Diploid origins, adaptation to polyploidy, and the beginning of rediploidization in allotetraploid *Arabidopsis suecica*

Robin Burns[1,*], Aboli Kulkarni[2], Anna Glushkevich[2], Uliana K Kolesnikova[2], Filip Kolář[3], Alison Dawn Scott[2], Polina Yu Novikova[2,*]

[1]Department of Plant Sciences, University of Cambridge, Cambridge, UK
[2]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, Cologne, 50829, Germany
[3]Department of Botany, Faculty of Science, Charles University, Benátská 2, Praha, CZ-128 00, Czech Republic
[*]co-corresponding authors: rb996@cam.ac.uk, pnovikova@mpipz.mpg.de

Abstract

Polyploidization or whole-genome duplication (WGD), followed by genome downsizing is a recurring evolutionary cycle with a significant impact on genome structure. Polyploidy originates from a cell cycle error and in species without a dedicated germline, mitotic as well as meiotic errors can be stably inherited. After the origin, polyploids face immediate challenges, particularly the need to adapt meiotic machinery to prevent improper chromosome pairing and segregation, such as multivalent formation. Over evolutionary timescales, surviving polyploids undergo rediploidization, reorganizing their genomes back to a diploid state. This study focuses on the evolution of a young allotetraploid *Arabidopsis suecica,* a hybrid species derived from *A. thaliana* and *A. arenosa*. Our genomic analyses reveal that *A. suecica* originated from diploid *A. arenosa*, contrary to earlier suggestions of a tetraploid origin, indicating that WGD in *A. suecica* likely occurred somatically rather than being inherited via unreduced gametes. Ecological niche modeling traces the geographic origin of *A. suecica* to the southeastern edge of the European glacier during the last ice age. We explore how *A. suecica* adapted to allopolyploidy from diploid ancestors and find evidence of positive selection on meiotic cell cycle genes in the *A. arenosa* sub-genome. Several of these genes overlap with those involved in autopolyploid adaptation in *A. arenosa*, but have distinct haplotypes. We also show that purifying selection is relaxed in *A. suecica*, however, functional compensation between homeologous gene pairs persists, with fewer gene pairs than expected having loss-of-function mutations in both copies. This is consistent with early signs of rediploidization just after ~16K years since the origin of *A. suecica*. Our study covers three key aspects of the evolution of allopolyploid *A. suecica*: origin, meiotic adaptation, and the beginning of rediploidization.

Author summary

Polyploidization, or whole-genome duplication (WGD), drastically reshapes genomes. Most eukaryotic genomes carry the scars of ancient polyploidy events, reflecting cycles of genome doubling followed by downsizing and eventual return to a diploid state. Here, we examine the cycle of polyploidy in *Arabidopsis suecica,* a hybrid polyploid (i.e. allopolyploid) of *A. thaliana* and *A. arenosa*. Starting at the origin, we reveal that the *A. arenosa* parent of *A. suecica* was diploid, not tetraploid as previously hypothesized. Given this, the most probable scenario is that two diploid parents formed a sterile diploid hybrid, which partially restored fertility via somatic WGD. Establishment of the new polyploid species requires further restoration of fertility via meiotic adaptation. We find signatures of positive selection in allopolyploid *A. suecica* acting on meiosis-related genes. Several of the meiosis-related genes overlap with those involved in meiotic adaptation in autotetraploid *A. arenosa*, but distinct alleles suggest an independent adaptation. Rediploidization concludes the evolutionary cycle of polyploidy, however, pinpointing the beginning of this process is extremely challenging. We show that despite relaxed purifying selection in *A. suecica* due to genome-wide genetic redundancy, selection acts to preserve one intact functional gene when the alternate gene is damaged in homeologous gene pairs. Within just 16,000 years, *A. suecica* appears to be undergoing early stages of rediploidization, with some genes already reverting to a single-copy functional state.

Introduction

Cycles of polyploidy, or recurrent whole genome duplications (WGDs), have profoundly shaped the genomes of most eukaryotes[1–3]. Particularly common in plants, allopolyploidy, the combination of WGD with interspecific hybridization, is associated with agricultural traits of interest[4,5] as well as tolerance to harsher climates[6]. Notably, key crops like wheat, cotton, *Brassica*, strawberry, peanut, tobacco, to name a few, are all allopolyploids[7–13]. The importance of allopolyploids requires an understanding of the complex series of events that lead to their origin and establishment, in addition to how their genomes are stabilized and restructured over long evolutionary time[14].

The origin of allopolyploids involves hybridization and WGD, and the timing of WGD—whether it occurs before or after hybridization—may significantly impact how the genome eventually stabilizes. For example, when WGD occurs after hybridization, as in *Primula kewensis[15]*, *Capsella bursa-pastoris[16]* and in F1 hybrids such as those between *Mimulus nelsoni* and *M. lewisii[17]* and between *A. thaliana* and *A. lyrata[18]*, the resulting neo-allopolyploid must adapt *de novo* to the challenges of orchestrating divergent genomes to perform basic cellular tasks. However, if WGD happens before hybridization via unreduced (2*n*) gametes, autopolyploid ancestors that are a natural source of 2*n* gametes may ease the transition to genome stability by providing alleles that are pre-adapted to the challenges of polyploidy.

One of the main challenges of polyploidy is the stabilization of meiosis, as the formation of multivalents can lead to a failure of proper chromosomal segregation[19,20]. Sub-genomic divergence helps mitigate the challenges of chromosome pairing and the risk of homeologous recombination[21–23], but it does not eliminate it entirely[19,20]. For instance, in *A. suecica*, with an estimated sub-genome divergence of 11.6%[24] recombination between sub-genomes and genome errors still frequently occur in the early generations of synthetic *A. suecica[24–26]*, highlighting that sub-genome divergence is insufficient to prevent meiotic errors on its

own. Autotetraploids also evolve to stabilize meiosis and avoid multivalent formation by redistributing and reducing the amount of crossovers per chromosome. This is a polygenic adaptation described in several species, and involves several genes related to the synaptonemal complex[27–29] but the extent to which these adaptations can benefit genome stabilization in neo-allopolyploids is unknown. Stable allotetraploids do appear to have evolved genetic control of sub-genome specific recombination. Notable examples include *Ph1* (*ZIP4*)[30] and *Ph2* (*MSH7-3D*)[31] loci in allohexaploid wheat, the *PrBn* locus in *Brassica napus[32]* and *BYS* in natural *A. suecica[33]*, all of which suppress homeologous recombination. On a transcriptional level, meiotic-related genes were also shown to be up-regulated on the *A. suecica A. thaliana* sub-genome[24].

Ancient WGDs in the ancestry of most diploids suggest that current polyploids will eventually return to a diploid-like state through a process known as rediploidization[34]. Rediploidization dramatically restructures polyploid genomes after WGD[35–40], as seen in ancient allopolyploids like maize[41] and current allopolyploids like broomcorn millet[42], and has contributed to distinct morphotypes in *Brassica[43]*. During rediploidization, relaxed purifying selection in allopolyploids allows mutations to accumulate in redundant gene copies, with functional homeologs compensating for non-functional homeologs, potentially marking the early stages of genome restructuring and downsizing, as observed in *Brachypodium hybridum[44]*. Additionally, genome restructuring can favor faster fractionation (gene loss) in one sub-genome over another[41,45], possibly due to differences in transposable elements and inherited expression dominance[46]. However, genome dominance is notably absent in several natural allopolyploids[24,42,44,47]. Gene function may also affect fractionation rates, with previous studies showing that genes related to genome integrity and organelle function revert to single copies faster than those involved in signaling, transport, and metabolism, likely due to gene dosage constraints[48].

The allotetraploid *A. suecica*, a hybrid between *A. thaliana* and *A. arenosa*, is an excellent model for studying the cycle of allopolyploid evolution: origin, establishment, and the beginning of rediploidization. Initially, it was hypothesized that *A. suecica* originated from the fusion of an unreduced egg from diploid *A. thaliana* with a 2*n* sperm from tetraploid *A. arenosa[49]*. However, recent evidence from sequence similarity of meiotic genes points to a diploid *A. arenosa* parent[50]. Knowing which cytotype contributed to the origin of *A. suecica* will help our understanding of the genetic basis underlying the evolution of a stable allotetraploid genome in natural *A. suecica[24]*. We use the recent genome assembly of *A. suecica[24]* and expanded sequence data from both parental lineages, including the full diversity of *A. arenosa* cytotypes[24,36,51–53] to reconstruct the evolutionary history of *A. suecica* and predict the parental cytotypes. To explore how *A. suecica* adapted to the allopolyploid cell state, we compare its sub-genomes with the defined parental lineages from our reconstruction and identify genes under selection. We notice an overlap between genes under selection in allotetraploid *A. suecica* and autotetraploid *A. arenosa* and explore if the haplotypes under selection are shared or independently evolved. Finally, we investigate whether *A. suecica*, an allotetraploid formed around 16 Kya, shows early signs of rediploidization. Specifically, we analyzed the distribution of inherited and *de novo* loss-of-function (LoF) mutations across its sub-genomes to see if they are negatively correlated. In other words, when one homeolog (e.g., the *A. thaliana* gene copy) has a damaging mutation, does the other (e.g., the *A. arenosa* gene copy) remain intact and functional? Such functional compensation would indicate the onset of rediploidization, and potentially reveal whether there is a non-random progression in this final stage of the polyploidization cycle.

Results

Genetic origins of *A. suecica*

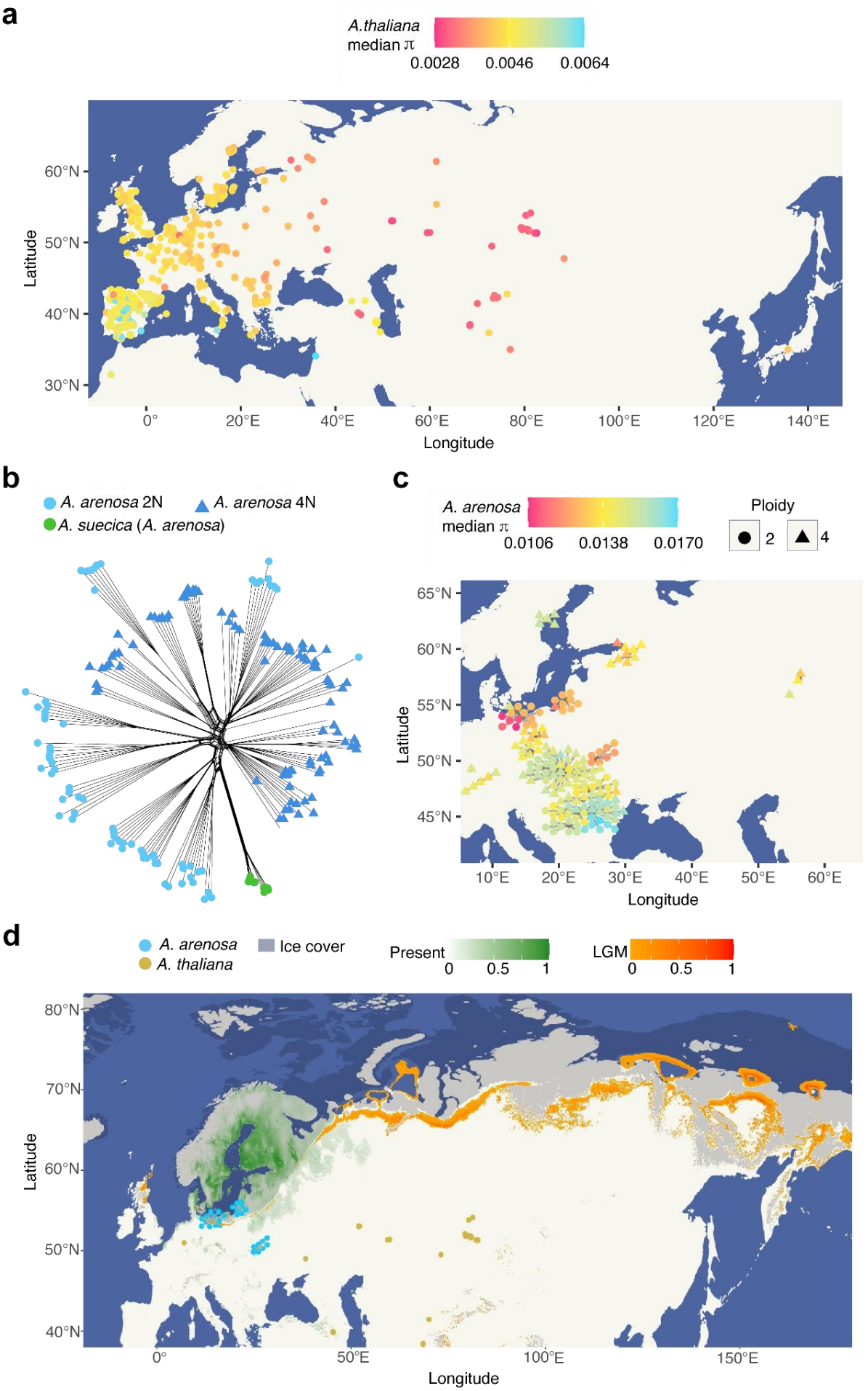
To reconstruct the evolutionary history of allotetraploid *A. suecica*, we first aimed to identify the closest parental accessions, using available data from the parental species, *A. thaliana* and *A. arenosa*, covering their vast geographic range. We mapped 970 *A. thaliana* accessions[53] and 163 *A. arenosa* individuals (93 tetraploids and 70 diploids), representing spatially close lineages of both diploid and autotetraploid cytotypes[36,51,54], to the *A. thaliana* and *A. arenosa* sub-genomes of *A. suecica[24]*. Genomic positions were conservatively filtered to ensure reliable variant calling (see Materials and Methods). In summary, callable sites included ~62% (73,813,230) of the *A. thaliana* sub-genome and ~34% (47,950,105) of the *A. arenosa* sub-genome, with 6,125,591 biallelic SNPs in *A. thaliana* and 9,664,305 in *A. arenosa.* Throughout this study, results generated for the sub-genomes of *A. suecica* are compared with corresponding data from the parental species*, A. thaliana* and *A. arenosa.*

Previously, it was shown that the closest *A. thaliana* accessions to *A. suecica* are found in Central Eurasia[49]. Here, eliminating the possibility of a reference bias artifact, we calculated median pairwise diversity (π) for each *A. thaliana* accession to our 16 *A. suecica* natural accessions and confirmed that result (Figure 1A).

To identify the *A. arenosa* individuals most closely related to *A. suecica* and to determine their cytotype (diploid or tetraploid), we expanded the previously released *A. arenosa* dataset[36,54]. We sequenced five additional autotetraploid *A. arenosa* individuals sampled from herbarium specimens collected in the southern Urals, representing the easternmost populations and those geographically closest to the parental *A. thaliana* accessions. The ploidy of most of the *A. arenosa* individuals was previously determined using flow cytometry[36,54]. For the additional five herbarium individuals, ploidy was inferred using the software nQuire[55] and counting the number of different *S*-alleles (see Materials and Methods). The combination of these two methods has been used previously to determine ploidy in *A. lyrata[56]*. We conducted a network analysis based on genetic distances (Nei's D[57]) that are suitable for mixed-ploidy datasets[58] (Figure 1B). We also calculated the median pairwise nucleotide diversity (π) for each *A. arenosa* individual against our 16 *A. suecica* natural accessions and placed it on a geographical map (Figure 1C). Our results indicate that diploid *A. arenosa* from southern Baltic sea coast[59] is genetically the closest to *A. suecica*, and certainly closer than any autotetraploid *A. arenosa*, including both the current sympatric tetraploid Ruderal lineage growing in Fennoscandia[36] and tetraploid populations from the southern Urals, i.e. the geographic region closest to putative ancestral *A. thaliana* lineage (Figure 1A). A full list of *A. thaliana* and *A. arenosa* accessions and their π values to *A. suecica* is available in Supplementary File 1. The vast majority of the SNPs (>99%) in *A. suecica* are shared with *A. arenosa* and *A. thaliana,* such ubiquitous SNP sharing supports the claim that multiple individuals of *A. arenosa* and *A. thaliana* contributed to the origin of *A. suecica[49]* (Supplementary Figure 1).

Altogether our results suggest a diploid rather than tetraploid origin for the *A. arenosa* sub-genome within *A. suecica*. This is in agreement with data showing that meiotic genes on *A. arenosa* sub-genome of *A. suecica* are diploid-like rather than tetraploid-like [50]. And it supports the original hypothesis by Hylander in 1957[60] that *A. suecica* originated from hybridization between diploid *A. thaliana* and diploid *A. arenosa*, with autotetraploid *A. arenosa*

later spreading across Fennoscandia, while the diploids remain mostly restricted to southeastern Europe and isolated Baltic areas[59].


a
A.thaliana
median π
0.0028
0.0046
0.0064
60°N
50°N
40°N
30°N
Latitude
0°
20°E
40°E
60°E
80°E
100°E
120°E
140°E
Longitude
b
A. arenosa 2N
A. arenosa 4N
A. suecica (A. arenosa)
c
A. arenosa
median π
0.0106
0.0138
0.0170
Ploidy
2
4
65°N
60°N
55°N
50°N
45°N
Latitude
10°E
20°E
30°E
40°E
50°E
60°E
Longitude
d
A. arenosa
A. thaliana
Ice cover
Present
0
0.5
1
LGM
0
0.5
1
80°N
70°N
60°N
50°N
40°N
Latitude
0°
50°E
100°E
150°E
Longitude

179

Figure 1: The genetic and geographic origin of *A. suecica*. a Median nucleotide diversity (π) of 970 *A. thaliana* accessions to 16 *A. suecica* accessions shows Central Eurasian *A. thaliana* are genetically closest **b** A genetic distance network (Nei's D) shows that *A. suecica* (green circles) is genetically closer to diploid *A. arenosa* (light blue circles) than autotetraploid *A. arenosa* (dark blue triangles). **c** Median nucleotide diversity (π) of *A. arenosa* individuals to 16 *A. suecica* accessions indicates that Southern Baltic *A. arenosa* diploids (red circles) are the genetically closest. Locations are represented by grey lines connecting *A. arenosa* individuals to their geographic origins.. **d** Niche modeling of *A. suecica* shows probability of occurrence for the current climate (green) and the Last Glacial Maximum (LGM) (orange). Closest parental populations include Southern Baltic and Ukrainian diploid *A. arenosa*, and the top 5% of *A. thaliana* accessions genetically closest to *A. suecica*.

Geographic origins of *A. suecica*

The origin of *A. suecica* is estimated to be around 16K years ago[49], when Fennoscandia, where *A. suecica* is currently distributed [50], was sheathed in ice. The genetically closest *A. thaliana* accessions to *A. suecica* are located in Central Eurasia (Figure 1C) and the closest *A. arenosa* individuals are located in the Southern Baltics (Figure 1B) To clarify the geographic origin of *A. suecica*, we conducted a niche-modeling analysis to compare the current distribution with the likely geographic range of *A. suecica* at the time of origin.

We used the occurrence data for *A. suecica* and corresponding climatic variables to predict the potential distribution of the species across Eurasia (see Methods). We included all of Eurasia to fully capture the geographic distribution of both parental species. The model showed that the present climate niche aligns well with the current distribution of *A. suecica*, which is restricted to Fennoscandia (Figure 1C, green layer).

Next, we projected the current niche of *A. suecica* onto the climate of the Last Glacial Maximum (LGM), the estimated time of the origin for *A. suecica[49]*, assuming that the niche remained similar. The projection resulted in a very narrow distribution following the edge of the ice sheet (Figure 1D, orange layer), which could be suitable for *A. suecica.* This projected niche partially overlaps with the current distribution of the diploid *A. arenosa* that is genetically closest to *A. suecica,* near the southern Baltic Sea coast (Figure 1C, red circles). While the location of the genetically closest *A. thaliana* accessions is far from the Baltics (Figure 1C, red circles), several *A. thaliana* accessions, included in the top 5% of the genetically closest to *A. suecica*, are located relatively near the southern Baltic region (Figure 1D, light brown circles, Figure 1A red circles).

Based on these findings we suggest that the likely geographic origin for *A. suecica* was near the southern Baltic coast during the last ice age, after which *A. suecica* migrated to Fennoscandia. The movement of *A. thaliana* eastwards is also consistent with the demographic post-glacial history of the species[61].

Genomic adaptation to allopolyploidy in *A. suecica*

Given that the genetically closest parents of *A. suecica* were diploids, *A. suecica* could not inherit any of the potentially pre-adaptive meiotic alleles from autotetraploid *A. arenosa*. Previously, we reported that meiosis-related genes on the *A. thaliana* sub-genome of *A. suecica* are up-regulated compared to parental transcriptional levels[24]. However, in the same study, we did not observe significant upregulation of meiosis-related genes on the *A. arenosa* sub-genome. Notably, the latter transcriptional comparison was made against

tetraploid *A. arenosa*, the previously assumed parent of *A. suecica*, whereas we now establish here that diploid *A. arenosa* is closer. Therefore, to uncover how *A. suecica* adapted to allopolyploidy from diploid ancestry, we searched for signatures of selection on genes, performing McDonald-Kreitman test between *A. suecica* and the identified closest progenitors (see Methods). Due to very low genetic diversity within *A. suecica*, in our gene selection scans we prioritized genes with higher non-synonymous than synonymous divergence and an absence of non-synonymous polymorphisms in *A. suecica*.

We identified signatures of selection on 35 genes in *A. thaliana* and 585 genes in the *A. arenosa* sub-genomes of *A. suecica* (Supplementary File 2). Due to the low number of founders contributing to the origin of *A. suecica[49]*, large parts of the genome, depleted of genetic diversity, are hard to distinguish from selection sweeps. However, the genes we identified with signatures of selection are also located outside of low-diversity regions (Supplementary Figure 2) and importantly are associated with specific gene ontology (GO) terms (Figures 2A and Supplementary File 2). This suggests that the observed selection signatures likely reflect genomic adaptations in *A. suecica*.

Genes under selection in the *A. thaliana* subgenome of *A. suecica* are enriched for several GO terms, notably including *polar nucleus fusion* (GO:0010197) and *positive regulation of DNA endoreduplication* (GO:0032877) (Figure 3A). In the *A. arenosa* subgenome, enrichment for *polar nucleus fusion* (GO:0010197) also appears, but involves different genes. Signatures of selection on the *A. arenosa* subgenome reveals GO terms enriched for *mitotic cohesin loading* (GO:0061780), suggesting genetic adaptations in *A. suecica* for chromosome cohesion during cell division—a key factor for genomic stability in polyploids[28]. Among the cohesin-related genes, *SISTER CHROMATID COHESION2* (*SCC2*) encodes adherin, a protein essential for cohesin loading onto chromosomes[62]. We also find enrichment for the *meiotic cell cycle* (GO:0051321), suggesting genetic adaptations in *A. suecica* for basic meiotic cell functions (Figure 3A). Genes identified as under selection on the *A. thaliana* sub-genome in our selection scans here, do not overlap with the transcriptionally upregulated meiosis-related genes reported in our previous study[24].

In the *A. arenosa* subgenome of *A. suecica*, *SCC2*, along with *ATRANGAP1*, *UBIQUITIN PROTEIN LIGASE 4*, *FRAGILE FIBER 1*, and *CYCD5*, are meiotic genes identified in our selection scan and implicated in polyploid adaptation[27][54].The limited overlap between selected genes in *A. suecica* and tetraploid *A. arenosa*, likely reflects the complexity of polyploid adaptation. To investigate whether the modest overlap was due to shared alleles or independent evolution, we performed phylogenetic analyses.

The analyses revealed no shared haplotypes between *A. suecica* and autotetraploid *A. arenosa*. Instead, *A. suecica* consistently clustered with diploid *A. arenosa* for the five genes examined (e.g., *SCC2* in Figure 3B; Supplementary Figure 3). Notably, amino acid variants distinguishing *A. suecica* from diploid *A. arenosa* were absent in both tetraploid *A. arenosa* and the outgroup *A. thaliana*, suggesting that these polymorphisms arose independently in *A. suecica* (e.g., *SCC2* variants, Figure 3C). We did not observe fixed amino acid differences between diploid and tetraploid *A. arenosa* in *SCC2*. This gene, along with other proteins involved in sister chromatid cohesion, has been shown to be under selection in specific diploid and tetraploid lineages of *A. arenosa* [63][64], emphasizing that the specific populations analyzed in diploid-tetraploid *A. arenosa* selection scans can influence which genes appear under selection. In our comparison with *A. suecica*, we included all available diploid lineages. Furthermore, selection scans comparing *A. suecica* to both diploid and tetraploid *A. arenosa* consistently identified the same meiotic genes as under selection in *A. suecica*

(Supplementary File 2). This consistency highlights the potential importance of these genes in the adaptation of *A. suecica*.

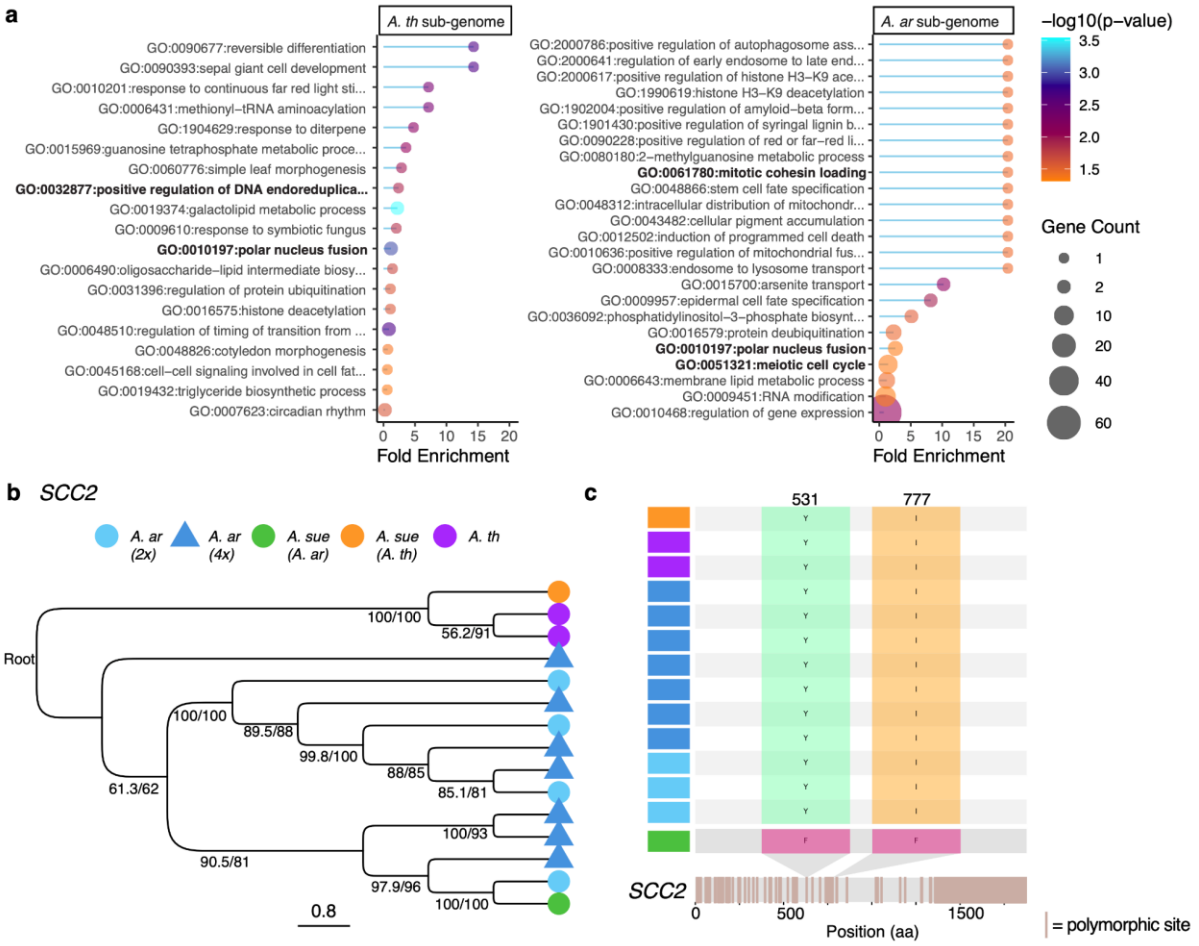


Figure 2: Genomic adaptation to allopolyploidy in *A. suecica* a GO enrichment for the 35 genes on the *A. thaliana* sub-genome and the 585 genes on the *A. arenosa* sub-genome. GO terms discussed in the text are highlighted in bold. **b** Phylogenetic tree from protein alignment of SCC2 showing that the *A. arenosa* homeolog in *A. suecica* clusters with diploid *A. arenosa*. Branch length is not drawn to scale to improve the analysis of the tree topology. Both the translated genomic and CDS sequence of the *SCC2* from *A. thaliana,* in addition to the *A. suecica A. thaliana* homeolog were used to root the tree. **c** Fixed amino acid differences in the *SCC2* between the *A. suecica A. arenosa* sub-genome and diploid *A. arenosa*. Tetraploid *A. arenosa* and *A. thaliana* share the same amino acid as diploid *A. arenosa*, the difference is unique to *A. suecica.*

Early signs of rediploidization in *A. suecica*

Polyploid organisms have a substantial amount of genetic redundancy, which can lead to relaxed purifying selection. This pattern of relaxed selection has been observed in autopolyploid *A. arenosa[38]* and allopolyploid *A. kamchatica[65]*. We examined *A. suecica* for similar signs of relaxed purifying selection and the potential implications this has for the early stages of rediploidization.

Analyzing over 3.5 million SNPs in the *A. arenosa* sub-genome and 1.2 million SNPs in the *A. thaliana* sub-genome, we found that both sub-genomes in *A. suecica* are evolving under purifying selection, however purifying selection appears more relaxed in *A. suecica*

compared to the parental species (Figure 3A). This relaxation in selection pressure likely represents the early stages of rediploidization. During this phase reduced selective constraints following polyploidization enables mutational robustness between the sub-genomes, since a LoF mutation in one sub-genome can be compensated by its homeolog on the alternate sub-genome.

To investigate the compensatory relationship between homeologs we identified 414 homeologous gene pairs with at least one gene carrying a LoF mutation. We observe a slight bias towards *A. arenosa* homeologs (269 vs 142). However, the majority of the LoF mutations are also segregating in the parental species and, therefore, are most probably inherited rather than evolved *de novo* after hybridization (Figure 3B). The distribution of *de novo* LoF mutations (mutations appearing after the allotetraploidization) in homeologous gene pairs was more balanced (49 in *A. arenosa*, 62 in *A. thaliana*), suggesting no sub-genome bias in the post-polyploidisation distribution of LoF mutations (Figure 3B). To investigate if these LoF homeologs overlap with genes identified in our selection scans, we compared the genes involved. However, we find little overlap between the gene lists (Supplementary Figure 4), indicating that a LoF occurring on one homeolog is not driving the signals of positive selection we observed.

We next asked if the distribution of LoF mutations between homeologs can be explained by selection acting to preserve at least one functional copy. By comparing the observed LoF distribution between homeolog pairs with randomly paired homeologs, we found significantly fewer cases in *A. suecica* where both homeologs carry LoF mutations than expected (Figure 3C and D, $p < 0.05$). This indicates that *A. suecica* tends to retain at least one functional homeolog, despite an overall relaxed selection pressure—a pattern also supported by in-silico simulations that showed dual LoF mutations occur less frequently in *A. suecica* than expected from random combinations of *A. thaliana* and *A. arenosa* genomes (Figure 3C and D, $p < 0.05$). These results align with the hypothesis that early stages of rediploidization in *A. suecica* involves selective pressures that favor maintaining at least one intact homeolog, helping to preserve essential gene functions. Notably, which homeolog remains functional is not biased toward one sub-genome.

Finally, we assessed whether dosage constraints on single-copy genes impact homeolog retention. If dosage constraints were at play, genes that evolve towards single-copy levels[48] would be overrepresented in our dataset, and LoF SNPs would show higher population frequencies in single-copy genes in *A. suecica* when compared to genes that exist as multi-copies. However, we observed no significant enrichment of single-copy genes among homeologs carrying a LoF mutation, nor do we see that LoF mutations within single copy genes exist at higher population frequencies than LoF mutations within multi-copy genes in *A. suecica* (Supplementary Figure 5). This suggests that dosage constraints are not a major driver of homeolog rediploidization at this stage of the genome fractionation process.

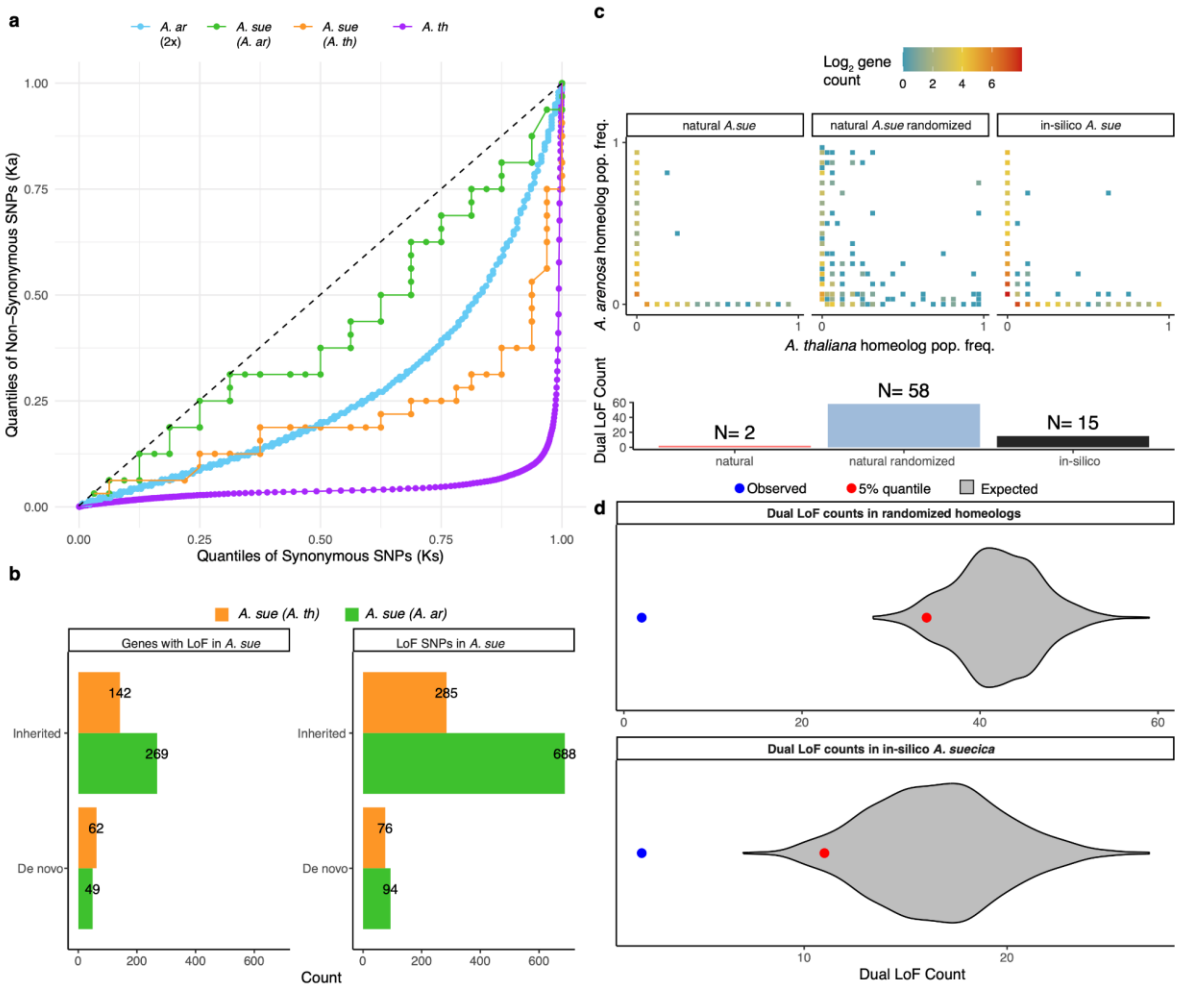


Figure 3: Relaxed Purifying Selection and early rediploidization of homeologs in *A. suecica*
a QQ plot of synonymous (Ks) and non-synonymous (Ka) SNP frequencies for *A. suecica* and its progenitors (*A. thaliana* and diploid *A. arenosa*). The dotted line shows neutral evolution. All species experience purifying selection, with *A. suecica* sub-genomes showing relaxed purifying selection relative to parents. **b** Bar plots show genes and total counts of LoF SNPs in *A. suecica*, the majority being inherited from the parents. **c** Heatmaps of homeolog pairs with at least one LoF SNP, with population frequencies for *A. thaliana* and *A. arenosa* homeologs on the x- and y-axes, respectively. Observed values for natural *A. suecica* are compared to shuffled and in-silico controls (random combinations of *A. thaliana* and *A. arenosa* individuals). **d** Comparison of natural *A. suecica* with 1,000 shuffled and in-silico controls, showing that *A. suecica* (blue dot) has fewer dual LoF SNPs in homeolog pairs than expected by chance (gray violin plot), a difference that is statistically significant (red dot denotes p-value = 0.05).

Discussion

A. suecica is a relatively young allopolyploid species that originated from the hybridization of *A. thaliana* and *A. arenosa*. Using population-level DNA sequencing for all three species, we revisited the genetic origins of *A. suecica* to understand the nature of WGD leading to the formation of the allopolyploid species and if the ancestry shaped adaptation to allopolyploidy.

Leveraging the recent genome assembly of *A. suecica[24]*, we identified the genetically closest *A. thaliana* and *A. arenosa* individuals to *A. suecica*. Our analysis also supports the conclusion that *A. suecica* had multiple origins from multiple hybridizations between *A.*

thaliana and *A. arenosa[49]*. For the *A. thaliana* sub-genome, our findings align with previous reports[49], and by using the *A. suecica* genome for SNP calling, we minimized reference bias. Importantly, we find that the *A. arenosa* individuals closest to *A. suecica* are diploid, not tetraploid, a key insight into the origin of *A. suecica*. Given that both parents were diploid, we propose that WGD in *A. suecica* occurred via somatic genome doubling post-hybridization rather than the fusion of unreduced gametes that were produced independently during meiosis in the parents. Similarly, another allotetraploid in the *Arabidopsis* genus, *A. kamchatica*, a hybrid between *A. lyrat*a and *A. halleri*, is genetically more closely related to diploid rather than autotetraploid *A. lyrata[56,66]*. This suggests that WGD occurring via somatic doubling in a diploid hybrid could be a common mechanism of allopolyploid formation.

Based on the revised ancestry, we could identify selection signatures on multiple genes in *A. suecica*, which showed overrepresentation in particular categories (Figure 2). It seems that polar nucleus fusion may play a critical role in reproductive success and endosperm development in polyploids. Notably, in equal ploidy crosses made in the lab between *A. thaliana* and *A. arenosa*, ~95% of the seeds are aborted, and one of the main characteristics of the hybrid seed failure is endosperm overgrowth[67]. The role of endoreduplication in polyploid adaptation remains less clear; while endoreduplication may increase cell size and enhance cellular function, its impact on polyploid fitness is uncertain. However, a significant reduction in endoreduplication frequency has been observed in natural *A. arenosa* autopolyploids[68] and synthetic *A. thaliana* autopolyploids[69], hinting at a possible cellular trade-off between polyploidization and endoreduplication. We also found that genes under selection in *A. suecica* are involved in meiosis, which reinforces the idea that divergence of the subgenomes alone can not guarantee the faithful disomic inheritance, and meiotic adaptation is still required. While meiotic genes were previously found to be upregulated on the *A. thaliana* subgenome of *A. suecica* by us[24], our selection scans of this sub-genome did not identify GO terms related to meiosis. This suggests that adaptation to polyploidy likely also involves regulatory changes rather than solely changes to the protein of a gene. Future studies that will also use the correct *A. arenosa* diploid parent will be crucial to disentangle the contributions of cis-regulatory changes versus protein-level changes for adaptation to polyploidization.

Our haplotype analysis of genes in the *A. arenosa* sub-genome of *A. suecica*—overlapping with genes identified in previous selection scans between autopolyploid and diploid *A. arenosa[27][54]*—suggests that adaptation to allopolyploidy in *A. suecica* involved many of the same genes, but with distinct, independent alleles. Additionally, amino acid changes in *A. suecica* differ from those both tetraploid from diploid *A. arenosa*, indicating independent evolution in *A. suecica*.

The limited contribution of autopolyploids to the origin of allopolyploids raises questions about their suitability as allopolyploid parents. Despite harboring alleles pre-adapted to polyploidy, we propose that autopolyploids may be constrained in two key ways. First, autopolyploids with four homologous chromosome copies can accumulate deleterious mutations due to relaxed purifying selection[38,70]. When a 2*n* gamete from an autopolyploid contributes to an allopolyploid, these mutations may become exposed and, if not fully compensated by a homeolog, could lead to reduced fitness and increased extinction risk. Second, autopolyploids like *A. arenosa* exhibit altered meiotic crossover patterns[71]. Whether these altered crossover patterns are beneficial or reversible or could be disruptive for homeolog segregation in allopolyploids remains unknown, but it may further reduce their viability as allopolyploid parents.

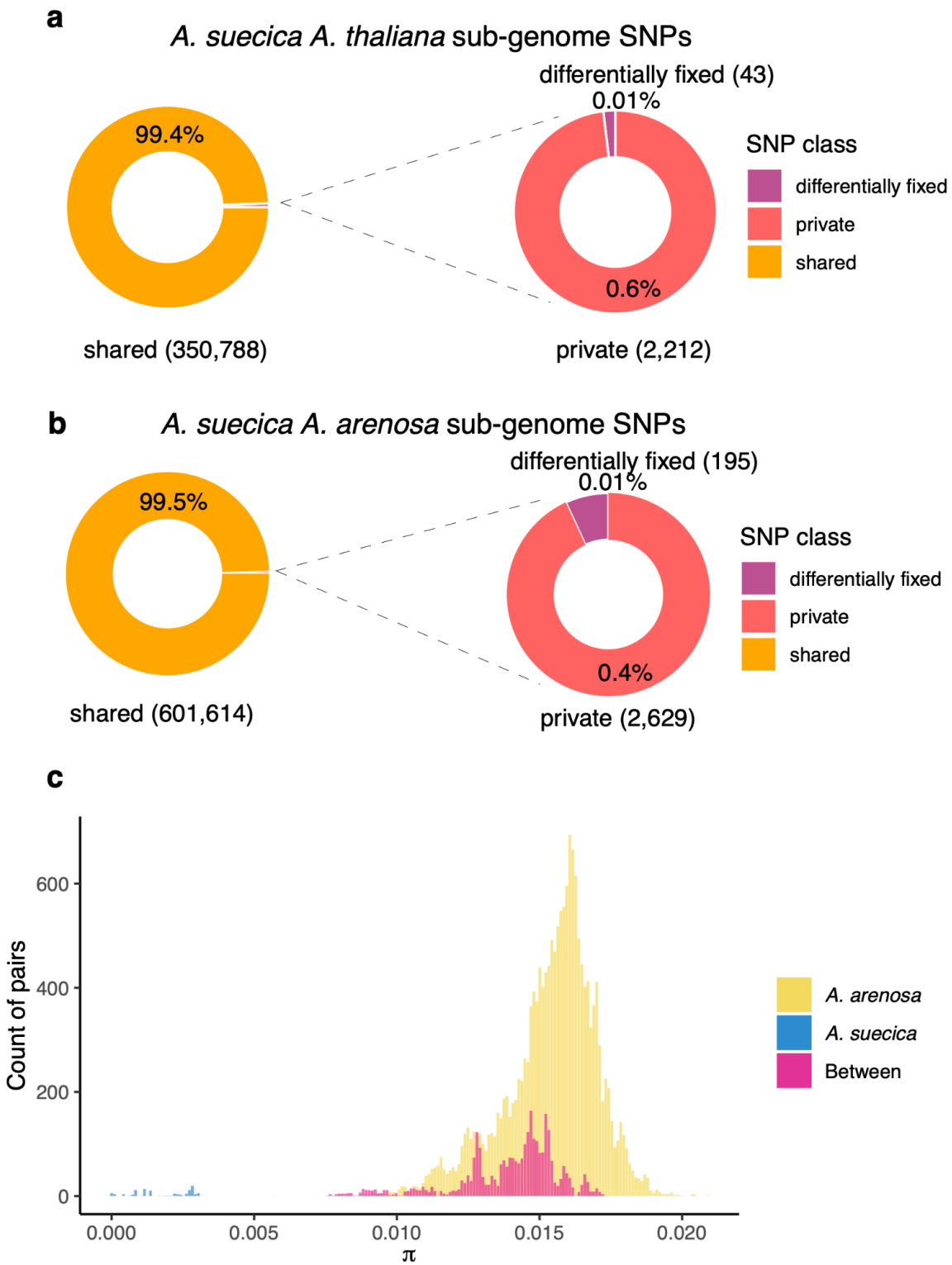
We find evidence of relaxed purifying selection in *A. suecica*, consistent with the idea that polyploidy confers increased mutational robustness through genetic redundancy. Previous studies have shown that diploid *A. thaliana* is more sensitive to mutagenesis than both natural and synthetic *A. suecica[37]*, supporting the claim that *A. suecica* has enhanced mutational robustness compared to its diploid parents.

We observed bias in *A*. *suecica* toward LoF mutations in the *A*. *arenosa* sub-genome, however, most of these LoF mutations also segregate in *A. arenosa* and are most probably inherited. This is likely due to the high genetic diversity and outcrossing nature of *A. arenosa*, which allows it to better mask deleterious mutations compared to the highly homozygous and predominantly selfing *A. thaliana*. *De novo* LoF mutations did not show subgenome bias, consistent with our previous study, showing absence of subgenome dominance on the transcription level[24].

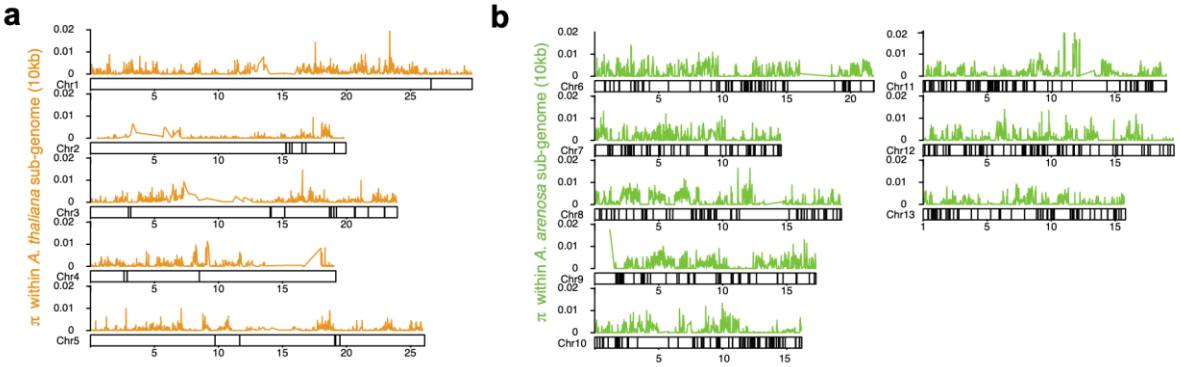
Our analysis of LoF mutations in *A. suecica* also reveals a robust gene compensatory mechanism, where the retention of at least one functional homeolog mitigates the impact of LoF mutations. Importantly when a LoF mutation occurs in one copy, the intact homeolog is effectively reverted back to a diploid-like state and will likely remain refractory throughout the genome downsizing processes, in order to maintain the gene function. In agreement, studies in yeast have shown that while additional gene copies do enhance mutational robustness, they are prone to rapid inactivation by LoF mutations rather than gradual subfunctionalization between duplicate genes[72].

We observe no bias toward one sub-genome accumulating more *de novo* LoF mutations since alloploidization, nor any overrepresentation for dosage-sensitive single-copy genes in homeologs that have been functionally rediploidized. These findings suggest that early stages of rediploidization may be primarily shaped by the random distribution of mutations across sub-genomes, with selection acting post-mutation to retain essential gene functions. As a result, *A. suecica* exhibits mutational robustness and maintains gene function as the genome gradually evolves back toward a diploid state—a process that will unfold over many generations.

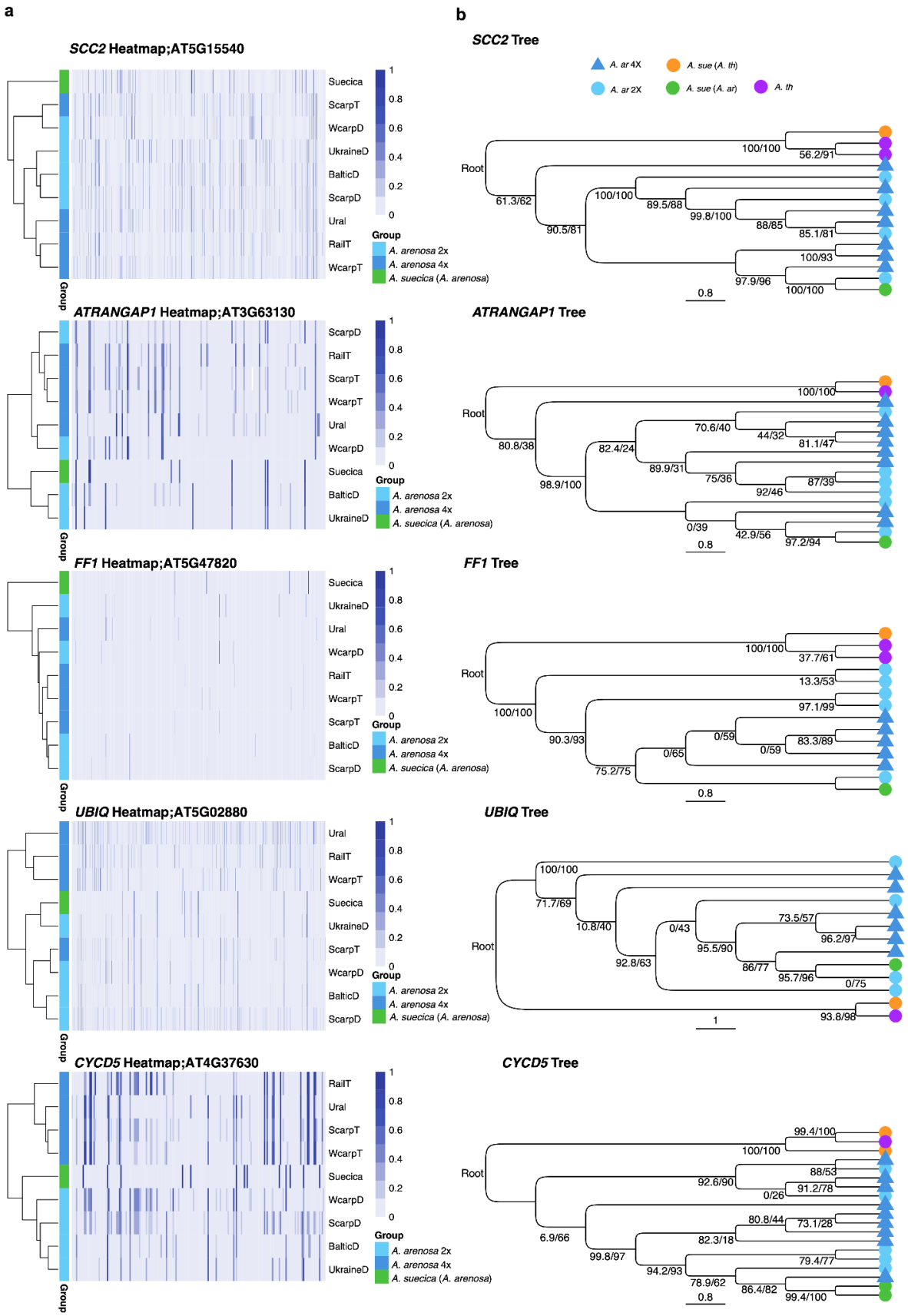
Supplemental Figures



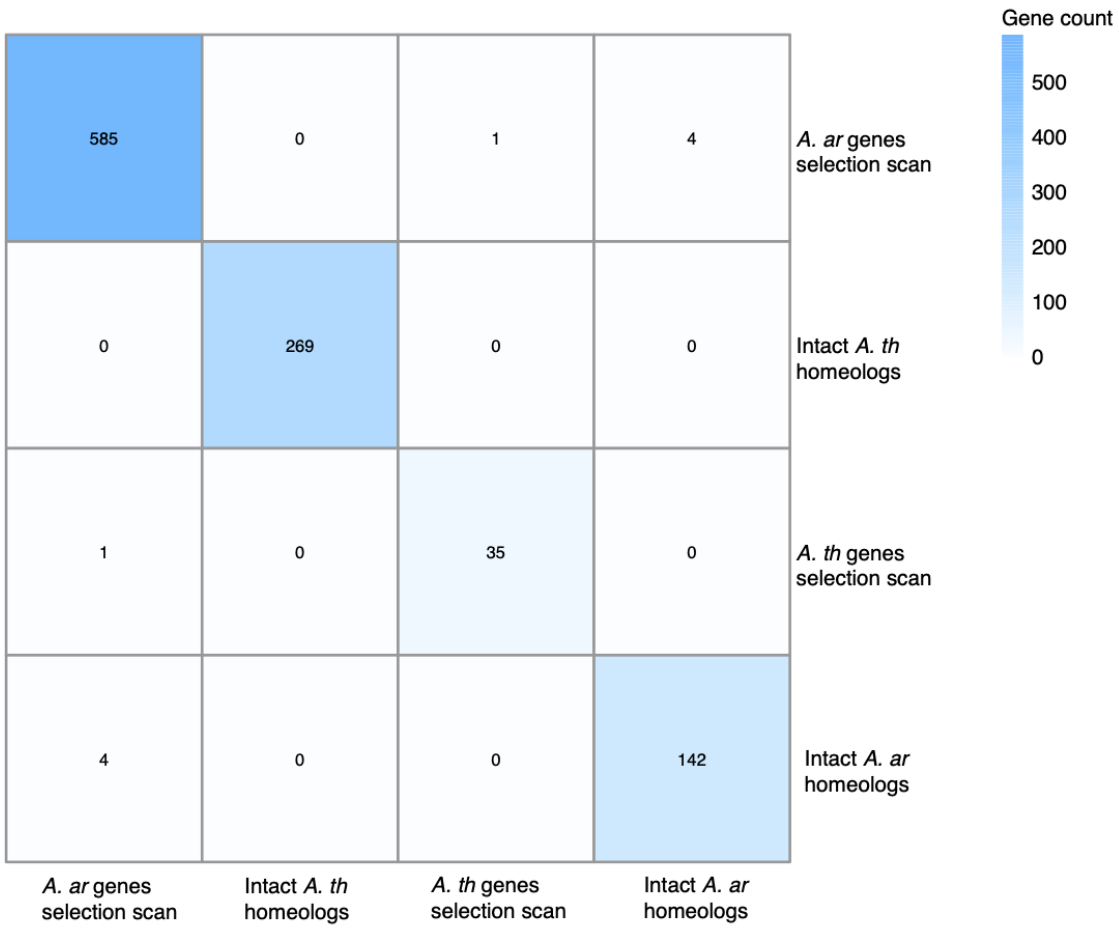
Supplementary Figure 1: SNP Sharing and Nucleotide Diversity (π) in *A. suecica*
a, b Over 99% of *A. suecica* SNPs are shared with progenitor species (*A. thaliana* and *A. arenosa*), indicating contributions from multiple individuals of both progenitors. **c** Nucleotide diversity (π) is significantly lower in *A. suecica* (blue) compared to *A. arenosa* (yellow) due to the allopolyploidy bottleneck and self-compatibility. In contrast, the obligate outcrosser *A. arenosa* displays greater diversity within species than between species (pink).



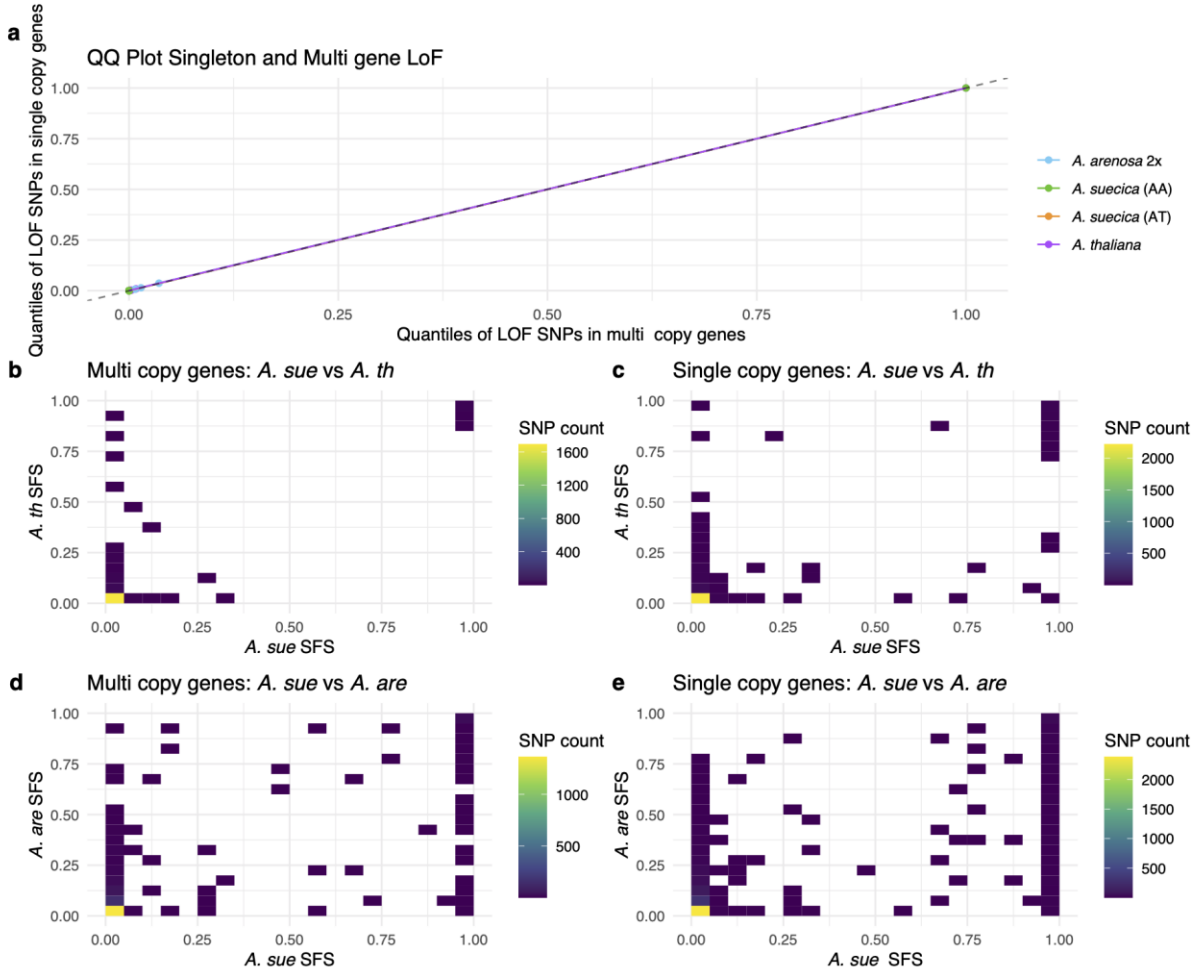
Supplementary Figure 2 a In the *A. thaliana* sub-genome (Chr1 to 5), 35 genes, and **b** in the *A. arenosa* sub-genome (Chr 6 to 13), 585 genes show signatures of positive selection. These genes are found outside of the areas of low nucleotide diversity (π) within *A. suecica*, meaning a founder effect or population bottleneck is not the likely explanation for the signature of positive selection.


a
b
SCC2 Heatmap;AT5G15540
Suecica
ScarpT
WcarpD
UkraineD
BalticD
ScarpD
Ural
RailT
WcarpT
Group
A. arenosa 2x
A. arenosa 4x
A. suecica (A. arenosa)
SCC2 Tree
A. ar 4X
A. sue (A. th)
A. ar 2X
A. sue (A. ar)
A. th
Root
ATRANGAP1 Heatmap;AT3G63130
ATRANGAP1 Tree
FF1 Heatmap;AT5G47820
FF1 Tree
UBIQ Heatmap;AT5G02880
UBIQ Tree
CYCD5 Heatmap;AT4G37630
CYCD5 Tree

Supplementary Figure 3: Haplotype analysis and gene trees of five meiosis-related genes from selection scans on the *A. arenosa* subgenome of *A. suecica*. a Haplotype analysis based on SNP frequency sharing, points to a unique allele in the *A. arenosa* sub-genome *A. suecica* that is not the same allele as tetraploid *A. arenosa* suggesting **b**. Phylogenetic analysis shows the *A. arenosa* homeolog in *A. suecica* clusters with diploid *A. arenosa*, suggesting independent evolution of meiosis-related genes in the *A. arenosa* subgenome of *A. suecica* during polyploid adaptation. Notably, CYCD5 in *A. suecica* clusters with a Ruderal autotetraploid *A. arenosa* within an otherwise diploid clade, this is explained by the fact that Ruderal *A. arenosa* tetraploids have introgressed with the baltic diploid *A. arenosa[36]*. Branch lengths are not drawn to scale.



Supplementary Figure 4 Overlap of genes in selection scans and homeologous gene pairs carrying one LoF mutation. Minimal overlap was observed between genes from selection scans and homeologous gene pairs with one intact copy and one carrying a LoF mutation, indicating that LoF on one homeolog is not driving the likely adaptive genetic changes on the other.



Supplementary Figure 5. Lack of enrichment for LoF mutations in single-copy genes in *A. suecica.* **a** QQ plots comparing the population frequencies of LoF SNPs in single-copy genes versus multi-copy genes in *A. suecica*. No significant increase in the population frequency of LoF mutations is observed in *A. suecica* for single-copy genes, indicating that both single- and multi-copy genes are under similar selective pressures. **b-e** Joint allele frequency spectra for LoF SNPs, comparing *A. suecica* and the parental species for multi-copy and single-copy genes. The data show an enrichment of singleton SNPs for all species and gene copy status, consistent with strong purifying selection acting on these deleterious mutations.

Materials and Methods

DNA extraction of herbarium samples and sequencing

Herbarium samples were obtained from Moscow University Herbarium[73]. Extraction of DNA from herbarium samples of five *A. arenosa* and one *A. suecica* accession was carried out under a PCR workstation equipped with UV sterilization lights. The workstation, tubes and equipment were UV-sterilized before beginning the extraction. Herbarium samples were weighed and photographed. The material was then crushed using a sterile plastic pestle in a 2ml tube and then weighed again. PTB buffer was added and the material was kept at 37°C overnight on continuous rotation. We extracted the DNA following the protocol of[74]. Libraries were sequenced in 50-bp paired-end mode on Illumina HiSeq 2000 Analyzers at the Vienna BioCenter Core Facilities (VBCF).

Read mapping and SNP calling

For *A. thaliana* accessions, 970 accessions of *A. thaliana* were used from the 1,001 genomes dataset[53], these accessions had a read length greater than 60bp and greater than 8X coverage. In addition to the herbarium samples aforementioned, we used 15 previously published *A. suecica* accessions[49]. For *A. arenosa* we used 158 previously published individuals[36,75] in addition to the 5 herbarium *A. arenosa* individuals from (giving a total of 163 *A. arenosa* individuals). The *A. arenosa* individuals used had a minimum of 5X coverage. A complete list of the *A. thaliana* and *A. suecica* accessions and the *A. arenosa* individuals used in this study is available as Supplementary File 1.

We mapped illumina paired-end short reads for the 16 *A. suecica* accessions, 970 *A. thaliana* accessions and 163 *A. arenosa* individuals to a chromosome-level genome assembly of *A. suecica[24]* using the BWA-MEM algorithm with an increased penalty for unpaired reads set to 15[76] (version 0.7.17). The sub-genomes of the *A. suecica* accessions were treated separately after partitioning the mapped short reads. The *A. thaliana* accessions and *A. arenosa* individuals were mapped to the *A. thaliana* and *A. arenosa* sub-genome, respectively. PCR duplicates were removed using the Samtools[77] (version 1.10) rmdup function and SNPS were called using Genome Analysis Toolkit[78] (GATK version 3.8) using the HaplotypeCaller function. Generated gVCFs were combined into a single merged VCF using the combineGVCFs and GenotypeGVCFs functions.

Variant filtration

We removed multi-allelic sites and indels from our analysis and focused on bi-alleic SNPs. Bi-allelic SNPs were filtered to remove repetitive regions in the *A. suecica* sub-genomes (TEs sequences, centromere repeats, simple repeats and rDNA repeats). SNPs and non-variant sites were removed if the site was less than -0.5x log median coverage or greater than 0.5x log median coverage.

Heterozygous SNPs in *A. thaliana* have previously been shown to often result from segregating gene duplications rather than being residual heterozygosity from outcrossing[79]. To avoid SNP-calling errors from segregating gene duplications in our VCF, we removed heterozygous SNPs that were present in *A. thaliana* and *A. suecica* at a population frequency greater than 5%, since both are highly selfing species. As *A. arenosa* is an obligate outcrossing species, to avoid segregating duplicates we removed heterozygous SNPs that were

heterozygous in more than 90% of the individuals in the *A. arenosa* populations we examined. We allowed for 10% missing sites in the *A. arenosa* VCF. We allowed for 15% missing sites on the *A. thaliana* VCF.

The final VCFs contained 9,664,305 biallelic SNPs for the *A. arenosa* sub-genome, and 6,125,591 biallelic SNPs for the *A. thaliana* sub-genome.

Determining the ploidy level of *A. arenosa* individuals

The cytotype of the published *A. arenosa* individuals ploidy were previously confirmed using FACs[36,75]. nQuire[55] was run on raw fastq files of the 5 herbarium *A. arenosa* individuals, using the 'denoised' function and the 'lrdmodel'. Ploidy of the sample was determined by taking the ploidy that gave the maximum log-likelihood and lowest delta log-likelihood. Ploidy of the herbarium *A. arenosa* individuals was further confirmed by counting the number of different *S*-alleles for each individual with tetraploids having up to four and diploids up to two different *S*-alleles. The S-alleles were identified as described in[66] using short-read data and NGSGenotyp pipeline[80] with known *Arabidopsis* SRK sequences as a database.

Nucleotide divergence analysis

Pairwise nucleotide diversity (π) was calculated using custom scripts (see Code availability) in python. In order to carry out the network based on Nei's D, we read the VCF into R using the package vcfR[81] and calculated Nei'D using StAMPP[58]. To visualize the network we used the R package tanngle[82] and ggtree[83]. Geographic maps were made using the ggspatial package in R[84].

Niche modeling of *A. suecica*

We modeled the niche of *A. suecica* using the current occurrence records and the selected Bioclim variables to project it onto the climate at LGM. We used workflow described in the ModleR package in R[85] to the entire modeling (set up, evaluation and projection).

We obtained occurrence data of *A. suecica* from the GBIF database (gbif.org). There are a total of 8451 records available of which we considered records only from preserved specimens having georeferencing data (1772 records). This data was supplemented with the published records[49]. After cleaning the data for duplicates and occurrences in the same pixel 483 clean data points were obtained. Then, species thinning was carried out using the geo_filt function embedded in ModleR. Finally, 197 data points remained after applying the geographic filter of 0.5 km. These occurrence points were considered for further steps. We created 10000 background points using the setup_sdmdata function in the ModleR package.

We obtained 19 bioclim variables from Worldclim climate data (Fick and Hijmans, 2017) at a resolution of 30 arc seconds. We extracted the data for our occurrence points using the R package raster[86]. Multicollinearity was checked for the bioclim variables in order to enhance the model consistency and remove the impact of highly correlated variables. Out of the 19 variables, eight uncorrelated variables (bio3, bio4, bio8, bio9, bio10, bio14, bio15) were then used for further analysis

We used maxent to fit our model, which uses the occurrence data and the corresponding climatic data to predict the probability of distribution of the target species. Twenty-five models were created by partitioning the data into five cross-validation partition types and five runs per partition. We evaluated the performance of these models using the true skill statistic (TSS), which is based on the confusion matrix[87]. The TSS scores span from 0 to 1, where 0 signifies

a model performing no better than chance, and 1 denotes a flawless alignment with the observed data. We had set a TSS value greater than 0.7 to prepare the final model. Since all our models showed TSS values >0.9, we considered all 25 models.
The final model was projected onto the past climate (LGM). Further, occurrence points of *A. thaliana* and *A. arenosa* were overlaid on the projection to understand the origin and distribution of *A. suecica* environmentally and spatially.

Selection scans

To examine polymorphism patterns in genes we first ran SnpEff for annotation, and we used a modified version of the McDonald Kreitman (MK) test in order to find genes showing signatures of selection in *A. suecica.* We used a modified version of the MK test as *A. suecica* is evolutionarily young (~16 Kya), meaning if we relied on ratios of divergence to polymorphism in the MK test we would often have division by zero. Therefore we focused instead on genes where nonsynonymous divergence (Dn) exceeded synonymous divergence (Ds), nonsynonymous polymorphisms (Pn) were fewer than synonymous polymorphisms (Ps) and where Pn was zero in *A. suecica.*

Gene ontology (GO) enrichment

We used the R package TopGO[88] to perform our GO enrichment analysis, using the ‘weight01’ algorithm to account for the hierarchical structure in GO terms. We kept GO terms with a Fisher test p-value lower than 0.05. We used the gene annotations of *A. thaliana* one-to-one orthologs of *A. suecica* genes. Gene annotations for *A. thaliana* were obtained using the R package biomaRt from Ensembl ‘biomaRt::useMart(biomart = ’plants_mart’, dataset = ’athaliana_eg_gene’, host = 'plants.ensembl.org').

SNP annotation, polarization and homeolog analysis

We annotated SNPs using SNPeff (version 5.2). We used 200 of the genetically closest *A. thaliana* accession to polarize annotated SNPs on the *A. arenosa* sub-genome. We used 29 individuals of the closest diploid *A. arenosa* to polarize annotated SNPs on the *A. thaliana* sub-genome. This resulted in 3,500,145 annotated and polarized SNPs for the *A. arenosa* sub-genome and 1,205,789 for the *A. thaliana* sub-genome. 11,812 and 11,983 SNPs, on the *A. arenosa* and *A. thaliana* sub-genome respectively, were annotated as predicted LoF polymorphisms that were later used in our homeolog analysis. We excluded SNPs with heterozygosity frequencies exceeding 1% in both *A. thaliana* and *A. suecica* to minimize the risk of bias in the LoF mutation analysis, which could be introduced by segregating gene duplicates leading to false signals of heterozygosity. We filtered for homeolog pairs where in *A. suecica* at least one homeolog in the pair carried a LoF mutation, amounting to 414 genes. In instances where a homeolog carried more than one LoF mutation, we selected the LoF mutation that was at a higher population frequency.

Permutation analysis

To assess the negative correlation of LoF mutations between the *A. thaliana* and *A. arenosa* sub-genomes in *A. suecica*, we conducted two types of permutation analyses. First, we compared the observed number of LoF mutations in homeolog gene pairs within *A. suecica* to randomly paired homeologs. From the 414 homeolog gene pairs, where at least one gene in

the pair carried a LoF polymorphism, we generated 1,000 permutations of random homeolog pairings. The observed number of homeolog pairs with LoF mutations in both sub-genomes was then compared to the distribution of random pairings to assess whether fewer dual LoF mutations occurred than expected by chance.

Second, we performed a permutation analysis using in-silico *A. suecica*. We randomly paired *A. thaliana* and *A. arenosa* genomes to simulate 1,000 in-silico *A. suecica* genomes. We used 200 of the genetically closest *A. thaliana* and 29 of the closest diploid *A. arenosa* individuals to generate the in-silico *A. suecica* combinations. The observed dual LoF count in *A. suecica* was compared to the distribution generated by the in-silico genomes.

Analysis of meiotic genes in the *A. suecica A. arenosa* sub-genome

Biallelic SNPs from CDS regions of the genes were used to generate SNP-frequency heatmaps. *A.arenosa* assemblies from[89] were scaffolded with RagTag[90] on the *A.lyrata* NT1 genome assembly[66] and annotated with Helixer[91]. *A. thaliana* protein sequences of target genes were used for protein BLAST[92] against annotated proteins extracted with AGAT[93]. Genes identified by BLAST were aligned with L-INS-i algorithm of MAFFT[94] and trees were constructed with IQTree[95,96] with fast bootstrap option equals 1000. Protein alignments were visualized using snipit[97]

Single and multi copy gene analysis

Single copy and multi copy genes were obtained from[48] and hypergeometric tests were performed to test for enrichment.

Code availability

Custom scripts used in this study are available at https://github.com/burnsro/ArabidopsisSuecica/tree/main/origin

Data availability

Raw fastq files of the Illumina sequenced herbarium samples for 5 tetraploid *A. arenosa* and 1 *A. suecica* generated in this study were deposited to the European Nucleotide Archive (ENA) under the project number PRJEB79527.

Author contributions

R.B and P.Y.N conceived the study. Herbarium samples were aquired by P.Y.N and genomic DNA was extracted by R.B. R.B, P.Y.N, A.K, A.G, U.K.K, A.D.S analyzed the data. R.B, P.Y.N, F.K and A.D.S wrote the paper.

Acknowledgements

This work was supported by an EMBO long-term postdoctoral fellowship ALTF224-2022 to R.B. P.Y.N. acknowledges the European Union (ERC, HOW2DOUBLE, 101041354). The views and opinions expressed are however those of the author (s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. F. K. and P.Y.N. acknowledge the DAAD exchange travel grant, project number 57601834. F. K. and P.Y.N. also acknowledge the joint Deutsche Forschungsgemeinschaft (DFG) and Czech Science Foundation (GACR) program, project number 490698526 and 22-29078K. We thank Heïdi Serra and Andrew Lloyd for helpful comments on the manuscript and discussions.

Competing interests

The authors declare no competing interests.

References

1. Aköz G, Nordborg M. The Aquilegia genome reveals a hybrid origin of core eudicots. Genome Biol. 2019;20: 256.

2. Hokamp K, McLysaght A, Wolfe KH. The 2R hypothesis and the human genome sequence. Genome Evolution. 2003; 95–110.

3. Van de Peer Y. Polyploidy: Methods and Protocols. Springer Nature; 2023.

4. Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM, et al. Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. Nat Genet. 2020;52: 525–533.

5. Li A-L, Geng S-F, Zhang L-Q, Liu D-C, Mao L. Making the Bread: Insights from Newly Synthesized Allohexaploid Wheat. Mol Plant. 2015;8: 847–859.

6. Dubcovsky J, Dvorak J. Genome plasticity a key factor in the success of polyploid wheat under domestication. Science. 2007;316: 1862–1866.

7. Liston A, Cronn R, Ashman T-L. Fragaria: a genus with deep historical roots and ripe for evolutionary and ecological insights. Am J Bot. 2014;101: 1686–1699.

8. Staudt G. STRAWBERRY BIOGEOGRAPHY, GENETICS AND SYSTEMATICS. International Society for Horticultural Science (ISHS), Leuven, Belgium; 2009.

9. U, Nagaharu. Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization. Japanese Journal of Botany. 1935;7: 389–452.

10. Feldman M, Levy AA. Allopolyploidy--a shaping force in the evolution of wheat genomes. Cytogenet Genome Res. 2005;109: 250–258.

11. Renny-Byfield S, Wendel JF. Doubling down on genomes: polyploidy and crop plants. Am J Bot. 2014;101: 1711–1725.

12. Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, et al. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat Genet. 2019;51: 877–884.

13. Sierro N, Auberson M, Dulize R, Ivanov NV. Chromosome-level genome assemblies of Nicotiana tabacum, Nicotiana sylvestris, and Nicotiana tomentosiformis. Sci Data. 2024;11: 135.

14. Rarely successful polyploids and their legacy in plant genomes. Curr Opin Plant Biol. 2012;15: 140–146.

15. Newton nW CF, Pellew N. Primula kewensis and its derivatives. J Genet. 1929;20: 405–467.

16. Bachmann JA, Tedder A, Fracassetti M, Steige KA, Lafon-Placette C, Köhler C, et al. On the origin of the widespread self-compatible allotetraploid Capsella bursa-pastoris (Brassicaceae). Heredity . 2021. doi:10.1038/s41437-021-00434-9

17. Wiiliam M, Malcolm A. Biosystematics, genetics, and physiological ecology of the Erythranthe section of Mimulus. 1971.

18. Nasrallah ME, Yogeeswaran K, Snyder S, Nasrallah JB. Arabidopsis species hybrids in the study of species differences and evolution of amphiploidy in plants. Plant physiology. 2000;124. doi:10.1104/pp.124.4.1605

19. Katche EI, Schierholt A, Schiessl S-V, He F, Lv Z, Batley J, et al. Genetic factors inherited from both diploid parents interact to affect genome stability and fertility in resynthesized allotetraploid Brassica napus. G3 . 2023;13. doi:10.1093/g3journal/jkad136

20. Katche EI, Schierholt A, Becker HC, Batley J, Mason AS. Fertility, genome stability, and homozygosity in a diverse set of resynthesized rapeseed lines. The Crop Journal. 2023. Available: https://www.sciencedirect.com/science/article/pii/S2214514122002082

21. Stebbins GL. Types of polyploids: their classification and significance. Adv Genet. 1947;1: 403–429.

22. Soltis DE, Soltis PS, Rieseberg LH. Molecular Data and the Dynamic Nature of Polyploidy. CRC Crit Rev Plant Sci. 1993;12: 243–273.

23. Ramsey J, Schemske DW. Neopolyploidy in Flowering Plants. Annu Rev Ecol Syst. 2002;33: 589–639.

24. Burns R, Mandáková T, Gunis J, Soto-Jiménez LM, Liu C, Lysak MA, et al. Gradual evolution of allopolyploidy in Arabidopsis suecica. Nat Ecol Evol. 2021. doi:10.1038/s41559-021-01525-w

25. Nibau C, Evans A, King H, Phillips DW, Lloyd A. Homoeologous crossovers are distally biased and underlie genomic instability in first-generation neo-allopolyploid Arabidopsis suecica. New Phytol. 2024 [cited 10 Sep 2024]. doi:10.1111/nph.20095

26. Chéron F, Petiot V, Lambing C, White C, Serra H. Incorrect recombination partner associations contribute to meiotic instability of neo-allopolyploid Arabidopsis suecica. New Phytol. 2023. doi:10.1111/nph.19487

27. Yant L, Hollister JD, Wright KM, Arnold BJ, Higgins JD, Franklin FCH, et al. Meiotic adaptation to genome duplication in Arabidopsis arenosa. Curr Biol. 2013;23: 2151–

2156.

28. Hollister JD, Arnold BJ, Svedin E, Xue KS, Dilkes BP, Bomblies K. Genetic adaptation associated with genome-doubling in autotetraploid Arabidopsis arenosa. PLoS Genet. 2012;8: e1003093.

29. Bohutínská M, Vlček J, Yair S, Laenen B, Konečná V, Fracassetti M, et al. Genomic basis of parallel adaptation varies with divergence in and its relatives. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2022713118

30. Martín AC, Alabdullah AK, Moore G. A separation-of-function ZIP4 wheat mutant allows crossover between related chromosomes and is meiotically stable. Sci Rep. 2021;11: 21811.

31. Serra H, Svačina R, Baumann U, Whitford R, Sutton T, Bartoš J, et al. Ph2 encodes the mismatch repair protein MSH7-3D that inhibits wheat homoeologous recombination. Nat Commun. 2021;12: 803.

32. Liu Z, Adamczyk K, Manzanares-Dauleux M, Eber F, Lucas M-O, Delourme R, et al. Mapping PrBn and other quantitative trait loci responsible for the control of homeologous chromosome pairing in oilseed rape (Brassica napus L.) haploids. Genetics. 2006;174: 1583–1596.

33. Henry IM, Dilkes BP, Tyagi A, Gao J, Christensen B, Comai L. The BOY NAMED SUE quantitative trait locus confers increased meiotic stability to an adapted natural allopolyploid of Arabidopsis. Plant Cell. 2014;26: 181–194.

34. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet. 2017;18: 411–424.

35. Conover JL, Wendel JF. Deleterious Mutations Accumulate Faster in Allopolyploid Than Diploid Cotton (Gossypium) and Unequally between Subgenomes. Mol Biol Evol. 2022;39: msac024.

36. Monnahan P, Kolář F, Baduel P, Sailer C, Koch J, Horvath R, et al. Pervasive population genomic consequences of genome duplication in Arabidopsis arenosa. Nat Ecol Evol. 2019;3: 457–468.

37. Spoelhof JP, Soltis DE, Soltis PS. Polyploidy and mutation in Arabidopsis. Evolution. 2021;75: 2299–2308.

38. Baduel P, Quadrana L, Hunter B, Bomblies K, Colot V. Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. Nat Commun. 2019;10: 5818.

39. Mandáková T, Lysak MA. Post-polyploid diploidization and diversification through dysploid changes. Curr Opin Plant Biol. 2018;42: 55–65.

40. Redmond AK, Casey D, Gundappa MK, Macqueen DJ, McLysaght A. Independent rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor. Nat Commun. 2023;14: 1–15.

41. Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. 2011;108: 4069.

42. Sun Y, Liu Y, Shi J, Wang L, Liang C, Yang J, et al. Biased mutations and gene losses

underlying diploidization of the tetraploid broomcorn millet genome. Plant J. 2023;113: 787–801.

43. Cheng F, Wu J, Wang X. Genome triplication drove the diversification of Brassica plants. Hortic Res. 2014;1: 14024.

44. Scarlett VT, Lovell JT, Shao M, Phillips J, Shu S, Lusinska J, et al. Multiple origins, one evolutionary trajectory: gradual evolution characterizes distinct lineages of allotetraploid Brachypodium. Genetics. 2023;223. doi:10.1093/genetics/iyac146

45. Ma P-F, Liu Y-L, Guo C, Jin G, Guo Z-H, Mao L, et al. Genome assemblies of 11 bamboo species highlight diversification induced by dynamic subgenome dominance. Nat Genet. 2024;56: 710–720.

46. Alger EI, Edger PP. One subgenome to rule them all: underlying mechanisms of subgenome dominance. Curr Opin Plant Biol. 2020;54: 108–113.

47. Penin AA, Kasianov AS, Klepikova AV, Omelchenko DO, Makarenko MS, Logacheva MD. Origin and diversity of Capsella bursa-pastoris from the genomic point of view. BMC Biol. 2024;22: 1–15.

48. Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms. Plant Cell. 2016;28: 326–344.

49. Novikova PY, Tsuchimatsu T, Simon S, Nizhynska V, Voronin V, Burns R, et al. Genome Sequencing Reveals the Origin of the Allotetraploid Arabidopsis suecica. Mol Biol Evol. 2017;34: 957.

50. Nibau C, Gonzalo A, Evans A, Sweet-Jones W, Phillips D, Lloyd A. Meiosis in allopolyploid Arabidopsis suecica. Plant J. 2022;111: 1110–1122.

51. Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, et al. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat Genet. 2016;48: 1077–1082.

52. Bohutínská M, Petříková E, Booker TR, Cobo CV, Vlček J, Šrámková G, et al. Repeated adaptation to whole genome duplication in outcrossing Arabidopsis is mediated by mosaic adaptive haplotypes. bioRxiv. 2023. p. 2023.01.11.523565. doi:10.1101/2023.01.11.523565

53. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016;166: 481–491.

54. Bohutínská M, Petříková E, Booker TR, Cobo CV, Vlček J, Šrámková G, et al. Novel reassortments of variation underlie repeated adaptation to whole genome duplication in two outcrossing Arabidopsis species. bioRxiv. 2024. p. 2023.01.11.523565. doi:10.1101/2023.01.11.523565

55. Weiß CL, Pais M, Cano LM, Kamoun S, Burbano HA. nQuire: a statistical framework for ploidy estimation using next generation sequencing. BMC Bioinformatics. 2018;19: 122.

56. Scott AD, Kolesnikova U, Glushkevich A, Steinmann L, Tikhomirov N, Pfordt U, et al. Multiple polyploidizations in Arabidopsis lyrata stabilized by long-range adaptive introgression across Eurasia. bioRxiv. 2024. p. 2024.08.27.609292.

doi:10.1101/2024.08.27.609292

57. Nei M. Genetic Distance between Populations. Am Nat. 1972;106: 283–292.

58. Pembleton LW, Cogan NOI, Forster JW. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Mol Ecol Resour. 2013;13: 946–952.

59. Kolář F, Fuxová G, Záveská E, Nagano AJ, Hyklová L, Lučanová M, et al. Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model Arabidopsis arenosa. Mol Ecol. 2016;25: 3929–3949.

60. Hylander N. Cardaminopsis suecica (fr.) Hiit., a Northern Amphidiploid species. Bulletin du Jardin botanique de l'État a Bruxelles. 1957;27: 591–604.

61. Lee C-R, Svardal H, Farlow A, Exposito-Alonso M, Ding W, Novikova P, et al. On the post-glacial spread of human commensal Arabidopsis thaliana. Nat Commun. 2017;8: 1–12.

62. Wang H, Xu W, Sun Y, Lian Q, Wang C, Yu C, et al. The cohesin loader SCC2 contains a PHD finger that is required for meiosis in land plants. PLOS Genetics. 2020;16: e1008849.

63. Wright KM, Arnold B, Xue K, Šurinová M, O'Connell J, Bomblies K. Selection on Meiosis Genes in Diploid and Tetraploid Arabidopsis arenosa. Mol Biol Evol. 2014;32: 944–955.

64. Bohutínská M, Handrick V, Yant L, Schmickl R, Kolář F, Bomblies K, et al. De Novo Mutation and Rapid Protein (Co-)evolution during Meiotic Adaptation in Arabidopsis arenosa. Mol Biol Evol. 2021;38: 1980–1994.

65. Paape T, Briskine RV, Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi R, Hatakeyama M, et al. Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. Nat Commun. 2018;9: 3909.

66. Kolesnikova UK, Scott AD, Van de Velde JD, Burns R, Tikhomirov NP, Pfordt U, et al. Transition to Self-compatibility Associated With Dominant S-allele in a Diploid Siberian Progenitor of Allotetraploid Arabidopsis kamchatica Revealed by Arabidopsis lyrata Genomes. Mol Biol Evol. 2023;40. doi:10.1093/molbev/msad122

67. Joseffson C, Dilkes B, Comai L. Parent-Dependent Loss of Gene Silencing during Interspecies Hybridization. Current Biology. 2006;16: 1322–1328.

68. Wos G, Macková L, Kubíková K, Kolář F. Ploidy and local environment drive intraspecific variation in endoreduplication in Arabidopsis arenosa. Am J Bot. 2022;109: 259–271.

69. Pacey EK, Maherali H, Husband BC. The influence of experimentally induced polyploidy on the relationships between endopolyploidy and plant function in. Ecol Evol. 2020;10: 198–216.

70. Bao Z, Li C, Li G, Wang P, Peng Z, Cheng L, et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. Mol Plant. 2022;15: 1211–1226.

71. Bomblies K, Jones G, Franklin C, Zickler D, Kleckner N. The challenge of evolving stable polyploidy: could an increase in "crossover interference distance" play a central role? Chromosoma. 2016;125: 287–300.

72. Mihajlovic L, Iyengar BR, Baier F, Barbier I, Iwaszkiewicz J, Zoete V, et al. A direct experimental test of Ohno's hypothesis. Elife. 2024;13. doi:10.7554/eLife.97216.2

73. Website. Available: Seregin A. 2020. Moscow University Herbarium (MW). Available from: http://dx.doi.org/10.15468/CPNHC

74. Gutaker RM, Reiter E, Furtwängler A, Schuenemann VJ, Burbano HA. Extraction of ultrashort DNA molecules from herbarium specimens. Biotechniques. 2017;62: 76–79.

75. Bohutínská M, Petříková E, Booker TR, Cobo CV. Repeated adaptation to whole genome duplication in outcrossing Arabidopsis is mediated by mosaic adaptive haplotypes. bioRxiv. 2023. Available: https://www.biorxiv.org/content/10.1101/2023.01.11.523565.abstract

76. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760.

77. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

78. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.

79. Jaegle B, Pisupati R, Soto-Jiménez LM, Burns R, Rabanal FA, Nordborg M. Extensive sequence duplication in Arabidopsis revealed by pseudo-heterozygosity. Genome Biol. 2023;24: 1–19.

80. Genete M, Castric V, Vekemans X. Genotyping and De Novo Discovery of Allelic Variants at the Brassicaceae Self-Incompatibility Locus from Short-Read Sequencing Data. Mol Biol Evol. 2020;37: 1193–1201.

81. Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour. 2017;17: 44–53.

82. Schliep K. ***tanggle***: Visualization of phylogenetic networks in a *ggplot2* framework. [cited 18 Apr 2024]. Available: https://klausvigo.github.io/tanggle/articles/tanggle_vignette.html

83. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. Curr Protoc Bioinformatics. 2020;69: e96.

84. Website. Available: Dunnington D (2023). ggspatial: Spatial Data Framework for ggplot2. https://paleolimbot.github.io/ggspatial/, https://github.com/paleolimbot/ggspatial.

85. Sánchez-Tapia A, Mortara SR, Bezerra Rocha DS, Mendes Barros FS, Gall G, de Siqueira MF. modleR: a modular workflow to perform ecological niche modeling in R. bioRxiv. bioRxiv; 2020. doi:10.1101/2020.04.01.021105

86. Hijmans RJ. raster: Geographic data analysis and modeling. R package version. 2018;2: 8.

87. Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol. 2006;43: 1223–1232.

88. Website. Available: Alexa A, Rahnenfuhrer J (2023). topGO: Enrichment Analysis for Gene Ontology. doi:10.18129/B9.bioc.topGO, R package version 2.54.0,

https://bioconductor.org/packages/topGO.

89. Bohutínská M, Petříková E, Booker TR, Cobo CV, Vlček J, Šrámková G, et al. Novel reassortments of variation underlie repeated adaptation to whole genome duplication in two outcrossing Arabidopsis species. 2023. doi:10.1101/2023.01.11.523565

90. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. 2022;23: 1–19.

91. Stiehler F, Steinborn M, Scholz S, Dey D, Weber APM, Denton AK. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. Bioinformatics. 2021;36: 5291–5298.

92. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.

93. Dainat J, Hereñú D, Murray KD, Davis E, Ugrin I, Crouch K, et al. NBISweden/AGAT: AGAT-v1.4.1. Zenodo; 2024. doi:10.5281/ZENODO.3552717

94. Katoh K, Kuma K-I, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005;33: 511–518.

95. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018;35: 518–522.

96. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274.

97. O'Toole Á, Aziz A, Maloney D. Publication-ready single nucleotide polymorphism visualization with snipit. Bioinformatics. 2024;40. doi:10.1093/bioinformatics/btae510