OPEN
Gradual evolution of allopolyploidy in Arabidopsis suecica

Robin Burns, Terezie Mandáková, Joanna Gunis, Luz Mayela Soto-Jiménez, Chang Liu, Martin A. Lysak, Polina Yu. Novikova and Magnus Nordborg

Most diploid organisms have polyploid ancestors. The evolutionary process of polyploidization is poorly understood but has frequently been conjectured to involve some form of ‘genome shock’, such as genome reorganization and subgenome expression dominance. Here we study polyploidization in Arabidopsis suecica, a post-glacial allopolyploid species formed via hybridization of Arabidopsis thaliana and Arabidopsis arenosa. We generated a chromosome-level genome assembly of A. suecica and complemented it with polymorphism and transcriptome data from all species. Despite a divergence around 6 million years ago (Ma) between the ancestral species and differences in their genome composition, we see no evidence of a genome shock: the A. suecica genome is colinear with the ancestral genomes; there is no subgenome dominance in expression; and transposon dynamics appear stable. However, we find changes suggesting gradual adaptation to polyploidy. In particular, the A. thaliana subgenome shows upregulation of meiosis-related genes, possibly to prevent aneuploidy and undesirable homeologous exchanges that are observed in synthetic A. suecica, and the A. arenosa subgenome shows upregulation of cyto-nuclear processes, possibly in response to the new cytoplasmic environment of A. suecica, with plastids maternally inherited from A. thaliana. These changes are not seen in synthetic hybrids, and thus are likely to represent subsequent evolution.
Fig. 1 | The genome of *A. suecica* is largely colinear with the ancestral genomes. a. Schematic depicting the origin of *A. suecica* and its current distribution in relation to the ice cover at the Last Glacial Maximum. ka, thousand years ago. Ice cover data are from Natural Resource Canada (https://open.canada.ca/data/en/dataset/a384bada-a787-5b49-9799-f5d589e97bd3). b. Chromosome-level assembly of the *A. suecica* genome with inner links depicting syntenic blocks between the *A. thaliana* and *A. arenosa* subgenomes of *A. suecica*. Histograms show the distribution of TEs (in blue) and protein-coding genes (in green) along the chromosomes. c. Synteny of the *A. thaliana* subgenome of *A. suecica* to the *A. thaliana* TAIR10 reference. In total 13 colinear synteny blocks were found. d. Synteny of the *A. arenosa* subgenome to *A. lyrata*. In total 40 synteny blocks were found, 33 of which were colinear. Of the remaining seven blocks, five represent inversions in the *A. arenosa* subgenome of *A. suecica* relative to *A. lyrata*, one is a translocation and one corresponds to a previously reported misassembly in the *A. lyrata* genome. Orange bars show the density of missing regions (‘N’ bases) in the *A. lyrata* genome.
the subgenomes facilitated the assembly. By contrast, assembling A. arenosa is complicated by high heterozygosity (around 3.5% nucleotide diversity) and high repeat content. Our assembly of a tetraploid A. arenosa individual, included in this study, is fragmented: 3,629 contigs; N50 of 331 kb. The A. suecica assembly, however, has an N50 contig size of 9.02 Mb. Contigs totalled 276 Mb (around 90% and around 88% of genome size estimated by flow cytometry and k-mer analysis, respectively; see Extended Data Fig. 1 and Methods). Contigs were placed into scaffolds using chromatin conformation capture (Hi-C) data and using the reference genomes of A. thaliana and Arabidopsis lyrata (as substitute for A. arenosa) as guides. This resulted in 13 chromosome-scale scaffolds (Extended Data Fig. 2). The placement and orientation of each contig within a scaffold was confirmed and corrected using a genetic map for A. suecica (see Methods and Extended Data Fig. 2). The final assembly (Fig. 1b) contains 262 Mb and has an N50 scaffold size of 19.59 Mb. The 5 + 8 chromosomes of the A. thaliana and A. arenosa subgenomes sum to 119 Mb and 143 Mb, respectively.

Of the A. thaliana and A. arenosa subgenomes of A. suecica, 108 Mb and 135 Mb are in large blocks syntenic to the genomes of the ancestral species: 13 and 40 blocks, respectively (Fig. 1c,d). The majority of these syntenic blocks are colinear, with the exception of five small-scale inversions (around 4.5 Mb) and one translocation (around 244 kb) on the A. arenosa subgenome—which may reflect differences between A. lyrata and A. arenosa, two highly polymorphic species separated by about a million years. We also corrected the described misassembly in the A. lyrata reference genome using our genetic map. Overall we find that approximately 93% of the A. suecica genome is syntenic to the ancestral genomes (Fig. 1c,d). This highlights the conservation of the A. suecica genome and contrasts with the major rearrangements that have been observed in several resynthesized polyploids, and some crops. Notably, major rearrangements have also been observed in synthetic A. suecica.

A total of 45,585 protein-coding genes were annotated for the A. suecica reference, of which 22,232 and 23,353 are located on the A. thaliana and A. arenosa subgenomes, respectively. We assessed completeness of the genome assembly and annotation with the BUSCO set for eudicots and found 2,088 (98.4%) complete genes for both the A. thaliana and A. arenosa subgenomes. Of the protein-coding genes, 18,023 had a one-to-one orthology between the subgenomes of A. suecica and 16,999 genes were conserved single-copy orthologues for each sub-genome of A. suecica and the ancestral species (Supplementary Data 2 and Extended Data Fig. 3). We annotated lineage-specific genes in A. suecica (that is, genes in A. suecica with no orthologue) using InterPro. We found significant enrichment in the A. thaliana subgenome for two gene ontology (GO) terms (GO:0008234 and GO:0015074) that are associated with repeat content (Supplementary Data 2). Ancestral genes missing in our annotation were overrepresented for functional categories of defence response. Examining DNA-sequencing coverage for these genes in the ancestral genomes did not confirm any gene loss, suggesting rather misassembly or misannotation, likely because of the repetitive and highly polymorphic nature of resistance genes (R-genes).

The ribosomal DNA clusters are highly variable. In eukaryotic genomes, genes encoding ribosomal RNA (rRNA) occur as tandem arrays in rDNA clusters. The 45S rDNA clusters are massive, containing hundreds or thousands of copies and spanning millions of base pairs. The site of pre-ribosome assembly (nucleolus), forms at these clusters if they are actively transcribed. In inter-specific hybrids it was previously observed that the rDNA of only one parent tended to be involved in nucleolus formation, a phenomenon known as ‘nucleolar dominance’. In A. suecica, it was observed that the rDNA clusters inherited from A. thaliana were silenced, and structural changes associated with these clusters were also suggested.

Given this, we examined the composition and transcription of 45S rDNA repeats. Although the large and highly repetitive 45S rDNA clusters are missing from the genome assembly, we can measure the copy number of A. thaliana and A. arenosa 45S rDNA genes using sequencing coverage (see Methods). We find that three accessions have experienced massive loss of the A. thaliana rDNA loci (Fig. 2a), which was confirmed for one of the accessions (AS90a) by fluorescence in situ hybridization (FISH) analysis (Fig. 2b,c). However, there is massive copy number variation for 45S rRNA genes in A. suecica (Fig. 2a), and some accessions (AS33) have a higher 45S rRNA copy number in A. thaliana than in A. arenosa (Fig. 2d,e).

We find nucleolar dominance to be variable in A. suecica (see Methods and Extended Data Fig. 3). The majority of accessions express both 45S rRNA alleles, five exclusively express A. arenosa 45S rRNA and one exclusively expresses A. thaliana 45S rRNA (Fig. 2a).

This extensive variation in 45S cluster size and expression is consistent with the intraspecific variation seen in A. thaliana and A. arenosa, and previous observations made in natural A. suecica. This suggests that nucleolar dominance may partly be explained by retained ancestral variation. However, the large decrease in rDNA cluster size observed in some accessions may be a consequence of allopolyploidization, as synthetic A. suecica sometimes shows loss of 45S rDNA (even as early as the F1 stage) that varies between siblings and generations (Extended Data Fig. 3). Elimination of rDNA loci has also been previously observed in synthetic wheat, and loss of rDNA sites has been reported in strawberry.

No evidence for abnormal transposon activity. The possibility that allopolyploidization leads to a ‘genome shock’ in the form of increased transposon activity has been discussed. Evidence for TE proliferation following hybridization has been found for Tyl/Gypsy retrotransposons in hybrid sunflower species, although this may be due to environmental change. Analysis of TE expression in F1 hybrids between A. thaliana and A. lyrata found strong correlation to the parent species, and little alteration of repressive chromatin marks—although the F1 generation may be too early to study TE misregulation. Here we examine TE dynamics in A. suecica.

In A. suecica there are almost twice as many annotated transposons in the A. arenosa compared to the A. thaliana subgenome (66,722 versus 33,420). The difference is likely to be greater given that the A. arenosa subgenome is less complete and TE annotation is biased towards A. thaliana. Whether the combination of these two genomes has led to increased transposon activity is unknown.

The A. thaliana subgenome contains around 3,000 more annotated transposons than the TAIR10 A. thaliana reference genome but could reflect a greater number of transposons in the A. thaliana ancestors rather than increased transposon activity in A. suecica. To investigate TE activity, unique TE jumps that occurred after the species separated are needed. We used the software PoPoolationTE to call presence–absence variation on a population level using 15 natural A. suecica accessions, 18 A. thaliana accessions genetically close to A. suecica, and 9 A. arenosa lines. Of the 24,569 insertion polymorphisms in the A. thaliana subgenome, 8,767 were shared between A. thaliana and A. suecica, 7,196 were unique to A. thaliana and 8,606 were unique to A. suecica. Of the 115,336 insertion polymorphisms in the A. arenosa subgenome, 13,177 were shared with A. arenosa, 83,964 were unique to A. arenosa and 18,195 were unique to A. suecica (Supplementary Data 1a,b and Extended Data Fig. 4). Considering the number of transposons per individual genome (Fig. 3a), most transposon insertions in a typical A. thaliana subgenome are also found in A. thaliana. The slightly higher
The relationship between expression levels (log$_2$(CPM)) and copy number of 45S rDNA shows extensive variation of 45S rDNA copy number and varying direction of ‘nucleolar dominance’. Grey lines connect subgenomes of the same accession. Values above the dashed line are taken as evidence for expression of a particular 45S rDNA allele, as this is above the maximum level of mis-mapping seen in the ancestral species (see Extended Data Fig. 3). b, c, FISH results of a natural A. suecica accession AS90a that has largely lost the rDNA cluster of the A. thaliana subgenome (8 copies calculated for the A. thaliana 45S rDNA and 159 copies of the A. arenosa 45S rDNA). d, e, FISH results of a natural accession ASS3 that has maintained both ancestral rDNA loci (174 copies calculated for the A. thaliana 45S rDNA and 104 copies of the A. arenosa 45S rDNA). Scale bars, 10 μm (b, d).

In summary, we see no evidence for a burst of transposon activity in A. suecica, a conclusion supported by the analysis of transposon expression for A. suecica, which shows no upregulation relative to the ancestor species (Extended Data Fig. 5). The frequency distribution of polymorphic transposon insertions unique to A. suecica is heavily skewed towards zero, likely because of purifying selection as the distribution is more similar to that of non-synonymous than synonymous single-nucleotide polymorphisms (SNPs) (Fig. 3b,c). However, for both subgenomes, A. suecica also contains a large number of fixed or nearly fixed insertions that are present in the ancestral species at a lower frequency (Fig. 3d,e). These are likely to have reached high frequency as a result of a bottleneck. Shared transposons are enriched in the pericentromeric regions, while unique transposon insertions, generally at low frequency, are more uniformly distributed across the genome, which is consistent with strong selection against transposon insertions in the gene-dense chromosome arms.

An interesting subset of recent transposon insertions unique to A. suecica are those that have jumped between the subgenomes. We searched for full-length transposon copies that are present in both subgenomes of A. suecica and assigned the transposon sequences to the A. thaliana or the A. arenosa ancestral genome (see Methods). We were able to assign 15 and 56 transposon sequences as belonging to the A. thaliana and A. arenosa ancestral genome, respectively. Using these sequences, we searched our transposon polymorphisms and identified 1515 A. arenosa transposon polymorphisms on the A. thaliana subgenome, and 496 A. thaliana transposon polymorphisms on the A. arenosa subgenome. Like other private polymorphisms, these are skewed towards rare frequencies, and are uniformly distributed across the (sub-)genome. Most of the transposons that have jumped into the A. thaliana subgenome are helitron and long terminal repeat (LTR) elements (Extended Data Fig. 4). LTR elements also make up most of the A. thaliana transposons segregating in the A. arenosa subgenome. Three times as many transposon jumps from A. arenosa to A. thaliana than vice versa is notable, and suggests higher transposon activity in the A. arenosa subgenome, but we must consider differences in genome size and transposon number. If no differences in activity exist, we would expect the number of jumps to be proportional to the number of potential source elements and the size of the target genome. As the A. arenosa subgenome contains roughly twice as many transposons as the A. thaliana subgenome and is about 20% larger, we expect a 1.7-fold difference, not a three-fold one.

In conclusion, transposon activity in A. suecica appears to be governed by the same processes as in the ancestral species.
No global dominance in expression between the subgenomes.

Over time the traces of polyploidy are erased through an evolutionary process referred to as fractionation or re-diploidization\(^{104-108}\). Analyses of retained homeologues in ancient allopolyploids such as *A. thaliana*\(^{109}\), maize\(^{110}\), *B. rapa*\(^{111}\) and *Gossypium raimondii*\(^{112}\) have revealed that one ‘dominant’ subgenome remains more intact, with more highly expressed homeologues compared to the ‘submissive’ genome(s)\(^{109}\). This pattern of biased fractionation has not been observed in ancient autopolyploids\(^{111,112}\), and is believed to be allopolyploid-specific.

Studying genome expression dominance in allopolyploids is useful for understanding or predicting which of the subgenomes will likely be refractory to, and which will likely experience this fractionation process more, over time\(^{113}\). Subgenome dominance in expression has been reported for a number of recent allopolyploids such as strawberry\(^6\), peanut\(^7\), *Spartina*\(^8\), *T. miscellus*\(^9\), monkeyflower\(^{12}\) and synthetic *B. napus*\(^11\). However, some allopolyploids display even subgenome expression: *Capsella bursa-pastoris*\(^6,12\), *Trifolium repens*\(^13\), *Arabidopsis kamachatka*\(^7\) and *Brachypodium hybridum*\(^14\).

Subgenome dominance is linked to differences in transposon content\(^2\) and/or large genetic differences between subgenomes\(^4\). This makes *A. suecica*, with 6 Ma divergence between the gene-dense *A. thaliana* and the transposon-rich *A. arenosa*, a promising candidate to study this phenomenon. Previous reports on subgenome dominance in *A. suecica* are conflicting, suggesting a bias to either the *A. thaliana*\(^{116}\) or the *A. arenosa*\(^{117}\) subgenome.

To investigate the evolution of gene expression in *A. suecica*, we generated RNA sequencing (RNA-seq) data for 15 natural *A. suecica* accessions, 15 closely related *A. thaliana* accessions, 4 *A. arenosa* individuals, a synthetically generated *A. suecica* from a lab cross (the second and third hybrid generations) and the parental lines of this individual, 15 closely related *A. arenosa* accessions, 4 *A. thaliana* individuals, a synthetically generated *A. suecica* from a lab cross (the second and third hybrid generations) and the parental lines of this individual. Each sample had 2–3 biological replicates (Supplementary Data 2). On average, we obtained 10.6 million raw reads per replicate, of which 7.6 million reads were uniquely mapped to the *A. suecica* reference genome and 14,041 homeologous gene pairs (see Methods). On average, around 1% of *A. thaliana* and around 6% of *A. arenosa* RNA reads cross-mapped between the subgenomes of *A. suecica*. The approximately 6% of cross-mapping in *A. arenosa* is likely to be because of the high level of polymorphism in this species.
outcrossing species. However, diversity within *A. suecica* is massively lower, meaning that transcripts from the *A. arenosa* subgenome will probably map correctly (see Methods and Extended Data Fig. 6).

Examining expression differences between homeologous gene pairs, we found no general bias towards a subgenome of *A. suecica* (that is, the mean log fold change is 0), for any sample or tissue, including synthetic *A. suecica* (Fig. 4a and Extended Data Fig. 7). This suggests that the expression differences between the subgenomes have not changed systematically through polyploidization, and is in contrast to previous studies, which reported a bias towards the *A. arenosa* subgenome, likely because RNA-seq reads were not mapped to an appropriate reference genome.

The set of genes that show large expression differences between the subgenomes are not biased towards any particular GO category, and are not consistent between accessions and individuals (Fig. 4b and Extended Data Fig. 7). This suggests that many large subgenome expression differences are due to genetic polymorphisms within *A. suecica* rather than fixed differences between the ancestral species.

Levels of expression dominance varied to vary across tissues in natural *C. bursa-pastoris* and resynthesized cotton. To test whether expression dominance can vary for tissue-specific genes, we examined homeologous gene pairs in which at least one gene in the pair showed tissue-specific expression, in whole rosettes and floral buds. We do not find evidence for dominance between subgenomes in tissue-specific expression (Fig. 4b). A total of 897 genes with significant expression in whole rosettes for both homeologues showed GO overrepresentation that included both photosynthesis- and chloroplast-related functions (Supplementary Table 1). This result suggests that the *A. arenosa* subgenome has established important cyto-nuclear communication with the chloroplast inherited from *A. thaliana*, rather than being silenced. A total of 2,176 gene pairs with floral-bud-specific expression for both homeologues were overrepresented for GO terms related to responses to auxin and jasmonic acid that may reflect early developmental changes in this young tissue (Supplementary Table 1). Although flowers of selfing *A. thaliana* and *A. suecica* are scentless and are much smaller than those of the outcrosser *A. arenosa*, this result suggests that the ‘selfing syndrome’ has not hugely affected the transcriptome of floral buds in *A. suecica* at this stage of development.

In summary, we find no evidence that one subgenome is dominant and contributes more to the functioning of *A. suecica*. On the contrary, homeologous gene pairs are strongly correlated in expression across tissues.

**Fig. 4 | Patterns of gene expression between the subgenomes of *A. suecica* in rosettes and floral buds.** a. Violin plots of the mean log fold change (logFC) between the subgenomes for the 15 natural *A. suecica* accessions and 2 synthetic lines for whole rosettes. The centre line is the 50th percentile or median. The box limits represent the interquartile range. The whiskers represent the largest and smallest value within 1.5 times the interquartile range above and below the 75th and 25th percentile, respectively. Mean log fold change for the two accessions (ASS3 and ASS30) for which transcriptome data for both whole rosettes and flower buds were available. All the distributions are centred around zero, suggesting even subgenome expression. b. Violin plots for the mean log fold change between the subgenomes for gene pairs with tissue-specific expression in at least one member of the pair.

### Evolving gene expression in *A. suecica*.

The previous section focused on differences in expression between the subgenomes within the same individual. This section will focus on differences between individuals. To provide an overview of expression differences between individuals we performed a principal component analysis (PCA) on gene expression separately for each (sub-) genome. For both subgenomes, the first principal component separates *A. suecica* from the ancestral species and the synthetic hybrid (Fig. 5a,b and Extended Data Fig. 8), suggesting that hybridization does not automatically result in large-scale transcriptional changes, and that gene expression changes in natural *A. suecica* have evolved over time. Given the limited time involved, and the fact that the genes that have changed expression are not random with respect to function (Fig. 5c), we suggest that the first principal component captures trans-regulated expression changes in *A. suecica* that are adaptive.

To further characterize expression changes in natural *A. suecica*, we analysed differentially expressed genes (DEGs) on each subgenome compared to the corresponding ancestral species. The
Fig. 5 | Differential gene expression analysis in A. suecica. Patterns of differential gene expression in A. suecica support adaptation to the whole-genome duplication for the A. thaliana subgenome and adaptation to the new plastid environment for the A. arenosa subgenome. a, PCA for A. thaliana and the A. thaliana subgenome of natural and synthetic A. suecica lines. Principal component 1 (PC1) separates natural A. suecica from the ancestral species and the synthetic lines. PC1 separates natural A. suecica from the ancestral species and the synthetic lines, whereas PC2 identifies outlier accessions discussed further below (see Fig. 6). b, PCA for A. arenosa and the A. arenosa subgenome of natural and synthetic A. suecica lines. b, c, d, Heat map of DEGs for the A. thaliana (c) and the A. arenosa (d) subgenome of A. suecica. Positive numbers (red colour) indicate higher expression. Genes and individuals have been clustered on the basis of similarity in expression, resulting in the clusters that are discussed in the text. e, GO enrichment for each cluster in c and d. Categories discussed in the text are highlighted. RNA Pol II, RNA polymerase II.
total number of DEGs was 4,186 and 4,571 for the A. thaliana and A. arenosa subgenome, respectively (see Methods and Supplementary Data 2). These genes were clustered on the basis of the pattern of change across individuals (Fig. 5c,d) and GO enrichment analysis was performed for each cluster (Fig. 5e and Supplementary Table 2).

For the A. thaliana subgenome, we identified three clusters. Cluster 1 comprised 2,135 genes that showed decreased expression in A. suecica compared to A. thaliana. These genes are enriched for transcriptional regulation, which may be expected as we are examining DEGs between the species. Enrichments for circadian rhythm function and phototropism may be related to the ecology of A. suecica and its post-glacial migration to Fennoscandia (Fig. 1a).

Cluster 2 consisted of 468 genes that are upregulated in both natural and synthetic A. suecica relative to A. thaliana. These expression changes most likely are an immediate consequence of trans-regulation in hybrids. Genes in this cluster are enriched for mRNA transport and protein folding. Adjustment of protein homeostasis has been reported previously in experimentally evolved stable polyploid yeast194. Notably, the synthetic lines used in the expression analysis did not show signs of aneuploidy (Extended Data Fig. 9).

Cluster 3 consisted of 1,583 genes that are upregulated in A. suecica compared to A. thaliana, and several of the enriched GO categories, such as microtubule-based movement, cytokinesis, meiosis and cell division, suggest that the A. thaliana subgenome of A. suecica is adapting to polyploidy at a cellular level. Selection for this seems likely given that aneuploidy is frequent in synthetic A. suecica (Extended Data Fig. 9), while natural A. suecica has a stable conserved karyotype. Independent evidence for adaptation to polyploidy via modifications of the meiotic machinery in the other ancestor of A. suecica, A. arenosa, also exists123,124, although we see very little overlap in the genes involved (Extended Data Fig. 9).

The nature of these changes in the A. thaliana subgenome of A. suecica will require further investigation, but we note the enrichment (see Methods and Supplementary Data 2) of MYB family transcription factor binding sites19 in cluster 3.

For the A. arenosa subgenome, we also found three clusters of DEGs (Fig. 5d) with GO enrichment for two clusters (Fig. 5e and Supplementary Table 2). Cluster 1 consisted of 1,278 genes that are upregulated in natural A. suecica compared to A. arenosa and synthetic A. suecica. We find enrichment for plastid-related functions that may be due to selection on the A. arenosa subgenome to restore communication with the maternally inherited plastids from A. thaliana. Twelve out of a total of 69 genes with structural evidence for direct plastid–nuclear interactions in A. thaliana overlap our genes in Cluster 1 using CyMIRA195 (P = 0.0072; one-sided Fisher’s exact test; Supplementary Data 2). Cluster 3 consists of 3,166 genes that are downregulated in A. suecica compared to A. arenosa and synthetic A. suecica. These genes were enriched for mRNA processing and epigenetic regulation of gene expression (Supplementary Table 2), and for positive regulation of transcription by RNA polymerase II, which suggests differences in the epigenetic regulation of expression between A. arenosa and A. suecica. Cluster 2 (127 genes), finally, did not have a GO overrepresentation and showed an intriguing pattern that will be discussed in the next section.

**Homeologous exchange contributes to variation in gene expression.** The second principal component for gene expression identified three outlier accessions of A. suecica: two for the A. thaliana subgenome (Fig. 5a) and one for the A. arenosa subgenome (Fig. 5b). While closely examining the latter accession, AS530, we realized that it is responsible for the cluster of genes with distinct expression patterns but no GO enrichment (Fig. 5d, Cluster 2). Genes from this cluster were significantly downregulated on the A. arenosa subgenome (Fig. 6a) and upregulated on the A. thaliana subgenome (Fig. 6b)—for AS530 only. The observation that 104 of the 127 genes (Extended Data Fig. 10) in the cluster are located in close proximity in the genome pointed to a structural rearrangement. The lack of DNA-sequencing coverage on the A. arenosa subgenome around these 104 genes, and the doubled coverage for their homeologues on the A. thaliana subgenome, suggested a homeologous exchange (HE) event resulting in AS530 carrying four copies of the A. thaliana subgenome and zero copies of the A. arenosa subgenome for this approximately 2.5 Mb region of the genome (Fig. 6c). This was further supported by Hi-C data, which showed clear evidence for interchromosomal contacts between A. thaliana subgenome chromosome 1 and A. arenosa subgenome chromosome 6 around the break points of the putative HE in AS530 (Fig. 6d,e), and by multiple discordant paired-end reads at the break points between the homeologous chromosomes, which independently support the HE event (Extended Data Fig. 10).

We examined the two outlier A. suecica accessions for the A. thaliana subgenome (Fig. 5a; AS150 and AS05) and found that they probably share a single HE event in the opposite direction (four copies of the A. arenosa subgenome and no copies of the A. thaliana subgenome for a region of around 1.2 Mb in size; see Extended Data Fig. 10). This demonstrates that HE occurs in A. suecica and contributes to the intraspecific variation in gene expression (Fig. 5a,b). HE in alloploids is a main source of diversity, causing extensive phenotypic changes144–146. However, the majority of HEs are probably deleterious as they will lead to gene loss: although the A. thaliana and A. arenosa genomes are largely syntenic, AS530 is missing 108 genes (Extended Data Fig. 10) that are only present on the A. arenosa subgenome segment that has been replaced by the homeologous segment from the A. thaliana subgenome, and AS150 and AS05 are missing 53 genes that were only present on the A. thaliana subgenome.

**Discussion**

This study has focused on the process of polyploidization in a natural allopolyploid species, A. suecica. Its ancestral species, A. thaliana and A. arenosa, differ substantially in genome size, chromosome number, ploidy, mating system and ecology.

Our main conclusion from this study is that the polyploid speciation leading to A. suecica appears to have been a gradual process rather than some kind of ‘event’. We confirmed previous results that genetic polymorphism is largely shared with the ancestral species, demonstrating that A. suecica did not originate through a single unique hybridization event, but rather through multiple crosses9. We also find no evidence for a ‘genome shock’, in the sense of major genomic changes linked to structural and functional alterations, which has often been suggested to accompany polyploidization and hybridization. Specifically, the genome has not been massively rearranged, transposable elements are not out of control and there is no subgenome dominance in expression. On the contrary, we find evidence of genetic adaptation to ‘stable’ life as a polyploid, in particular changes to the meiotic machinery and in interactions with the plastids. These findings made in natural (but not synthetic) A. suecica, together with the observation that experimentally generated A. suecica is often unviable and does exhibit evidence of genome rearrangements, similar to the young allopolyploid species in Tragopogon and monkeyflower, suggest that the most important bottleneck in polyploid speciation may be selective. If this is true, domesticated polyploids may not always be representative of natural polyploidization. Darwin famously argued that evolution is gradual147–149—we suggest that natural polyploids are no exception from this, but note that many more species will have to be studied before it is possible to draw general conclusions.

**Methods**

PacBio sequencing of A. suecica. We used genomic DNA from whole rosettes of one A. suecica (ASS3) accession to generate PacBio sequencing data. DNA was extracted using a modified PacBio protocol for preparing Arabidopsis

---

1374 | NATURE ECOLOGY & EVOLUTION | VOL 5 | OCTOBER 2021 | 1367-1381 | www.nature.com/natecolevol
Fig. 6 | Homeologous exchange contributes to expression variance within A. suecica. a, Cluster 2 of Fig. 5d explains the outlier accession AS530, which is not expressing a cluster of genes on the A. arenosa subgenome. b, Homeologous genes of this cluster on the A. thaliana subgenome of A. suecica show the opposite pattern and are more highly expressed in AS530 compared to the rest of the population. c, Of the 122 genes from cluster 3, 97 are close to each other on the reference genome but appear to be deleted in AS530 on the basis of sequencing coverage. d, The A. thaliana subgenome homeologues have twice the DNA coverage, suggesting that they are duplicated. e, f, Hi-C data show (spurious) interchromosomal contacts at 25 kb resolution between chromosome 1 and chromosome 6 around the break point of the cluster of 97 genes in AS530 (f) but not in the reference accession ASS3 (e).
genomic DNA for size-selected approximately 20-kb SMRTbell libraries. In brief, whole-genome DNA was extracted from 32 g of 3–4-week-old plants, grown at 16 °C and subjected to a 2-day dark treatment. This generated 23 μg of purified genomic DNA with a fragment length of more than 40kbp for A. suecica. We assessed DNA quality with a Qubit fluorometer and a Nanodrop analysis and ran the DNA on a gel to visualize fragmentation. Genomic libraries and single-molecule real-time (SMRT) sequence data were generated at the Functional Genomics Center Zurich (FGCZ). The PacBio RSII instrument was used with P6-C4 chemistry and an average movie length of 6 hours. A total of 12 SMRT cells were processed, generating 16.3 Gb of DNA bases with an N50 read length of 20 kb and median read length of 14 kb. Using the same genomic library, an additional 3.3 Gb of data was generated by a PacBio Sequel instrument at the Vienna BioCenter Core Facilities (VBCF), with a median read length of 10 kbp.

**A. suecica genome assembly.** To generate the A. suecica assembly we first used FALCON+ (v.0.3.0) with a length cut-off for seed reads set to 1 kb in size. The assembly produced 828 contigs with an N50 of 5.81 Mb and a total assembly size of 271 Mb. In addition, we generated a Canu (v.1.3.0) assembly using default settings, which resulted in 260 contigs with an N50 of 6.65 Mb and a total assembly size of 267 Mb. Then we merged the two assemblies using the software quickmerge. The resulting merged assembly consisted of 929 contigs with an N50 of 9.02 Mb and a total draft assembly size of 276 Mb. We polished the assembly using Arrow (smrtlink release 5.0.0.6792 and Pilon (v.1.22). For Pilon115, 100 bp (with PCR duplicates removed), and a second PCR-free 250 bp Illumina paired-end reads were used that had been generated from the reference A. suecica accession ASS3.

**PacBio sequencing of A. arenosa.** A natural Swedish autotetraploid A. arenosa accession, Aa4, was inbred in a lab for two generations to reduce heterozygosity. We extracted whole-genomic DNA from 64 g of 3-week-old plants in the same way as described for A. suecica, generating 50 μg of purified genomic DNA with a fragment length of more than 40kbp in length. The whole genome of A. arenosa and SMRT sequence data were generated at the VBCF. A PacBio Sequel instrument was used to generate a total of 22 Gb of data from five SMRT cells, with an N50 of 9.02 Mb and a total draft assembly size of 276 Mb. We polished the assembly using Arrow (smrtlink release 5.0.0.6792 and Pilon (v.1.22). For Pilon115, 100 bp (with PCR duplicates removed), and a second PCR-free 250 bp Illumina paired-end reads used that had been generated from the reference A. suecica accession ASS3.

**Assembly of autotetraploid A. arenosa.** We assembled a draft contig assembly for the autotetraploid A. arenosa accession Aa4 using FALCON (v.0.3.0) as for A. suecica. The assembly produced 3,629 contigs with an N50 of 331 kbp, a maximum contig size of 2.5 Mb and a total assembly size of 461 Mb. The assembly size is greater than the calculated haploid size of 330 Mb using the genome size of A. arenosa. To generate the draft assembly, we used paired-end Illumina reads from the Hi-C genome, produced with LACHESIS. Reference-guided scaffolding of the draft assembly was the paternal plant in this cross. We managed to obtain very few F1 hybrid plants, which after one round of selfing set higher levels of seeds.

**Construction of the A. suecica genetic map.** We crossed the different A. suecica accession AS150 with the reference accession ASS3. The cross was uni-directional with AS150 as the maternal and ASS3 as the paternal plant. A total of 192 F1 plants were collected. We multiplexed the samples on 96-well plates using 75-bp paired-end reads and generated data of 1–2x coverage per sample. Samples were mapped to the draft-repeated masked chromosome of the reference A. suecica genome using BWA-MEM123 (v.0.7.15). SAMtools124 (v.0.1.19) was used to filter reads for proper pairs and a minimum mapping quality of 5. We called single nucleotide polymorphisms (SNPs) using Samtools mpileup giving a total of 590,337 SNPs. We required sites to have non-zero coverage in a minimum of 20 individuals and filtered SNPs to have frequency between 0.45–0.55 in the F1 population. We removed F1 individuals that did not have genotype calls for more than 90% of the data. This resulted in 183 individuals with genotype calls for 334,257 SNPs.

We applied a hidden Markov model implemented in the R package HMM125 to classify SNPs as homozygous or heterozygous for each of our F1 lines. We then divided the genome into 500-kb non-overlapping windows, and classified each window as homozygous or heterozygous. This was done per chromosome and the resulting file for each chromosome and their markers were processed in the R package qtl126, to generate a genetic map. Markers genotyped in fewer than 100 F1 individuals were excluded from the analysis. Linkage groups were assigned with a minimum log of the odds (LOD) score of 8 and a maximum recombination fraction of 0.05. We defined the flanking marker order by the best LOD score and the lowest number of crossover events.

We corrected the erroneous placement of a contig at the beginning of chromosome 1 of the A. arenosa subgenome. The misplaced contig was relocated from chromosome 1 to the pericentromeric region of chromosome 2 of the A. arenosa subgenome in A. suecica. Chromosome 2 of the A. suecica subgenome was previously shown to be largely devoid of intraspecific variation, resulting in few markers for this chromosome. Therefore, this chromosome-scale scaffold was assembled by the manual inspection of 3D-proximity information based on our Hi-C sequencing using the software Juicebox127.

**Gene prediction and annotation of the A. suecica genome.** We combined de novo and evidence-based approaches to predict protein-coding genes. For de novo prediction, we trained AUGUSTUS128 on the set of conserved single-copy genes using BUSCO129 separately on A. thaliana and A. arenosa subgenomes of A. suecica. The evidence-based approach included both homology to the protein sequences of the ancestral species and the transcriptome of A. suecica. We aligned the peptide sequences from the A. suecica genome to the A. arenosa subgenome of A. suecica, while the peptides from A. lyrata annotation130 (Alyrata_384_v2.1) were aligned to the A. arenosa subgenome of A. suecica using GenomeThreader131 (v.1.7.0). We mapped the RNA-seq reads from the reference accession of A. suecica (ASS3) from the rosettes and flower bud tissues to the reference genome using TopHat132 and generated intron hints from the split reads using the R package GenomeThreader. We split the alignment into A. thaliana and A. arenosa subgenomes and assembled the transcriptome of A. suecica for each subgenome separately in the genome-guided mode with Trinity133 (v.2.6.6). Separately for each of the subgenomes, we filtered the assembled transcripts using a transcripts per million (TPM) cut-off set to 1, collapsed similar transcripts using CD-HIT134 and chose the longest open reading frame from the six-frame translation. We algorithmically annotated the proteins from A. thaliana and A. arenosa parts of A. suecica to the corresponding subgenomes using GenomeThreader (v.1.7.0). We ran AUGUSTUS using retrained parameters from BUSCO and merged hints from all three sources, these being: (1) intron hints from A. suecica RNA-seq; (2) homology hints from ancestral proteins; and (3) hints from A. suecica proteins.

RepeatModeler135 (v.0.11.0) was used to build a de novo TE consensus library for A. suecica and identify repetitive elements based on the genome sequence. Genome locations of the identified TE repeats were determined by using RepeatMasker136 (v.4.0.7) and filtered for full-length matches using a cutoff described previously137. Helitrons are the most abundant TE family in both subgenomes.

**Synthetic A. suecica lines.** To generate synthetic A. suecica we crossed a natural tetraploid A. thaliana accession (6978, also known as Wa-1) to a natural Swedish autotetraploid A. arenosa accession (6978, also known as Wa-1). We aligned the draft A. arenosa genome assembly to the A. thaliana TAIR10 reference and to our A. arenosa draft contig assembly, simultaneously. We used the MUMer command mdnadiv to produce 1-to-1 alignments. As the subgenomes are only around 86% identical, the majority of contigs could be conclusively assigned to either subgenome by examining how similar the alignments were. Contigs that could not be assigned to a subgenome on the basis of percentage identity were examined manually, and the length of the alignment used to determine subgenome assignment.

Finally, we used the software LACHESIS138 (v.1.0.0) to scaffold our draft assembly, using the reference genomes of A. thaliana and A. lyrata as a guide to assist with scaffolding the contigs (we used A. lyrata here instead of our draft A. arenosa contig assembly, as A. lyrata is a chromosome-level assembly). This produced a 13 scaffold chromosome-level assembly for A. suecica.
The resulting synthetic line was able to self-fertilize. F1 seeds were descended from a common F1, and were similar to natural A. suecica in appearance. We further continued the synthetic line to F2 (selfed third generation).

**Synteny analysis.** We performed an all-against-all BLASTP search using CDS sequences for the reference A. suecica genome and the ancestral genomes, A. thaliana and A. lyrata. We used the SynMap tool from the online CoGe portal. We examined synteny using the default parameters for DAGChanger (maximum distance between two matches = 20 genes; minimum number of aligned pairs = 5 genes).

**Estimating the copy number of rDNA repeats using short DNA reads.** To measure the copy number of 45S rRNA repeats in our populations of different species, we aligned short DNA reads to a single reference 45S consensus sequence of A. thaliana. An A. arenosa 45S rRNA consensus sequence was constructed by aligning the best hit of BLAST with our A. arenosa contig draft assembly. This hit matched position 1571–8522 bp of the A. thaliana consensus sequence, was 6,647 bp in length and is 97% identical to the A. thaliana 45S rRNA consensus sequence. The aligned regions of these two 45S rRNA consensus sequences, determined by BLAST, were used in copy number estimates, to ensure that the sizes of the sequences were equal. The relative increase in sequence coverage of these loci, when compared to the mean coverage for the reference genome, was used to estimate copy number.

**Plant material for RNA-seq.** Transcripтомic data generated in this study included 15accessions of A. suecica, 16 accessions of A. thaliana, 4 accessions of A. arenosa and 2 generations of an artificial A. suecica line (the second and third selfed generation). The sibling of a paternal A. arenosa parent (Aa4) and the maternal tetraploid A. thaliana parent (978, Wa-1) of our artificial A. suecica line were included as part of our samples (Supplementary Data 1). Each accession was replicated three times. Seeds were stratified for 4 days at 4 °C in 1 ml of sterilized water. Seeds were then transferred to pots in a controlled growth chamber at 21 °C. Humidity was kept constant at 60%. Pots were thinned to two to three seedlings after one week. Pots were re-randomized each week in their trays. Whole rosettes were collected when plants reached the seven-to-nine true-leaf stage of development. Samples were collected between 14:00 and 17:00 and flash-frozen in liquid nitrogen.

**RNA extraction and library preparation.** For each accession, two to three whole rosettes in each pot were pooled and total RNA was extracted using the ZR Plant RNA MiniPrep kit. We treated the samples with DNase and purified isolation of mRNA using the AMPure XP magnetic beads and the Poly(A) RNA Selection Kit from Lexogen. RNA quality and degradation were assessed using the RNA Fragment Analyzer (DNR-471 stranded sensitivity RNA analysis kit, 15 nt). The concentration of RNA per sample was measured using the Qubit fluorometer. Library preparation was carried out following the NEBNext Ultra II RNA Library Prep Kit for Illumina. Barcoded adapters were ligated to the 3′ ends of RNA. 150 bp paired-end sequencing was carried out on the VCRF on Illumina (HiSeq 2500) using multiplexing.

**RNA-seq mapping and gene expression analysis.** We mapped 125-bp paired-end reads to the de-novo assembide A. suecica reference using STAR.25 and we filtered for primary and uniquely aligned reads using the parameters-outfilterMultimapNmax 1–outSamprimaryFlag OneBestScore. We quantified reads assigned to the A. thaliana genome using–quantMode GeneCounts.

**Cross-mapping of short reads.** Cross-mapping of short RNA reads between the subgenomes of A. suecica was measured by mixing the RNA reads between A. thaliana and A. arenosa individuals to generate ‘in-silico’ A. suecica individuals. We mapped reads from 10 in-silico A. suecica individuals to the A. suecica genome. We compared different RNA-seq pipelines to determine cross-mapping error rates. We mapped reads using STAR (v. 2.7), HISAT2 (v. 2.1) and EAGLE (v. 1.0). Around 1% of the A. arenosa reads map to the A. thaliana subgenome and around 6% of the A. arenosa reads map to the A. thaliana subgenome, regardless of mapping strategy or pipeline (see Extended Data Fig. 6).

**Expression analysis of rRNA.** RNA reads were mapped in a similar manner to DNA reads for the analysis of rDNA copy number. Expression analysis was performed in a similar manner to protein-coding genes, in edgeR. We defined the exclusive expression of a particular 45S rRNA gene by taking a cut-off of 15 for log,(CPM) as this was the maximum level of cross-mapping we observed for the ancestral species (see Extended Data Fig. 3).

**Expression analysis of transposable elements.** To analyse the expression of transposable elements between species, the annotated TE consensus sequences in A. suecica were aligned using BLAST, v. all. Highly similar TE sequences (more than 85% similar for more than 85% of the TE sequence length), were removed, leaving 813 TE families out of 1,215. Filtered A. suecica TEs were aligned to annotated A. thaliana (TAIR10) and A. arenosa (the PacBio contig assembly presented in this study) TE sequences to assign each family to an ancestral species using BLAST. A total of 208 TE families were assigned to the A. thaliana parent and 171 TE families were assigned to the A. arenosa parent.

**RNA reads were mapped to TE sequences using a similar approach as for gene expression analysis using edgeR that showed expression of more than 85% of the reads map to the TE sequences.** A total of 121 A. thaliana TE sequences and 93 A. arenosa TE sequences passed this threshold. We took the mean of replicates per accession for further downstream analyses.

**GO enrichment analysis.** We used the R package TopGO to conduct GO enrichment analysis. We used the ‘weight01’ algorithm when running TopGO, which accounts for the hierarchical structure of GO terms and thus implicitly corrects for multiple testing. GO annotations were based on the A. thaliana orthologue of A. suecica genes. Gene annotations for A. thaliana were obtained using the R pattern biomaRt from Ensembbl.‘biomaRt::useMart(biomart = ‘plants_mart’, dataset = ‘athaliana_eg_gene’, host = ‘plant.ensembl.org’).

**Genome size measurements.** We measured genome size for the reference A. suecica accession AS53 and the A. arenaosa accession used for PacBio (Aa4), using Solanum lycopersicum cv. Stupicke (2C = 1.96 pg DNA) as the standard.

**GO enrichment analysis.** We used the R package TopGO to conduct GO enrichment analysis. We used the ‘weight01’ algorithm when running TopGO, which accounts for the hierarchical structure of GO terms and thus implicitly corrects for multiple testing. GO annotations were based on the A. thaliana orthologue of A. suecica genes. Gene annotations for A. thaliana were obtained using the R pattern biomaRt from Ensembbl.‘biomaRt::useMart(biomart = ‘plants_mart’, dataset = ‘athaliana_eg_gene’, host = ‘plant.ensembl.org’).

**Genome size measurements.** We measured genome size for the reference A. suecica accession AS53 and the A. arenaosa accession used for PacBio (Aa4), using Solanum lycopersicum cv. Stupicke (2C = 1.96 pg DNA) as the standard.

**GO enrichment analysis.** We used the R package TopGO to conduct GO enrichment analysis. We used the ‘weight01’ algorithm when running TopGO, which accounts for the hierarchical structure of GO terms and thus implicitly corrects for multiple testing. GO annotations were based on the A. thaliana orthologue of A. suecica genes. Gene annotations for A. thaliana were obtained using the R pattern biomaRt from Ensembbl.‘biomaRt::useMart(biomart = ‘plants_mart’, dataset = ‘athaliana_eg_gene’, host = ‘plant.ensembl.org’).

**Genome size measurements.** We measured genome size for the reference A. suecica accession AS53 and the A. arenaosa accession used for PacBio (Aa4), using Solanum lycopersicum cv. Stupicke (2C = 1.96 pg DNA) as the standard.

**GO enrichment analysis.** We used the R package TopGO to conduct GO enrichment analysis. We used the ‘weight01’ algorithm when running TopGO, which accounts for the hierarchical structure of GO terms and thus implicitly corrects for multiple testing. GO annotations were based on the A. thaliana orthologue of A. suecica genes. Gene annotations for A. thaliana were obtained using the R pattern biomaRt from Ensembbl.‘biomaRt::useMart(biomart = ‘plants_mart’, dataset = ‘athaliana_eg_gene’, host = ‘plant.ensembl.org’).

**Genome size measurements.** We measured genome size for the reference A. suecica accession AS53 and the A. arenaosa accession used for PacBio (Aa4), using Solanum lycopersicum cv. Stupicke (2C = 1.96 pg DNA) as the standard.
to that of each sample to calculate the 2C DNA content of that sample using the equation:  
\[ \text{Sample 2C DNA content} = \frac{\text{(sample G1 peak mean) × standard G1 peak mean)}}{\text{standard 2C DNA content}} \]

We also measured genome size for the reference species and suitable slides were pretreated by RNase (100 µg ml⁻¹) and digested by a 0.3% mix of pectolytic enzymes use. Selected inflorescences were rinsed in distilled water and citrate buffer and 

\[ \text{Chromosome preparation and FISH.} \]

We used PoPoolationTE20 (v1.10.04) to identify TE insertions in the genome of each of our species, A. suecica, A. thaliana and A. arenosa. We used a mapping quality of 10 for the read in discordant read pairs to calculate the location and abundance of a TE in the genome for an accession of interest. We mapped 100-bp Illumina DNA reads from previous studies, in addition to our newly generated synthetic A. suecica using BWA-MEM. We as the data format required by PoPoolationTE2. Reads were given an increased penalty of 15 for being unpaired. Reads were de-duplicated using SAMTools and rmdup (v.1.9). The resulting bam files were then provided to PoPoolationTE2 to identify TE insertions in the genome of each of our species, A. suecica, A. thaliana and A. arenosa. We used a mapping quality of 10 for the read in discordant read pairs to calculate the location and abundance of a TE in the genome. We used the 'gene' filters (AHF Analysentechnik). The monochromatic images were pseudo-coloured and merged using Adobe Photoshop CS6 software (Adobe Systems).

DAP-seq enrichment analysis for transcription factor target genes. We downloaded the target genes of transcription factors from the plant cistrome database (http://neomorph.salk.edu/dap_web/pages/index.php), which is a collection of transcription-factor-binding sites and their target genes, in A. thaliana, based on DAP-seq. To test for enrichment of a gene set (for example, the genes in A. thaliana cluster 2 on Fig. 5) for target genes of a particular transcription factor, we performed a hyper-geometric test in R. As a background we used the total 14,041 genes used in our gene expression analysis. We then performed FDR correction for multiple testing to calculate an accurate P value of the enrichment.

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

Data availability

Genome assemblies and raw short reads can be found in the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/home), The genome assembly for A. suecica ASS3 can be found under the BioProject number PRJEB42198, assembly accession GCA_905175345. The raw reads for the A. suecica genome assembly generated by PacBio RSII can be found under ERR5037702 and those from Sequel under ERR5031296. The Hi-C reads used for scaffolding the A. suecica assembly can be found under ERR5032369. The contig assembly for tetraploid A. arenosa (ssp. arenosa) can be found under the BioProject number PRJEB42276, assembly accession GCA_905175405. The raw reads for the A. arenosa A4 contig assembly generated by Sequel can be found under ERR5031542 and the reads generated by Nanopore under ERR5031541. The Hi-C reads for the A. arenosa assembly can be found under ERR5032367. Hi-C sequencing data for the ancestral species, the outlier accession AS530 and synthetic A. suecica can be found under the BioProject number PRJEB42290, DNA sequencing data of synthetic A. suecica and parents generated in this study can be found under the BioProject number PRJEB42291. The RNA-seq reads are under the BioProject number PRJEB42277. TE presence or absence calls for A. suecica and the ancestral species can be found in Supplementary Data 1. A list of DEGs, orthologues, enriched DAP-seq transcription factors, CyMIRA gene overlaps and RNA-seq mapping statistics can be found in Supplementary Data 2. Log fold change and CPM for genes on the A. thaliana and A. arenosa subgenomes can be found in Supplementary Data 3. The gene annotation (gff3 file) of the A. suecica genome can be found in Supplementary Data 4. TE consensus sequences and a hierarchy file of TE order for A. suecica can be found in Supplementary Data 5.

Received: 11 September 2020; Accepted: 1 July 2021; Published online: 19 August 2021

References

1. DeBruyn, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. Nat. Rev. Genet. 18, 411–424 (2017).
2. Ortiz, S. P. & Solís, D. E. Ancient WGD events as drivers of key innovations in angiosperms. Curr. Opin. Plant Biol. 30, 159–165 (2016).
3. Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3, e314 (2005).
4. Li, Z. et al. Multiple large-scale gene and genome duplications during the evolution of hexapods. Proc. Natl Acad. Sci. USA 115, 4713–4718 (2018).
5. Chen, J. Z. et al. Genome diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. Nat. Genet. 52, 525–533 (2020).
6. Edger, P. P. et al. Origin and evolution of the octoploid strawberry genome. Nat. Genet. 51, 541–547 (2019).
7. Ramirez-González, R. H. et al. The transcriptional landscape of polyploid wheat. Science 361, eaar6089 (2018).
8. Zhang, W. et al. The genome of cultivated peanut provides insights into legume karyotypes, polyploid evolution and crop domestication. Nat. Genet. 51, 865–876 (2019).
9. Bertoli, D. J. et al. The genome sequence of segmental allopolyploid peanut Arachis hypogaea. Nat. Genet. 51, 877–884 (2019).

10. Kasinov, A. S. et al. High-quality genome assembly of Capsella bursa-pastoris reveals symmetry of regulatory elements at early stages of polyploid genome evolution. Plant J. 91, 278–291 (2017).

11. Kryvokhyzha, D. et al. Towards the new normal: transcriptomic convergence and genomic legacy of the two subgenomes of an allopolyploid weed (Capsella bursa-pastoris). PLoS Genet. 15, e1008131 (2019).

12. Douglas, G. M. et al. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris. Proc. Natl Acad. Sci. USA 112, 2806–2811 (2015).

13. Griffiths, A. G. et al. Breaking free: the genomes of allopolyploidy-facilitated niche expansion in white clover. Plant Cell 31, 1466–1487 (2019).

14. Gordon, S. P. et al. Gradual polyploid genome evolution revealed by pan-genomic analysis of Brachypodium distachyon, B. stacei, and B. hybridum (Poaceae). An. Jard. Bot. Made. 73, e028 (2016).

15. Catalán, P., López-Álvarez, D., Bellosta, C. & Villar, L. Updated taxonomic descriptions, iconography, and habitat preferences of Brachypodium distachyon, B. stacei, and B. hybridum (Poaceae). An. Jard. Bot. Made. 73, e028 (2016).

16. Paape, T. et al. Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. Nat. Commun. 9, 3909 (2018).

17. Edger, P. P. et al. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. Plant Cell 29, 2150–2167 (2017).

18. Solís, D. E. et al. Recent and recurrent polyploidy in Tragopogon (Asteraceae): cytogenetic, genome size, and genetic comparisons. Biol. J. Linn. Soc. 82, 485–504 (2004).

19. te Beest, M. et al. The more the better? The role of polyploidy in facilitating plant invasions. Annu. Bot. 109, 19–45 (2012).

20. Novikova, P. Y. et al. Genome sequencing reveals the origin of the allopolyploid Arabidopsis suecica. Mol. Biol. Evol. 34, 957–968 (2017).

21. Fowler, N. L. & Levin, D. A. Ecological constraints on the establishment of a novel polyploid in competition with its diploid progenitor. Nat. Ann. 124, 703–711 (1994).

22. Bomblies, K. & Madlung, A. Polyploidy in the Arabidopsis genus. Chromosoma Res. 22, 117–134 (2014).

23. Hollister, J. D. et al. Genetic adaptation associated with genome-doubling in autotetraploid Arabidopsis arenosa. PLoS Genet. 8, e1003093 (2012).

24. 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 125, 287–300 (2016).

25. Leitch, A. R. & Leitch, I. J. Genomic plasticity and the diversity of polyploid plants. Science 320, 481–483 (2008).

26. Bottani, S., Zabet, N. R., Wendel, J. F. & Veitia, R. A. Gene expression dominance in allopolyploids: hypotheses and models. Trends Plant Sci. 23, 593–592 (2018).

27. Parissod, C. et al. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol. 186, 37–45 (2010).

28. McClintock, B. The significance of responses of the genome to challenge. Science 226, 792–801 (1984).

29. Feldman, M. et al. Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. Genetics 147, 1381–1387 (1997).

30. Zhang, H. et al. Transcriptome shock invokes disruption of parental expression-conserved genes in tetraploid wheat. Science 320, 481–483 (2008).

31. Wang, X. et al. Transcriptome asymmetry and expression level dominance in resynthesized allopolyploid Brassica napus. BMC Genomics 19, 586 (2018).

32. Zhao, T. et al. IncRNAs in polyploid cotton interspecific hybrids are functionally constrained. Genome Biol. 19, 195 (2018).

33. Yoo, M.-J., Szadkowski, E. & Wendel, J. F. Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome 44, 321–330 (2001).

34. Kryvokhyzha, D. et al. Towards the new normal: transcriptomic activation of retrotransposons alters the expression of adjacent genes in wheat. Nat. Genet. 33, 102–106 (2003).

35. Kiraštein, Z., Yaakov, B., Khadain, V. & Kaskhul, K. Genetic and epigenetic dynamics of a retrotransposon after polyploidization of wheat. Genetics 186, 801–812 (2010).

36. Yang, Z. et al. Extensive intraspecific gene order and gene structural alterations during allopolyploid speciation. Mol. Biol. Evol. 34, 957–968 (2017).

37. Snabre, J. C., Springer, N. M. & Freling, M. Differentiation of the maize subgenomes by genomic imprinting. Mol. Biol. Evol. 34, 957–968 (2017).

38. Qian, W. et al. Phylogenomic analysis of wild and cultivated allotetraploid Brassica napus. Nat. Commun. 5, 512–517 (2014).

39. Wang, M. et al. Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. Nat. Plants 4, 90–97 (2018).

40. Cheng, F. et al. Biased gene fractionation and dominant gene expression among the subgenomes of Brassica napus. PLoS One 7, e36442 (2012).

41. Qin, L., Qian, W. & Snowdon, R. J. Sub-genomic selection patterns as a signature of breeding in the allopolyploid Brassica napus. BMC Genomics 15, 1170 (2014).

42. Yang, Z. et al. The interplay of demography and selection during maize domestication and expansion. Genome Biol. 18, 215 (2017).

43. Chen, Y.-C. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 181, 145–161 (2020).

44. Liu, Y. et al. Pan-genome of wild and cultivated soybean. Cell 182, 162–176 (2020).

45. Zhou, Y.-C. et al. The population genetics of structural variants in grapevine and their effects on genomic variation in crop domestication. Nat. Plants 5, 950–953 (2019).

46. Wu, J. et al. Homoeolog expression bias and expression level dominance in resynthesized allopolyploid Brassica napus. BMC Genomics 19, 586 (2018).

47. Chelaifa, H., Monnier, A. & Ainouche, M. Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina townsendii and Spartina anglica (Poaceae). New Phytol. 186, 161–174 (2010).
95. Vicient, C. M. & Casacuberta, J. A. Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* **120**, 195–207 (2017).

96. Ungerer, M. C., Strakosh, S. C. & Zhen, Y. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Carr. Biol.* **16**, R872–R873 (2006).

97. Rieseberg, L. H. et al. Ecological transitions in wild sunflowers facilitated by hybridization. *Science* **301**, 1211–1216 (2003).

98. Cavrak, V. V. et al. How a retrotransposon exploits the plant's heat stress response for its activation. *Plos Genet.* **10**, e1004115 (2014).

99. Göbel, U. et al. Robustness of transposable element regulation but no genomic shock observed in interspecific Arabidopsis hybrids. *Genome Biol. Evol.* **10**, 1403–1415 (2018).

100. Koller, R., Gomez-Sanchez, D. & Schlotterer, C. PoPoolationTE2: Comparative population genomics of transposable elements using Pool-Seq. *Mol. Biol. Evol.* **33**, 2759–2764 (2016).

101. Lockton, S. & Gaut, B. S. The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. *BMC Evol. Biol.* **10**, 10 (2010).

102. Quadrana, L. et al. The *Arabidopsis thaliana* mobilome and its impact at the species level. *elife* **5**, e15716 (2016).

103. Stuart, T. et al. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *elife* **5**, e20777 (2016).

104. Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2**, 333–341 (2001).

105. Conant, G. C., Birchler, J. A. & Pires, J. C. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for random gene duplication. *Curr. Opin. Plant Biol.* **19**, 91–98 (2014).

106. Akioz, G. & Nordborg, M. The *Aquilegia* genome reveals a hybrid origin of core eudicots. *Genome Biol.* **20**, 256 (2019).

107. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).

108. Soltis, P. S., Marchant, D. R., Van de Peer, Y. & Soltis, D. E. Polyplody and genome evolution in plants. *Curr. Opin. Genet. Dev.* **35**, 119–125 (2015).

109. Thomas, B. C., Pedersen, B. & Freeling, M. Following tetraploidy in *Arabidopsis*: comparative population genomics of transposable elements using Pool-Seq. *Mol. Biol. Evol.* **30**, 1063–1071 (2015).

110. Gamsmeur, O. et al. Two evolutionarily distinct classes of paleopolyploids. *Mol. Biol. Evol.* **31**, 448–454 (2014).

111. Li, Q. et al. Unbiased subgenome evolution following a recent whole-genome duplication in pear (*Pyrus bretschneideri* Rehd.). *Hortic. Res.* **6**, 34 (2019).

112. Han, J. et al. Replaying the evolutionary tape to investigate subgenome dominance in *Arabidopsis thaliana*.* New Phytol.* **230**, 354–371 (2021).

113. Alcoy, B. et al. On the origin of and the genomic consequences of rapid mating system evolution. *Curr. Biol.* **32**, 1063–1071 (2015).

114. Han, J. et al. Comparative population genomics of transposable elements using Pool-Seq: evidence from nuclear DNA markers. *Curr. Opin. Plant Biol.* **23**, 1217–1231 (2006).

115. Novkova, P. Y. et al. Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).

116. Cate, M. et al. *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45**, 831–835 (2013).

117. Liu, S. et al. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.* **5**, 3930 (2014).

118. Madlung, A. et al. Genomic changes in synthetic *Arabidopsis* polyploids. *Plant J.* **41**, 221–230 (2005).

119. Copenhaver, G. P. & Pikaard, C. S. Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. *Plant J.* **9**, 273–282 (1996).

120. Navashin, M. Chromosome alterations caused by hybridization and their bearing upon certain general genetic problems. *Cytologia* **5**, 169–203 (1934).

121. Tucker, S., Vitins, A. & Pikaard, C. S. Nuclear dominance and ribosomal RNA gene silencing. *Curr. Opin. Cell Biol.* **22**, 351–356 (2010).

122. Maciak, S., Michalak, K., Kale, S. D. & Michalak, P. Nuclear dominance and repression of 43S ribosomal RNA genes in hybrids between *Xenopus borealis* and *X. muelleri* (*2n = 36*). *Cytogetic Genet. Res.* **149**, 290–296 (2016).

123. Kräuszyck, T. et al. Immediate unidirectional epigenetic reprogramming of NORs occurs independently of rDNA rearrangements in synthetic allopolyploid *Arabidopsis thaliana*. *Curr. Opin. Genet. Dev.* **23**, 371–377 (2013).

124. Chen, Z. J., Comai, L. & Pikaard, C. S. Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nuclear dominance) in *Arabidopsis* allopolyploids. *Proc. Natl Acad. Sci. USA* **95**, 14891–14896 (1998).

125. Pontes, O. et al. Postembryonic establishment of megabase-scale gene silencing in nuclear dominance. *PLoS One* **2**, e1157 (2007).

126. Lewis, M. S. & Pikaard, C. S. Restricted chromosomal silencing in nuclear dominance. *Proc. Natl Acad. Sci. USA* **98**, 14536–14540 (2001).

127. Pontes, O. et al. Chromosomal locus rearrangements are a rapid response to formation of the allopolyploid *Arabidopsis suecica* genome. *Proc. Natl Acad. Sci. USA* **101**, 18240–18245 (2004).

128. Long, Q. et al. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.* **45**, 884–890 (2013).

129. Rabanal, F. A. et al. Epistatic and allelic interactions control expression of ribosomal RNA gene clusters in *Arabidopsis thaliana*. *Genome Biol.* **18**, 75 (2017).

130. Pontes, O. et al. Natural variation in nuclear dominance reveals the relationship between nucleolar organizer chromatin topology and rRNA gene transcription in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **100**, 11418–11423 (2003).

131. Guo, X. & Han, F. Asymmetric epigenetic modification and elimination of rDNA sequences by polyploidization in wheat. *Plant Cell* **26**, 4311–4327 (2014).

132. Liu, R. & Davis, T. M. Conservation and loss of ribosomal RNA gene sites in diploid and polyploid *Fragaria* (Rosaceae). *BMCI Plant Biol.* **11**, 157 (2011).

133. Steiger, K. A. & Slotte, T. Genomic legacies of the progenitors and the evolutionary consequences of allopolyploidy. *Curr. Opin. Plant Biol.* **30**, 89–93 (2016).

134. Vicient, C. M. & Casacuberta, J. A. Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* **120**, 195–207 (2017).
Extended Data Fig. 1 | Measuring genome sizes of *Arabidopsis* species using flow cytometry. a, FACS sorting of *Solanum lycopersicum* cells from 3-week-old leaf tissue for two replicates. G1 represents the peak denoting the G1 phase of the cell cycle. Cells in the G1 phase have 2 C DNA content (that is a 2 N genome). b, *A. thaliana* 'CVI' accession c, *A. lyrata* 'MN47' (the reference accession) d, *A. suecica* 'ASS3' (the reference accession) e autopolyloid *A. arenosa* accession 'Aa4' f, Bar chart shows calculated genome sizes (rounded to the nearest whole number) for each species using *Solanum lycopersicum* as the standard.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Hi-C and a genetic map analysis for the A. suecica genome. **a,** Hi-C contact map for the genome of *A. suecica.* **b,** Mixing of *A. thaliana* and *A. arenosa* Hi-C reads suggest interchromosomal contacts between homeologous chromosomes is a result of mis-mapping for Hi-C reads. **c,** Accession ‘AS530’ with the region of HE highlighted with an arrow (Fig. 6), no other rearrangements were observed. **d,** Hi-C of synthetic *A. suecica* (third selfed generation). **e** and **f** physical distance (Mb) vs genetic distance (cM) is plotted for the *A. thaliana* and *A. arenosa* subgenome, respectively. Chromosome 2 is not plotted as there are too few SNPs on this chromosome in our cross, due to the recent bottleneck in *A. suecica.*
Extended Data Fig. 3 | Genome composition and analysis of orthologues and the rDNA.  

**a**. Genome composition of the A. suecica subgenomes and the ancestral genomes of A. thaliana and A. lyrata (here a substitute reference for A. arenosa because it is annotated).  

**b**. Counts of orthologous genes between the subgenomes of the reference A. suecica genome and the reference A. thaliana and A. lyrata genome.  

**c**. Copy number of A. thaliana and A. arenosa rDNA in natural A. suecica, ancestral species and synthetic lines. Blue triangles represent the A. thaliana and A. arenosa parent lines of the synthetic A. suecica cross. AT represents results when mapping to the A. thaliana consensus sequence and AA to the A. arenosa consensus sequences for the 45S rRNA.  

**d**. Expression (log$_2$(CPM)) of A. thaliana and A. arenosa rDNA in natural A. suecica, ancestral species and synthetic lines. Accessions with log$_2$(CPM) of $>15$ was taken as evidence for expression for the A. thaliana and A. arenosa 45S rRNA in A. suecica, as this CPM value was above the maximum level of mis-mapping observed in the ancestral species (A. thaliana mapping to the A. arenosa 45S rRNA).
Extended Data Fig. 4 | Population frequency and genomic location of transposon polymorphisms. Shared TE SFS for the a, A. thaliana and b, A. arenosa subgenome. Private TE SFS for the c, A. thaliana and d, A. arenosa subgenome. e, TEs ancestrally from A. arenosa that are present in the A. thaliana subgenome of A. suecica and TEs ancestrally from A. thaliana that are present in the A. arenosa subgenome of A. suecica. f, Shared TEs in the population between A. thaliana and the A. thaliana subgenome of A. suecica. Shared TEs are likely older than private TEs and are enriched around the pericentromeric regions in the A. thaliana subgenome. Private TEs are enriched in the chromosomal arms for both species, where protein-coding gene density is higher (Fig. 1b). g as in f but examining TEs in the population of A. arenosa and the A. arenosa part of A. suecica. Note the region between 5 and 10 on chromosome 2 was not included in the analysis as this region shows synteny with an unplaced contig.
Extended Data Fig. 5 | Transposable element expression analysis. Patterns of TE expression in natural and synthetic A. suecica show that allopolyploidy is not accompanied by an overall upregulation in TE expression as predicted by the ‘genome shock’ hypothesis. a, Heat map of TE expression for the A. thaliana subgenome of A. suecica (dark green) synthetic A. suecica (cyan) and A. thaliana (light green). b, Heat map of TE expression for the A. arenosa subgenome of A. suecica (dark purple) synthetic A. suecica (pink) and A. arenosa (light purple). c and d the breakdown of TE families expressed in each cluster, with helitrons being the most abundant class on the A. thaliana subgenome and TEs of an unknown family being the most abundant in the A. arenosa subgenome.
Extended Data Fig. 6 | Cross-mapping of RNA-seq short reads. a, Box plots of cross-mapping RNA short reads. This was examined by mixing reads in-silico between A. thaliana and A. arenosa. On average, ~6% of A. arenosa reads map to A. thaliana subgenome instead of the A. arenosa subgenome, and ~1% vice versa. Mapping these reads to the combined reference genomes of A. thaliana and A. lyrata (box plot 4 in a) shows that reads map more precisely to the A. suecica reference and that cross-mapping is not due to unreported HE. b, LogFC of log2(CPM) read counts for A. arenosa (CPM of A. arenosa subgenome genes when reads are mapped only to A. arenosa subgenome of A. suecica/CPM of A. arenosa subgenome genes when reads are mapped to the full genome) show only a small effect of mapping strategy to estimate gene expression on the A. arenosa subgenome. c, Pairwise percentage differences ($\pi$) for each group measured for the exons of the 14,041 genes in the expression analysis. High levels of $\pi$ in A. arenosa overlaps with the distribution of $\pi$ between A. thaliana and A. arenosa. This explains why there is more cross-mapping for A. arenosa than for A. thaliana in a Importantly, lower $\pi$ within A. suecica for both subgenomes means that measurements for subgenome dominance are not biased by cross-mapping, as we expect less cross-mapping since the distribution of $\pi$ overlaps less with $\pi$ between A. thaliana and A. arenosa.
Extended Data Fig. 7 | Expression differences between subgenomes in natural and synthetic A. suecica. The distribution of expression differences across homeologous gene pairs in natural and synthetic A. suecica. b, A heat map of expression for genes in the top 5% biased toward the A. arenosa subgenome. The gene must be in the 5% quantile for at least 1 accession. c, The same as in b but for the A. thaliana subgenome. Correlations of log fold change for genes in the tails of the distribution (top 5% quantile) for the A. arenosa subgenome d and the A. thaliana subgenome e.
**Extended Data Fig. 8 | Comparison of genetic and expression distance.**

**a**, PCA plot of biallelic SNPs in the population of *A. thaliana* and *A. suecica* for the *A. thaliana* subgenome of *A. suecica* (N = 345,075 biallelic SNPs), of the analysed 13,647 genes in gene expression in addition to 500 bp up and downstream of each gene sequence.

**b**, Correlation of \( \pi \) (pairwise genetic differences) and expression distance (that is, euclidean distance) for 14,041 genes (* = Bootstrapped 1000 times).

**c**, PCA plot of biallelic SNPs in the population of *A. arenosa* (N.B. we had DNA sequencing for only 3 of the 4 accessions used in the expression analysis) and *A. suecica* for the *A. arenosa* subgenome of *A. suecica* (N = 1,761,708 biallelic SNPs), of the analysed 14,041 genes in gene expression in addition to 500 bp up and downstream of each gene sequence.

**d**, Correlation of Pi (pairwise genetic differences for mapped genomic regions) and expression distance (that is, euclidean distance) for 14,041 genes (* = Bootstrapped 1000 times). *A. arenosa* was too few samples to give reliable correlations and therefore is NA. Grey bars represent the 95% confidence intervals.
Extended Data Fig. 9 | Aneuploidy is frequently observed in synthetic *A. suecica*. a, Comparison of FISH analyses of the reference natural *A. suecica* ‘ASS3’ and synthetic *A. suecica*. Synthetic *A. suecica* shows aneuploidy in both subgenomes in the F₂ generation (gain of one chromosome on the *A. thaliana* subgenome (N=11) and loss of one chromosome on the *A. arenosa* subgenome (N=15)). Natural *A. suecica* shows a stable karyotype. b, DNA-sequencing coverage in the reference natural *A. suecica* accession ‘ASS3’. c and d, DNA-sequencing coverage in siblings of F₁ synthetic *A. suecica* show different cases of aneuploidy (indicated with arrow) in synthetic *A. suecica*, chromosome 4 in c and chromosome 11 in d e overlap of genes involved in cell division from Fig. 5e and genes previously shown to play a role in the adaptation to autopolyploidy in *A. arenosa*¹²¹. The little overlap in genes between *A. suecica* and *A. arenosa* highlights that successful meiosis in polyploids is likely a complex trait.
Extended Data Fig. 10 | See next page for caption.
Extended Data Fig. 10 | Evidence of HE in A. suecica. Reads mapped to the beginning of the HE event in chromosome 6 (~15.9 Mb) in ‘AS530’. Arrows point to the direction of the break. Discordant reads map between the A. arenosa subgenome on chromosome 6 and the read pair maps to the homeologous chromosome 1 on the A. thaliana subgenome (~5 Mb) in b. The end of the HE event in chromosome 6 (~18.4 Mb). Discordant reads map between the A. arenosa subgenome in c and the read pair maps to chromosome 1 (~2.8 Mb) on the A. thaliana subgenome in d, e. Gene counts between the syntenic regions. 431 have a 1:1 relationship, 108 genes are specific to the A. arenosa subgenome and 105 genes are specific to the A. thaliana subgenome. f, Composition of the syntenic regions between the two subgenomes. g, The top 5% quantiles (N=702) for variation in gene expression for the A. thaliana subgenome shows in cluster 7 (N=111) the two outlier accessions (AS150 and ASÖ5) are expressing genes differently to the rest of the population. h, Homeologous genes of this cluster on the A. thaliana subgenome of A. suecica show that these genes are not expressed in these two accessions while i shows they are upregulated in ‘AS150’ and ‘ASÖ5’. j and k 101/111 genes in cluster 7 are located on chromosome 4 in close proximity to each other on the A. thaliana subgenome of the A. suecica reference genome and appear to be deleted in AS5Ö5 and AS150. The A. arenosa subgenome homeologues (located on chromosome 11) have twice the DNA coverage, suggesting they are duplicated, in agreement with expectations of HE event.
Reporting Summary

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

Statistics

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

- □ n/a
- □ Confirmed

☐ □ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ □ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ □ The statistical test(s) used AND whether they are one- or two-sided

☐ □ Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ □ A description of all covariates tested

☐ □ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ □ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ □ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.

☐ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ □ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

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

Software and code

Policy information about availability of computer code

Data collection

No software was used

Data analysis

- FALCON (version 0.3.0), Arrow (snrtrack release 5.0.0.6792), PILON (version 1.22), HICUP (version 0.6.1), MUMmer (version 3.23), LACHESIS (version 1.0.0), BWA-MEM (version 0.7.15), samtools (version 0.1.19), Augustus, GenomeThreader (version 1.7.0), RepeatModeler (version 1.0.11), RepeatMasker (version 4.0.7), STAR (version 2.7.5a), HISAT2 (version 2.1), EAGLE, POPsodateurTE2 (version v1.10.04), BLAST, R package HMM, R package TopGO, R package EdgeR, R package biomaRt, R package flowCore

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

Data

Policy information about availability of data

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

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

Genome assemblies and raw short reads can be found in the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/home). The genome assembly for A. sucorea ASS3 can be found under the BioProject number PRJEB42198. Assembly accession GCA_900175345. The raw reads for the A. sucorea genome assembly generated by Pacbio RSII can be found under ERR5037702 and those from Seqel under ERR5031296. The HiC reads used for scaffolding the A. sucorea assembly can be found under ERR5032369.
The contig assembly for tetraploid A. arenosa (ssp. arenosa) can be found under the BioProject number PRJEB42276, assembly accession GCA_905175405. The raw reads for the A. arenosa A4x contig assembly generated by Sequel can be found under ERR5031542 and the reads generated by Nanopore under ERR5031541. HIC reads for the A. arenosa assembly can be found under ERR5032370.

DNA sequencing data for the ancestral species, the outlier accession ASG30 and synthetic A. suetica can be found under the BioProject PRJEB42290.

The RNA-seq reads are under the BioProject number PRJEB42277.

TE presence/absence calls for A. suetica and the ancestral species can be found in Supplementary Data 1.

A list of DEGs, orthologs, enriched DAP-seq transcription factors, CyMIRA gene overlaps and RNA-seq mapping statistics can be found in Supplementary Data 2.

Log fold change and CPM (counts per million) for genes on the A. thaliana and A. arenosa subgenome can be found in Supplementary Data 3.

The gene annotation (gff3 file) of the A. suetica genome can be found in Supplementary Data 4.

TE consensus sequences and a hierarchy file of TE order for A. suetica can be found in Supplementary Data 5.

Field-specific reporting

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

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

| Sample size | No sample size calculation was performed |
| Data exclusions | No data was excluded from the analysis |
| Replication | Attempts at replication were successful |
| Randomization | Samples were allocated to species groups |
| Blinding | Not relevant to study |

Reporting for specific materials, systems and methods

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

### Materials & experimental systems

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

### Methods

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