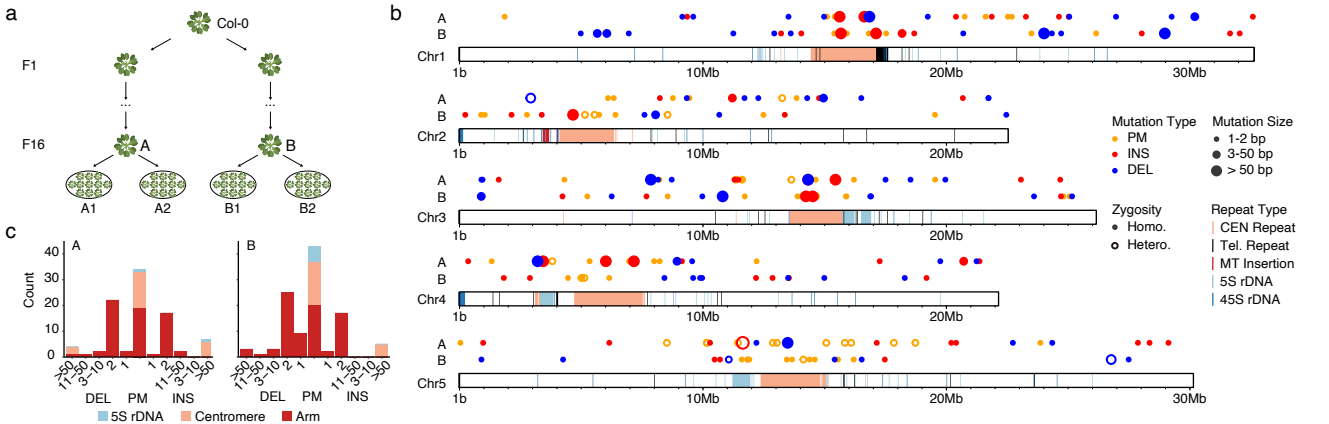
40 Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome biology* **20**, 1-13 (2019).

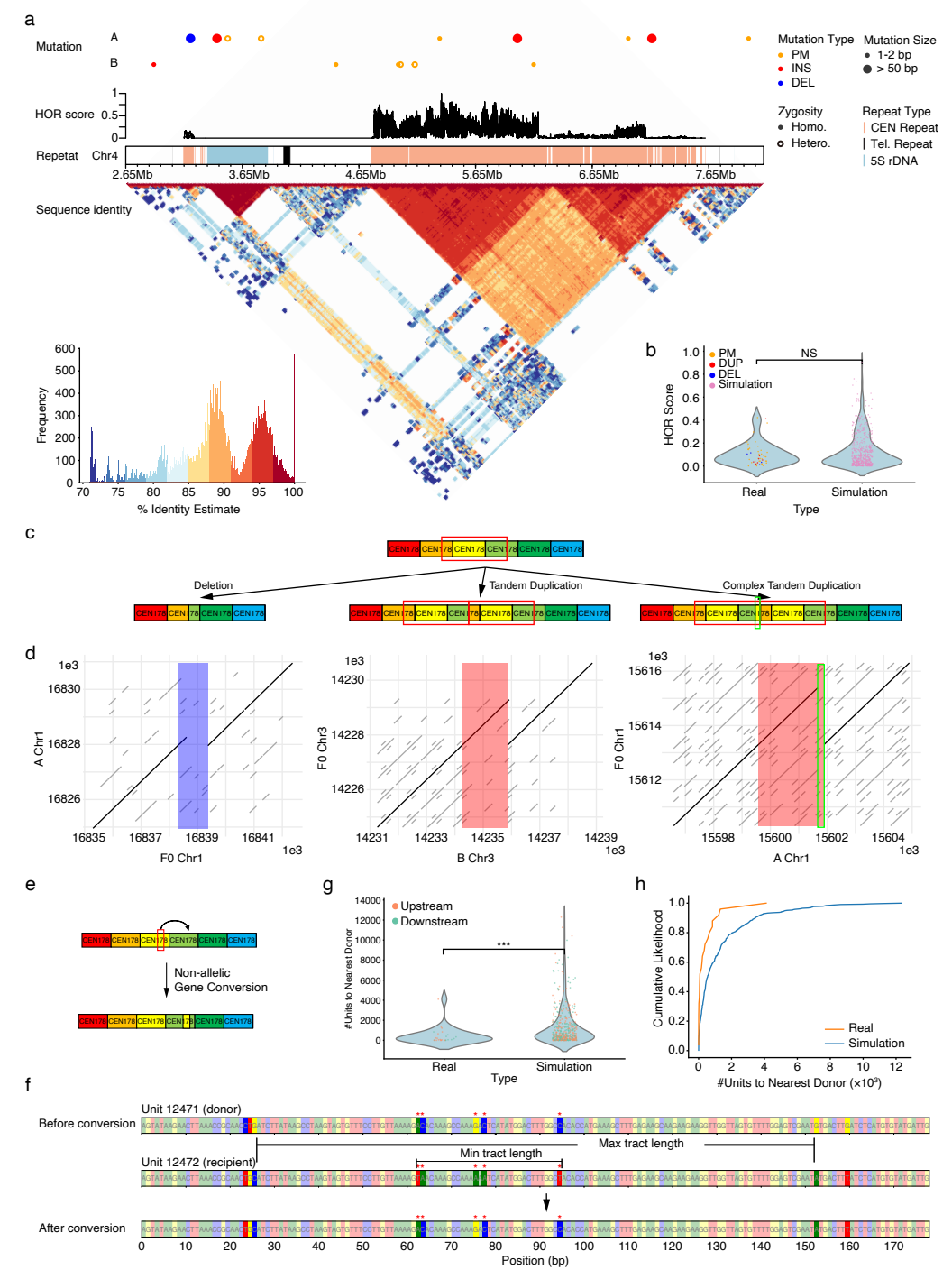
41 Li, Heng. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." *arXiv preprint arXiv:1303.3997* (2013).

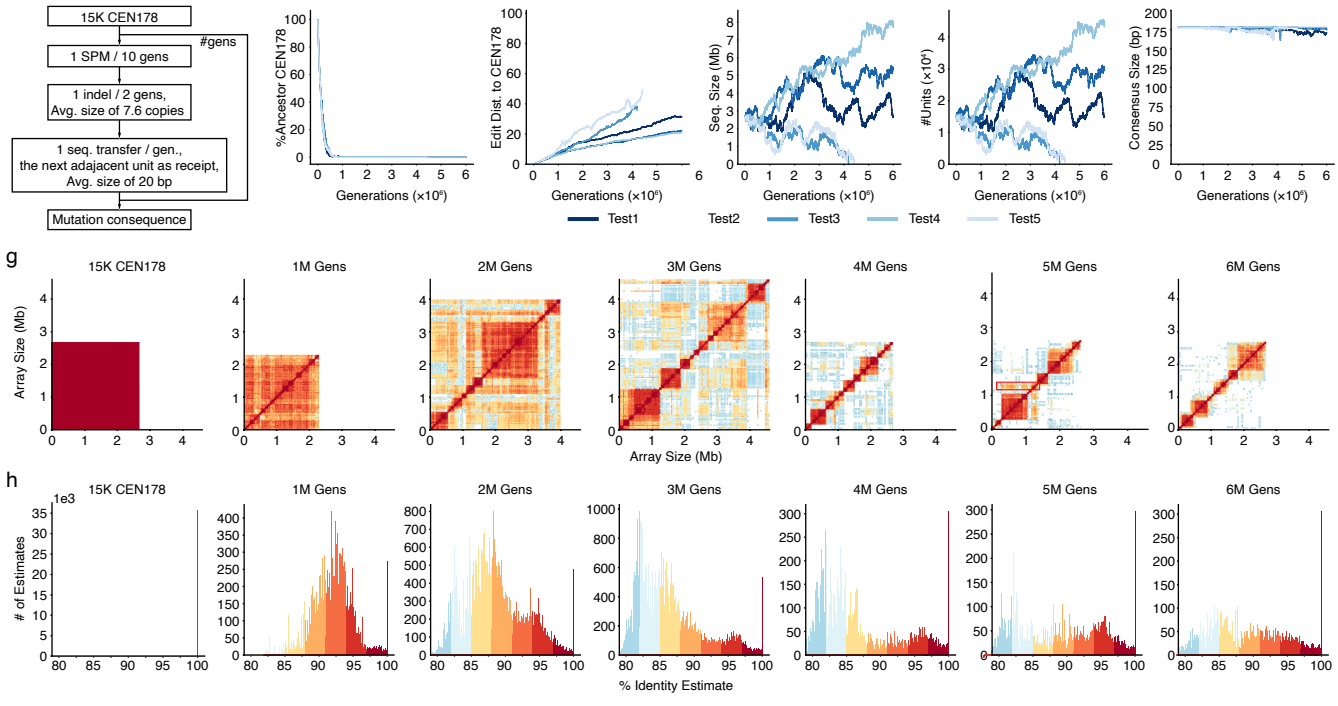
42 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

43 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26 (2011).

44 Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology* **14**, e1005944 (2018).


a
Col-0
F1
F16
A
B
A1
A2
B1
B2
c
A
B
Count
40
30
20
10
0
>50
11–50
3–10
2
1
DEL
PM
INS
5S rDNA
Centromere
Arm
b
A
B
Chr1
Chr2
Chr3
Chr4
Chr5
1b
10Mb
20Mb
30Mb
Mutation Type
PM
INS
DEL
Mutation Size
1-2 bp
3-50 bp
> 50 bp
Zygosity
Homo.
Hetero.
Repeat Type
CEN Repeat
Tel. Repeat
MT Insertion
5S rDNA
45S rDNA


a
Mutation
A
B
Mutation Type
PM
INS
DEL
Mutation Size
1-2 bp
> 50 bp
HOR score
Zygosity
Homo.
Hetero.
Repeat Type
CEN Repeat
Tel. Repeat
5S rDNA
Repetat
Chr4
2.65Mb
3.65Mb
4.65Mb
5.65Mb
6.65Mb
7.65Mb
Sequence identity
Frequency
% Identity Estimate
b
PM
DUP
DEL
Simulation
NS
HOR Score
Real
Simulation
Type
c
Deletion
Tandem Duplication
Complex Tandem Duplication
CEN178
d
A Chr1
F0 Chr1
F0 Chr3
B Chr3
A Chr1
e
Non-allelic
Gene Conversion
g
Upstream
Downstream
***
#Units to Nearest Donor
Real
Simulation
Type
h
Cumulative Likelihood
Real
Simulation
#Units to Nearest Donor (×10³)
f
Unit 12471 (donor)
Before conversion
Unit 12472 (recipient)
Max tract length
Min tract length
After conversion
Position (bp)


a
15K CEN178
1 SPM / 10 gens
#gens
1 indel / 2 gens,
Avg. size of 7.6 copies
1 seq. transfer / gen.,
the next adajacent unit as receipt,
Avg. size of 20 bp
Mutation consequence
b
%Ancestor CEN178
Generations (×10^6)
c
Edit Dist. to CEN178
Generations (×10^6)
d
Seq. Size (Mb)
Generations (×10^6)
e
#Units (×10^3)
Generations (×10^6)
f
Consensus Size (bp)
Generations (×10^6)
Test1
Test2
Test3
Test4
Test5
g
15K CEN178
1M Gens
2M Gens
3M Gens
4M Gens
5M Gens
6M Gens
Array Size (Mb)
Array Size (Mb)
h
15K CEN178
1M Gens
2M Gens
3M Gens
4M Gens
5M Gens
6M Gens
1e3
# of Estimates
% Identity Estimate