Gradual evolution of allopolyploidy in *Arabidopsis suecica*

Robin Burns[1], Terezie Mandáková[2], Joanna Jagoda[1], Luz Mayela Soto-Jiménez[1], Chang Liu[3], Martin A. Lysak[2], Polina Yu. Novikova[4,5*] and Magnus Nordborg[1*]

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna BioCenter, Vienna, Austria.
[2]CEITEC - Central European Institute of Technology, and Faculty of Science, Masaryk University, Brno, Czech Republic.
[3]Center for Plant Molecular Biology (ZMBP), University of Tübingen, Auf der Morgenstelle 32, 72076 Tübingen, Germany.
[4]VIB-UGent Center for Plant Systems Biology, Ghent, Belgium.
[5]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany.

***Corresponding authors: pnovikova@mpipz.mpg.de, magnus.nordborg@gmi.oeaw.ac.at**

Abstract

The majority of diploid organisms have polyploid ancestors. The evolutionary process of polyploidization (and subsequent re-diploidization) is poorly understood, but has frequently been conjectured to involve some form of "genome shock" — partly inspired by studies in crops, many of which are polyploid, and in which polyploidy has frequently been linked to dramatic genomic changes such as subgenome expression dominance and genome reorganization. It is unclear, however, whether domesticated polyploids are representative of natural ones. Here, we study polyploidization in *Arabidopsis suecica* (n = 13), a post-glacial allopolyploid species formed via hybridization of *A. thaliana* (n = 5) and *A. arenosa* (n = 8). We generated a chromosome-level genome assembly of *A. suecica* and complemented it with polymorphism and transcriptome data from multiple individuals of all species. Despite a divergence of ~6 Mya between the two ancestral species and appreciable differences in their genome composition, we see no evidence of a genome shock: the *A. suecica* genome is highly colinear with the ancestral genomes, there is no subgenome dominance in expression, and transposable element dynamics appear to be stable. We do, however, find strong evidence for changes suggesting gradual adaptation to polyploidy. In particular, the *A. thaliana* subgenome shows upregulation of meiosis-related genes, possibly in order to prevent aneuploidy and undesirable homeologous exchanges that are frequently observed in experimentally generated *A. suecica*, and the *A. arenosa* subgenome shows upregulation of cyto-nuclear related processes, possibly in response to the new cytoplasmic environment of *A. suecica,* with plastids maternally inherited from *A. thaliana*.

Introduction

Ancient polyploidization or whole-genome duplication is a hallmark of most higher-organism genomes[1,2], including our own[3,4]. While most of these organisms are now diploid and show only traces of polyploidy, there are many examples of recent polyploidization, in particular among flowering plants[5–9]. These examples are important because they allow us to study the process of polyploidization, rather than just inferring that it happened and trying to understand its evolutionary importance.

The existence of wide-spread recent polyploids, like *Capsella bursa-pastoris* (Shepherd's Purse)[10–12], *Trifolium repens* (white clover)[13], and *Brachypodium hybridum*[14,15], demonstrates that polyploid species can quickly become very successful. That said, new polyploid species face numerous challenges, ranging from those on a population level, such as bottlenecks[13,16] and competition with their diploid progenitors[17], to those on a cellular level, such as chromosome segregation[18–20] and changes to genome structure and regulation[21] — the so-called "genome shock"[22]. However, genomic changes resulting from the hybridization of two (or more) diverged genomes have primarily been studied in crops, such as cotton[5], strawberry[6], wheat[7] and peanut[8,9], which confounds polyploidization with domestication.

Here we focus on a young, natural allopolyploid species, *Arabidopsis suecica* (2n = 4x = 26), formed through the hybridization of *A. thaliana* (2n = 10) and *A. arenosa* (2n = 2x/4x = 16/32), circa 16 kya, during the Last Glacial Maximum[16] and now widely established in northern Fennoscandia (Fig. 1a). The ancestral species diverged around 6 Mya[23], and, based on mitochondrial and chloroplast sequences, it is clear that *A. thaliana* is the maternal and *A. arenosa* the paternal parent of the hybrid[24], a scenario also supported by the fact that *A. arenosa* itself is a ploidy-variable species, so that *A. suecica* could readily be generated through the fertilization of an unreduced egg cell (2n = 2x) from *A. thaliana* by a sperm cell (n = 2x) from autotetraploid *A. arenosa*[16,25]. We have previously shown that, although *A. suecica* shows clear evidence of a genetic bottleneck[16], it shares most of its variation with the ancestral species, demonstrating that the species was formed through a hybridization and polyploidization process that involved many crosses and individuals.In order to study genomic change in *A. suecica*, we used long-read sequencing to generate a high-quality, chromosome-level assembly of a single individual, taking advantage of the fact that *A. suecica*, like *A. thaliana*, is highly selfing, making it possible to sequence naturally inbred individuals. The genome sequence was complemented by a partial assembly of a tetraploid outcrosser *A. arenosa*, and by short-read genome and transcriptome sequencing data from many individuals of all three species — including "synthetic" *A. suecica* generated *de novo* in laboratory crosses.

Results and discussion

1. The genome is conserved

We assembled a reference genome from a naturally inbred *A. suecica* accession ("ASS3"), using 50x long-read PacBio sequencing (PacBio RS II). The absence of heterozygosity and the substantial (~11,6%) divergence between the subgenomes greatly facilitated the assembly. In contrast, assembling even a diploid *A. arenosa* genome is complicated by high heterozygosity (nucleotide diversity around 3.5%[26]) coupled with a relatively high level of repetitive sequences (compared to the gene-rich *A. thaliana* genome). Our attempt to assemble a tetraploid *A. arenosa*, which is also included here, led to a very fragmented assembly of 3,629 contigs with an N50 of 331 Kb. In contrast, the *A. suecica* assembly has an N50 contig size of 9.02 Mb. The assembled contigs totaled 276 Mb (~90% of the 305 Mb genome size estimated by flow cytometry — see Supplementary Fig. 1; ~88% of the 312Mb genome size estimated by kmer analysis). Contigs were placed into scaffolds using high-coverage chromosome conformation capture (HiC) data and by using the reference genomes of *A. thaliana* and *A. lyrata* (here the closest substitute for *A. arenosa*) as guides. This resulted in 13 chromosome-scale scaffolds (Supplementary Fig. 2a). The placement and orientation of each contig within a scaffold was confirmed and corrected using a genetic map for *A. suecica* (see Methods, Supplementary Fig. 3, Supplementary Fig. 4). The resulting chromosome-level assembly (Fig. 1b) contains 258 Mb, and has an N50 scaffold size of 19.59 Mb. The five chromosomes of the *A. thaliana* subgenome and the eight chromosomes of the *A. arenosa* subgenome sum to 119 Mb and 139 Mb, respectively.

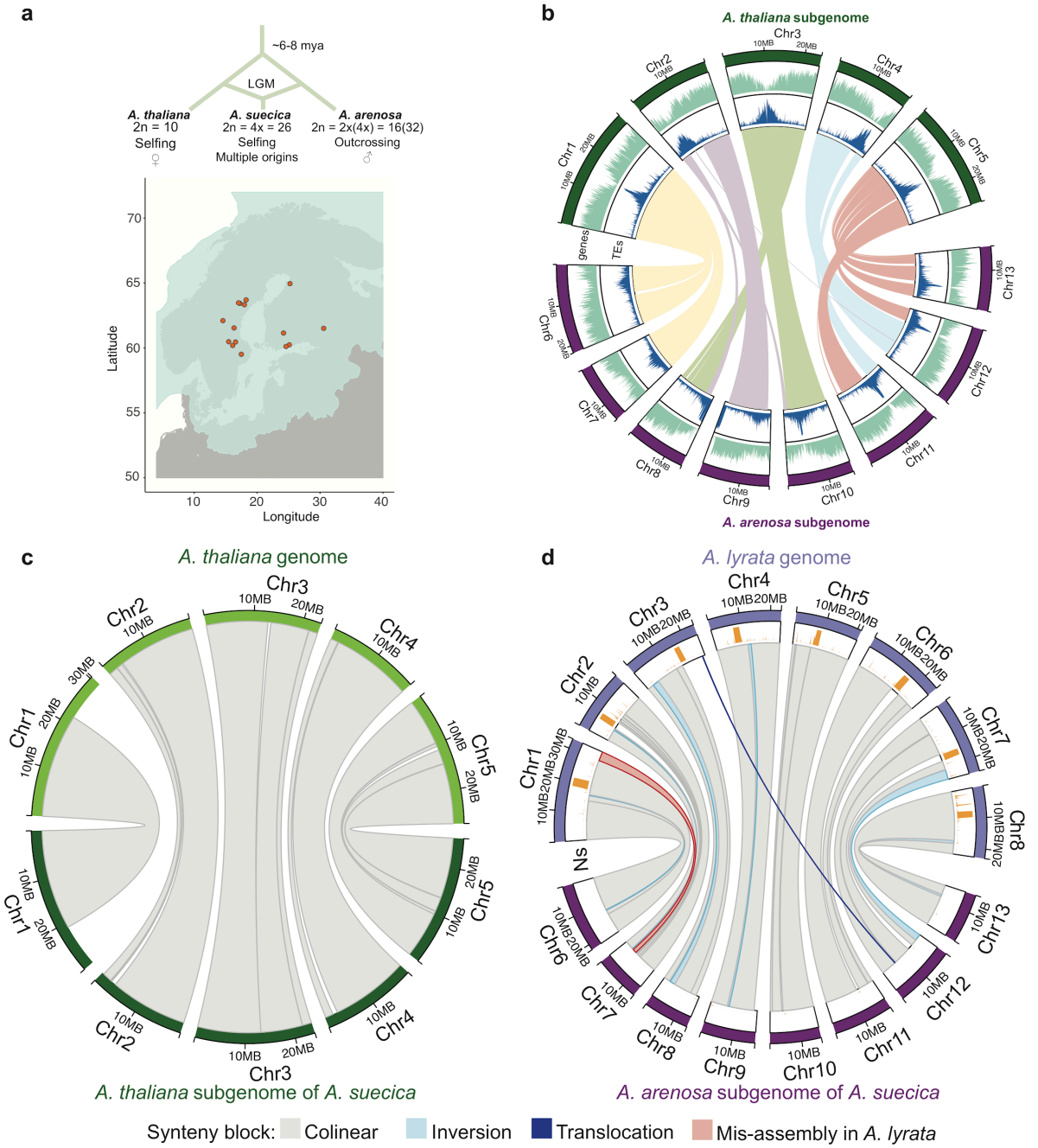


Figure 1. The genome of *A. suecica* is largely colinear with the ancestral genomes. **a** Schematic depicting the origin of *A. suecica* and its current distribution in the relation to the last glacial maximum (LGM). **b** The chromosome-level assembly of the *A. suecica* genome with inner links depicting syntenic blocks between the *A. thaliana* and *A. arenosa* subgenomes of *A. suecica*. The blue histogram represents the distribution of TEs along the genome and the green histogram corresponds to the distribution of protein-coding genes. **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 37 synteny blocks were found, 28 of which were colinear. Of the remaining 9 blocks, 7 represent inversions in the *A. arenosa* subgenome of *A. suecica* compared to *A. lyrata*, 1 is a translocation, and one corresponds to a previously reported mis-assembly in the *A. lyrata* genome[27]. Orange bars represent a density plot of missing regions ("N" bases) in the *A. lyrata* genome.

Approximately 108 and 131 Mb of the *A. thaliana* and *A. arenosa* subgenomes of *A. suecica* are in large syntenic blocks to the genomes of the ancestral species: 13 and 37 blocks, respectively (Fig. 1c,d). The vast majority of these syntenic blocks are themselves also colinear, with the exception of five small-scale inversions (~3.7 Mb) and one translocation (~1Mb) on the *A. arenosa* subgenome— which may well (indeed probably do) reflect differences between *A. lyrata* and *A. arenosa*, two species separated by about a million years[23,26]. We also corrected for the described[27] mis-assembly in the *A. lyrata* reference genome using our genetic map. Overall we find that approximately 92% of the *A. suecica* genome is syntenic to the ancestral genomes, the 13 chromosomes of *A. suecica* having remained almost completely colinear (Fig. 1c,d). This highlights the conservation of the *A. suecica* genome and contrasts with the major rearrangements that have been observed in several polyploid crops[8,9,28–30] and which have often been associated with a "genome shock". Interestingly, major rearrangements have also been observed in synthetic *A. suecica*[31], and we see clear evidence of aneuploidy in ours — a topic to which we shall return below.

A total of 45,048 protein-coding genes were annotated for the *A. suecica* reference, of which 22,383 and 22,665 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 2036 (96%) and 2007 (94.6%) complete genes for the *A. thaliana* and *A. arenosa* subgenomes, respectively (Supplementary Fig. 5c,d). Of the protein-coding genes, 17,814 had a one-to-one orthology between the subgenomes of *A. suecica* and 16,781 genes were conserved in a 1:1:1:1 relationship for both subgenomes of *A. suecica* and the ancestral species (using *A. lyrata* as a substitute for *A. arenosa*) (Supplementary Data 2, Supplementary Fig. 5b). We functionally annotated lineage-specific genes in *A. suecica* using InterPro, and only found significant enrichment in *A. thaliana* subgenome of *A. suecica* for two GO terms (GO:0008234 and GO:0015074), both of which are associated with repeat content (Supplementary Data 2). Ancestral genes not found in the *A. suecica* genome annotation were overrepresented for functional categories of plant defense response. However, checking coverage for these genes by mapping the raw *A. suecica* whole-genome resequencing data to the ancestral genomes did not confirm their loss, suggesting rather misassembly or misannotation, which is expected due to the repetitive and highly polymorphic nature of R-genes in plants.

2. The rDNA clusters are highly variable

In eukaryotic genomes, genes encoding ribosomal RNA (rRNA) occur as tandem arrays in rDNA clusters. The 45S rDNA clusters are particularly large, containing hundreds or thousands of copies, spanning millions of base pairs[32]. The nucleolus, the site of pre-ribosome assembly, forms at these clusters, but only if they are actively transcribed, and it was observed long ago that only one parent's rDNA tended to be involved in nucleolus formation in inter-specific hybrids, a phenomenon known as "nucleolar dominance"[33,34]. 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[35].

Given this, we examined the composition of 45S rDNA repeats as well as their transcription. While the large and highly repetitive 45S rDNA clusters are not part of the genome assembly, it is possible to measure the copy number of *A. thaliana* and *A. arenosa* 45S rRNA genes using sequencing coverage (see Methods), and we find three accessions to have experienced massive loss of the *A. thaliana* rDNA loci (Fig. 2a), which we confirmed for

one of the accessions ("AS90a") by FISH analysis (Fig. 2d,e). However, there is massive copy number variation for 45S rRNA genes in *A. suecica* (Fig. 2a), and some accessions (e.g., the reference accession "ASS3") have higher *A. thaliana* than *A. arenosa* 45S rRNA copy number (Fig. 2b,c).

Turning to expression, we also find nucleolar dominance to be segregating in *A. suecica* (see Methods and Supplementary Fig. 6), with the majority of accessions expressing both 45S rRNA alleles, five exclusively expressing *A. arenosa* 45S rRNA, and one exclusively expressing *A. thaliana* 45S rRNA (Fig. 2a).

This extensive variation in 45S cluster size and expression is reminiscent of the intraspecific variation seen in *A. thaliana* (where different accessions express either the chromosome 2 or chromosome 4 rDNA cluster, or both[36,37]), and suggests that the phenomenon of nucleolar dominance can at least partly be explained by retained ancestral variation. However, the large-scale decrease in rDNA cluster size observed in some accessions may be a direct consequence of allopolyploidization itself, as synthetic *A. suecica* sometimes shows immediate loss of 45S rDNA (even as early as the F1 stage) and this too varies between siblings and generations (Supplementary Fig. 6a).

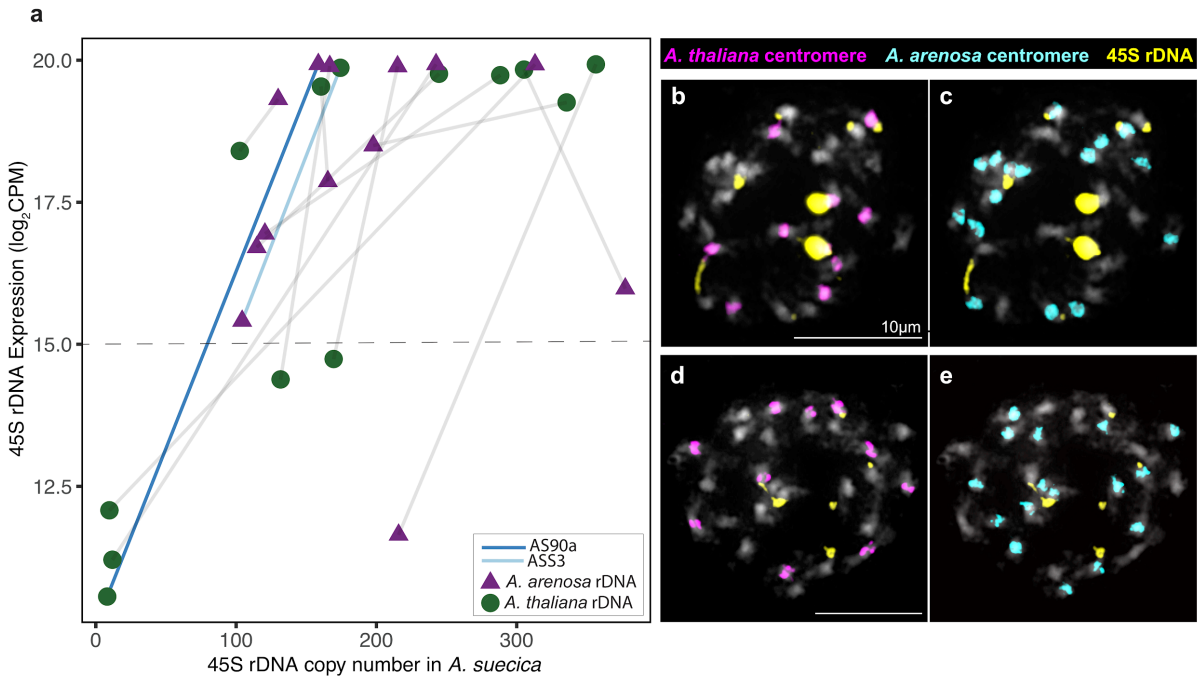


Figure 2. Expression and copy number variation of 45S rDNA in *A. suecica*. a The relationship between expression levels ($\log_2$ CPM) and copy number of 45S rRNA shows extensive variation of 45S rDNA copy number and varying direction of "nucleolar dominance". Grey lines connect subgenomes of the same accession. Values above the dashed line are taken as evidence for the expression of a particular 45S rDNA allele, as this is above the maximum level of mis-mapping seen in the ancestral species here used as a control (see Supplementary Figure 6b). **b** and **c** FISH results of a natural *A. suecica* 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). **d** and **e** FISH result of a natural 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).

188 # 3. No evidence for abnormal transposon activity

189 The possibility that hybridization and polyploidization leads to a "genome shock" in the form of
190 increased transposon activity has been much discussed [21,22,38,39]. Certainly the two
191 subgenomes of *A. suecica* differ massively in transposon content: there are almost twice as
192 many annotated transposons in the *A. arenosa* as in the *A. thaliana* subgenome (64,409 vs
193 33,420; see Supplementary Figs. 5a and 7), and the true difference is almost certainly greater
194 given that the *A. arenosa* subgenome assembly is less complete (and many of the missing
195 regions are likely to be repeat-rich) and that the transposon annotation is biased towards *A.*
196 *thaliana.* Has the combination of two such different genomes lead to increased transposon
197 activity?

198 Our assembled *A. thaliana* subgenome does contain roughly 3,000 more annotated
199 transposons than the TAIR10 *A. thaliana* reference genome, but this could reflect greater
200 transposon number in the *A. thaliana* ancestors of this genome rather than increased
201 transposon activity in *A. suecica.* In order to gain insight into transposon activity in *A. suecica*,
202 we need to identify jumps that occurred after the species separated (and are thus only found
203 in this species). We used the software PopoolationTE2[40] to call presence-absence variation
204 on a population-scale level using genome re-sequencing datasets for 15 natural *A. suecica*
205 accessions, 18 *A. thaliana* accessions genetically close to *A. suecica*, and 9 *A. arenosa* lines.
206 Of the 25,677 insertion polymorphisms called with respect to the *A. thaliana* subgenome,
207 9,354 were shared between *A. thaliana* and *A. suecica*, 8,068 were only found in *A. thaliana*,
208 and 8,255 were only found in *A. suecica*. Of the 100,849 insertions on the *A. arenosa*
209 subgenome of *A. suecica,* 11,024 were shared with *A. arenosa*, 69,041 were private to *A.*
210 *arenosa*, and 20,784 were private to *A. suecica* (Supplementary Data 1a,b; Supplementary
211 Figs. 8,9). Considering the number of transposons per individual genome (Fig 4a), we see that
212 most transposon insertions in a typical *A. thaliana* subgenome are also found in *A. thaliana*,
213 and that the slightly higher transposon load in the *A. thaliana* subgenome is mainly due to
214 these. The reason for this is likely a population bottleneck, as we shall see below. In contrast,
215 the number of recent insertions (that are private to the species) is not higher in the *A. thaliana*
216 subgenome, suggesting that transposon activity in this subgenome is not increased.

217 Turning to the *A. arenosa* subgenome, we see that a typical *A. suecica* contains only about
218 half the number of transposons of a typical *A. arenosa* individual (Fig 4a). However, the latter
219 is an outcrossing tetraploid, and it is thus fairer to compare with the number of transposons in
220 four randomly chosen *A. arenosa* subgenomes. This largely accounts for the observed
221 difference, but there are still clearly fewer transposons in *A. suecica*. A population bottleneck
222 (see below) likely explains much of this, but it is impossible to rule out a contribution of
223 decreased transposon activity in *A. suecica* as well.

224 To sum up, we see no evidence for a burst of transposon activity accompanying
225 polyploidization in *A. suecica*, a conclusion also supported by a lack of increase in transposon
226 expression for both synthetic and natural *A. suecica* compared to the *A. thaliana* and *A.*
227 *arenosa* on both subgenomes (Supplementary Fig. 9). We do see clear traces of the
228 population bottleneck accompanying the origin of *A. suecica*, however. The frequency
229 distribution of polymorphic transposon insertions private in *A. suecica* is skewed towards zero
230 — almost certainly because of purifying selection because the distribution is more similar to
231 that of non-synonymous SNPs than to that of synonymous SNPs (Fig 4b and c). However,
232 for both subgenomes, *A. suecica* also contains a large number of fixed or nearly-fixed
233 insertions that are present in the ancestral species at lower frequency (Fig 4d and e). These
234 are likely to have reached high-frequency as a result of a bottleneck. Shared transposons are

enriched in the pericentromeric regions of the genome depleted of protein-coding genes, while private transposons insertions, which are generally at low frequency, show a more uniform distribution across the genome, consistent with evidence for stronger selection against transposon insertion in the relatively gene-dense chromosome arms[41,42] (Supplementary Fig. 10).

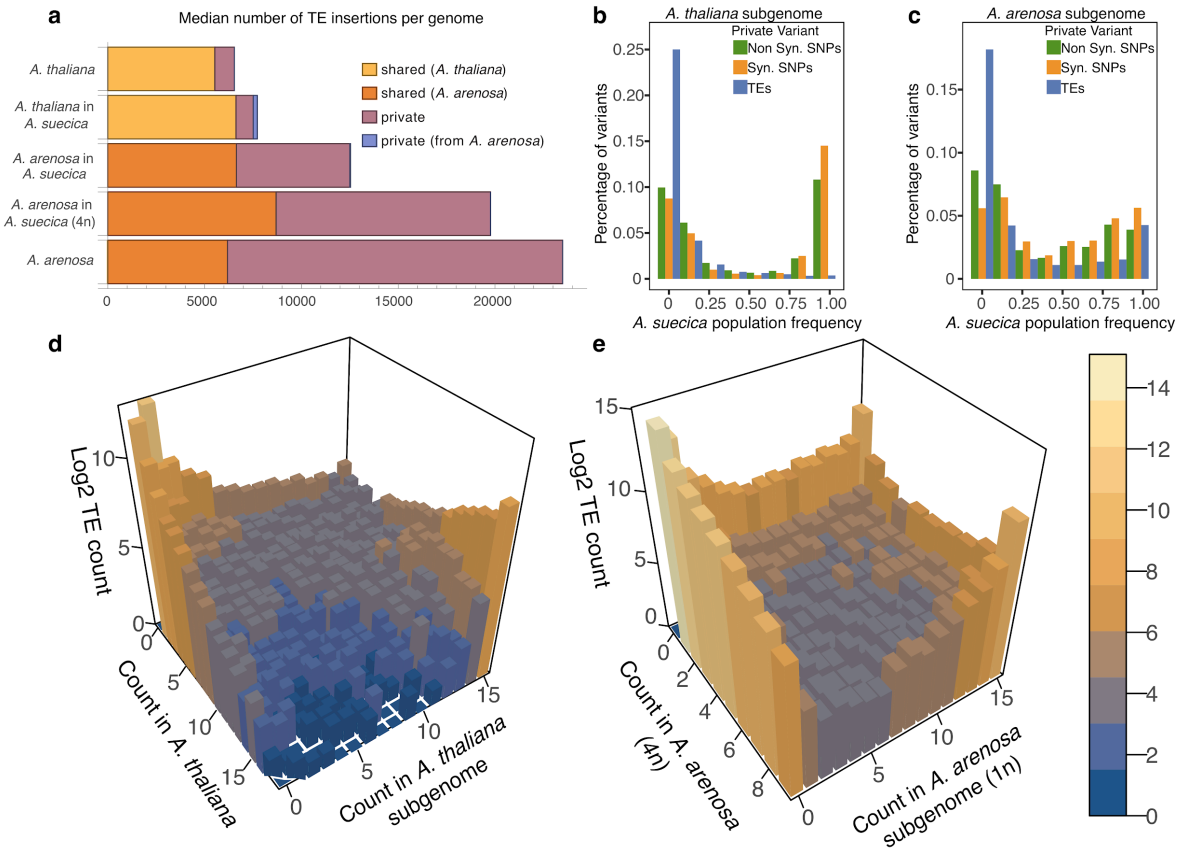


Figure 3. TE dynamics in *A. suecica* reveal no evidence for abnormal transposon activity. a Median TE insertions per genome. As the *A. arenosa* population is an autotetraploid outcrosser, 4 randomly chosen haploid *A. arenosa* subgenomes of *A. suecica* were combined to make a 4n *A. suecica*. *A. suecica* does not show an increase in private TE insertions compared with the ancestral species for both subgenomes, and shared TEs constitute a higher fraction of TEs in *A. suecica* reflecting the strong population bottleneck at its origin. Site-frequency spectra of non-synonymous SNPs, synonymous SNPs and TEs in the **b** *A. thaliana* and **c** *A. arenosa* subgenomes of *A. suecica* suggest that TEs are under purifying selection on both subgenomes. **d** 3D histogram of a joint TE frequency spectrum for *A. thaliana* on the x-axis and the *A. thaliana* subgenome of *A. suecica* on the y-axis **e** 3D histogram of a joint TE frequency spectrum for *A. arenosa* on the x-axis and the *A. arenosa* subgenome of *A. suecica* on the y-axis. **d** and **e** show stable dynamics of private TEs in *A. suecica* and a bottleneck effect on the ancestral TEs (shared) at the origin of the *A. suecica* species.

An interesting subset of recent transposon insertions private to *A. suecica* are those that have jumped between the two subgenomes. We searched for full-length transposon copies that are present in both subgenomes of *A. suecica* and then assigned the resulting consensus sequences to either the *A. thaliana* or the *A. arenosa* ancestral genomes using BLAST (see Methods). We were able to assign 15 and 55 consensus sequences as being specific to the *A. thaliana* and *A. arenosa* ancestral genomes, respectively. Using these sequences, we searched our transposon polymorphism data for corresponding polymorphisms, and identified 1,823 *A. arenosa* transposon polymorphisms on the *A. thaliana* subgenome, and 461 *A.*

262 *thaliana* transposon polymorphisms on the *A. arenosa* subgenome. Like other private
263 polymorphisms, these are skewed towards rare frequencies, and are uniformly distributed
264 across the (sub-)genome. Most of the transposons that have jumped into the *A. thaliana*
265 subgenome are helitrons and LTR elements (Supplementary Figure 12). LTR (copia) elements
266 also make up most of the *A. thaliana* transposons segregating in the *A. arenosa* subgenome,
267 with the most abundant element having a closest match in the *A. thaliana* mitochondrial
268 genome. The fact that nearly four times as many new insertions appear to have resulted from
269 jumps from *A. arenosa* to *A. thaliana* than the other way around is notable. It could reflect
270 higher transposon activity in the *A. arenosa* subgenome, but can mostly be explained by
271 differences in genome size and transposon number. If there were no differences in activity,
272 we would expect the number of cross-subgenome jumps to be proportional to the number of
273 potential source elements and the size of the target genome. As we have seen, the *A. arenosa*
274 subgenome contains roughly twice as many transposons as the *A. thaliana* subgenome, and
275 is about 17% larger. These factors alone would account for a 3.4-fold difference in cross-
276 subgenome jumping.
277 In conclusion, transposon activity in *A. suecica* appears to be governed largely by the
278 same processes that governed it in the ancestral species.

279 4. No bias in expression between subgenomes

280 Over time the traces of polyploidy are erased through an evolutionary process involving gene
281 loss, often referred to as fractionation or re-diploidization[43,4445–47]. Analyses of retained
282 homeologs in ancient allopolyploids such as *A. thaliana*[48], *Zea mays*[49], *Brassica rapa*[50] and
283 *Gossypium raimondii*[51] have revealed that one "dominant" subgenome remains more intact,
284 with more highly expressed homeologs compared to the "submissive" genome(s)[48]. This
285 pattern of "biased fractionation" has not been observed in ancient autopolyploids[52], such as
286 pear[53], and is believed to be allopolyploid-specific. Subgenome dominance in expression has
287 been reported for a number of more recent allopolyploids such as strawberry[6] and both natural
288 and synthetic allopolyploids such as *Brassica napus*[54], and monkeyflower[55]. However, some
289 allopolyploids display even subgenome expression, such as *C. bursa-pastoris*[10,12] and *B.*
290 *hybridum*[14].
291 Subgenome dominance is often linked to differences in transposon content[6] and/or large
292 genetic differences between subgenomes[56]. This makes *A. suecica*, with 6 Mya divergence
293 between the gene-dense *A. thaliana* and the transposon-rich *A. arenosa*, a promising
294 candidate to study this phenomenon at unprecedented resolution. Previous reports on
295 subgenome dominance in *A. suecica* are conflicting, suggesting a bias to either the *A.*
296 *thaliana*[57] or the *A. arenosa*[58] subgenome.
297 To investigate the evolution of gene expression in *A. suecica*, we generated RNA-seq data
298 for 15 natural *A. suecica* accessions, 15 closely related *A. thaliana* accessions, 4 *A. arenosa*
299 individuals, a synthetically generated *A. suecica* from a lab cross (the 2nd and 3rd hybrid
300 generations) and the parental lines of this cross. Each sample had 2-3 biological replicates
301 (Supplementary Data 2). On average, we obtained 10.6 million raw reads per replicate, of
302 which 7.6 million reads were uniquely mapped to the *A. suecica* reference genome and 13,647
303 homeologous gene pairs (see Methods, Supplementary Fig. 13).
304 Considering the difference in expression between homeologous genes, we found no bias
305 towards one or the other subgenome of *A. suecica*, for any sample or tissue, including
306 synthetic *A. suecica* (Fig. 4a and Supplementary Fig. 14a). This strongly 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. thaliana*[57] or the *A. arenosa*[58] 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 appears not to be biased towards any particular gene ontology (GO) category, and is furthermore not consistent between accessions and individuals (Fig. 4b, Supplementary Fig. 14b,c). This suggests that many large subgenome expression differences are due to genetic polymorphisms within *A. suecica* rather than fixed differences relative to the ancestral species.

Levels of expression dominance were reported to vary across tissues in *C. bursa pastoris*[11]. To test whether expression dominance can vary for tissue-specific genes, we examined homeologous gene-pairs where at least one gene in the gene pair showed tissue specific expression, in whole-rosettes and floral buds. We do not find evidence for dominance between subgenomes in tissue specific expression either (Figure 4b). Interestingly, the 402 genes with significant expression in whole rosettes for both homeologs 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. 1009 genes with floral bud specific expression for both homeologous gene pairs were overrepresented for GO terms related to responses to chemical stimuli, such as auxin and jasmonic acid, which may reflect early developmental changes in this young tissue (Supplementary Table 1).

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.

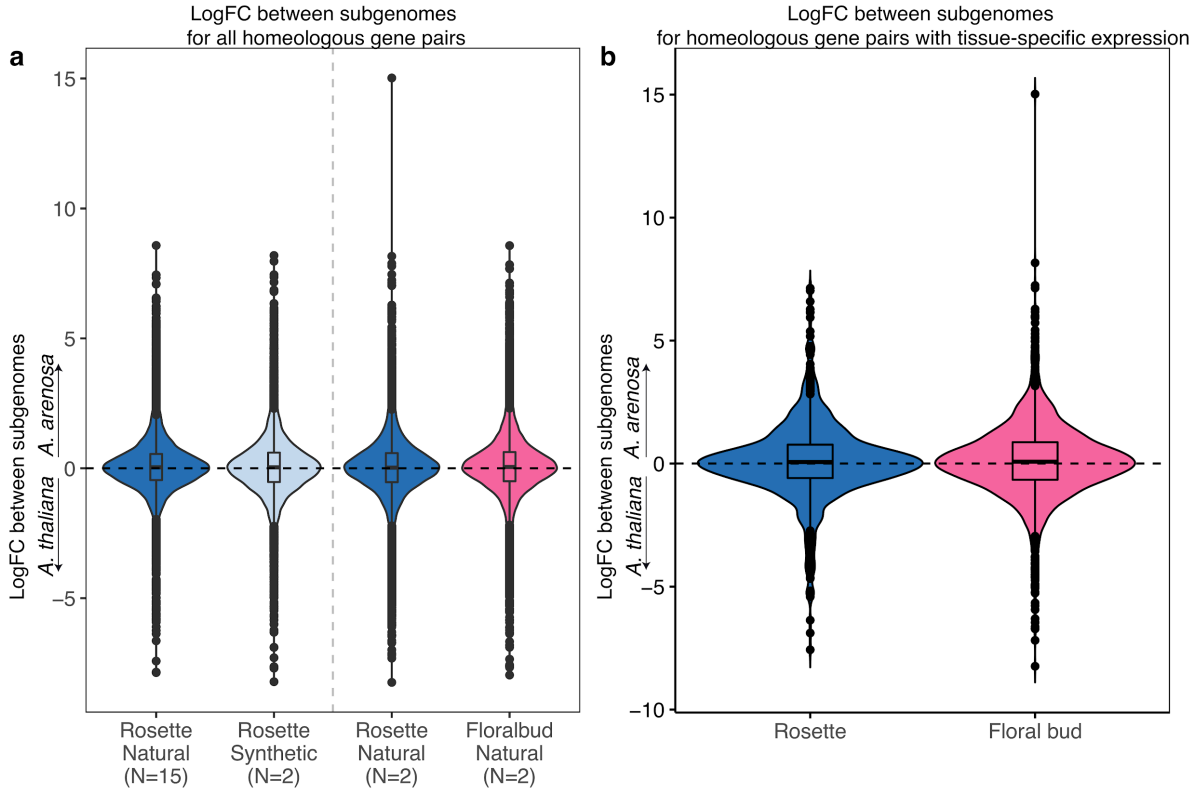


Figure 4. Patterns of gene expression between the subgenomes of *A. suecica* in rosettes and floral buds. **a** Violin plots of the mean log fold-change between the subgenomes for the 15 natural *A. suecica* accessions and two synthetic lines for whole rosettes. Mean log fold-change for the two accessions ("ASS3" and "AS530") where transcriptome data for both whole rosettes and flower buds were available. All the distributions are centered around zero suggesting even subgenome expression. **b** Violin plots for the mean log fold-change between the subgenomes for genes with tissue-specific expression. At least one gene in a homeologous gene pair was required to show tissue-specific expression.

5. Evolving gene expression in *A. suecica*

The previous section focused on differences in expression between the subgenomes, between homeologous copies of the same gene within the same individual. This section will focus on differences between individuals, between homologous copies of genes that are part of the same (sub-)genome. 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), suggesting that hybridization does not automatically result in large-scale transcriptional changes, and that altered gene expression changes in natural *A. suecica* have evolved over time. Given the limited time involved, the lack of correlation between expression and sequence divergence within and between species (Supplementary Fig. 15), and the fact the genes that have changed expression are far from random with respect to function (see below), we suggest that the first principal component primarily captures trans-regulated expression changes in *A. suecica* that are likely adaptive.

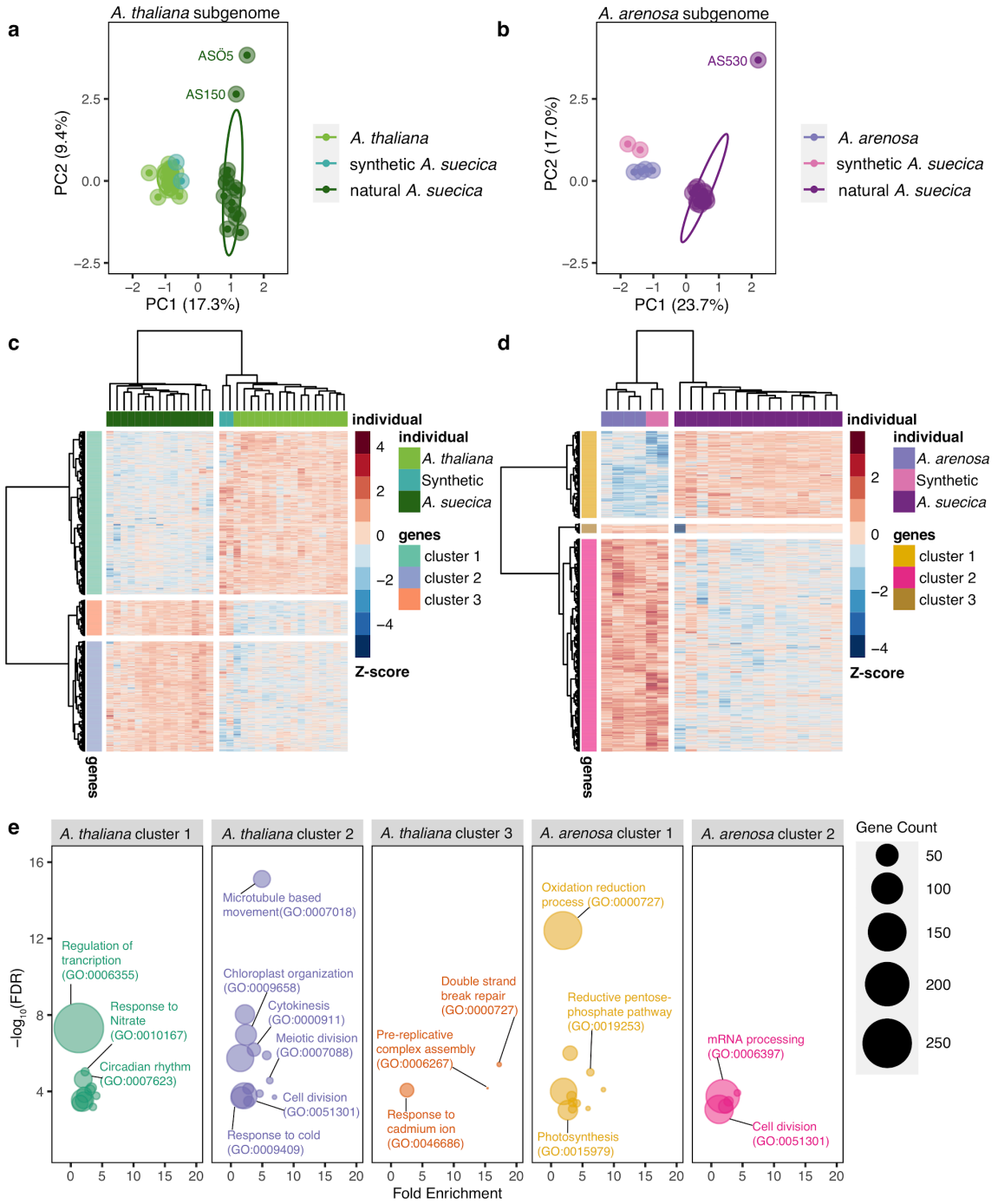


Figure 5. Differential gene expression analysis in *A. suecica*. Patterns of differential gene expression in *A. suecica* support adaptation to the whole-genome duplication for the *A. thaliana* subgenome and adaptation to the new plastid environment for the *A. arenosa* subgenome. **a** PCA for *A. thaliana*, *A. thaliana* subgenome of natural and synthetic *A. suecica* lines. PC1 separates natural *A. suecica* from the ancestral species and the synthetic lines. **b** PCA for *A. arenosa*, *A. arenosa* subgenome of natural and synthetic *A. suecica* 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). **c, d** Heatmap of DEGs for the two subgenomes of *A. suecica*. Positive numbers (red color) indicate higher expression. Genes and individuals have been clustered based on similarity in expression, resulting in clusters discussed in text. **e** Gene ontology enrichment for each cluster in **c** and **d**. Categories discussed in the text are highlighted.

371 To further characterize expression changes in natural *A. suecica* we analyzed differentially
372 expressed genes (DEGs) on both subgenomes compared to the corresponding ancestral
373 species. The total number of DEGs was 3,854 and 4,136 genes for the *A. thaliana* and *A.*
374 *arenosa* subgenomes, respectively (see Methods). These genes were clustered based on the
375 pattern of change across individuals (Fig. 5c,d) and GO enrichment analysis was carried out
376 for each cluster (Fig. 5e, Supplementary Table 2).
377 For the *A. thaliana* subgenome, we identified three clusters. Cluster 1 comprised 2,041
378 genes that showed decreased expression in *A. suecica* compared to *A. thaliana*. These genes
379 are strongly enriched for transcriptional regulation, which may be expected as we are
380 examining DEGs between the species. Also notable are enrichments for circadian rhythm
381 function, and response to nitrate, both of which may be related to the ecology of *A. suecica*
382 and its post-glacial migration to the Fennoscandinavia region (Fig. 1a).
383 Cluster 2 consisted of 1,376 genes that show increased expression in *A. suecica*
384 compared to *A. thaliana*, and several of the enriched GO categories, such as microtubule-
385 based movement, cytokinesis, meiosis and cell division, suggest that the *A. thaliana*
386 subgenome of *A. suecica* is adapting to polyploidy at the level of basic cell biology. That there
387 has been strong selection for this seems likely given that aneuploidy is frequent in synthetic
388 *A. suecica* (Supplementary Fig. 16), while natural *A. suecica* has a stable and conserved
389 karyotype. Importantly, there is independent evidence for adaptation to polyploidy via
390 modifications of the meiotic machinery in the other ancestor of *A. suecica*, *A. arenosa*, as
391 well[19,59,60], although we see very little overlap in the genes involved (Supplementary Fig. 16).
392 The nature of these changes in the *A. suecica A. thaliana* subgenome will require further
393 investigation, but we note that there is enrichment (see Methods, Supplementary Data 3) for
394 Myb family transcription factor binding sites[61] among upregulated genes in cluster 2. There is
395 also enrichment for "response to cold" genes in cluster 2, a category which was also over-
396 expressed in rosettes relative to floral buds (Supplementary Table 1), and which could again
397 be related to the post-glacial migration of *A. suecica*.
398 Cluster 3, finally, consists of 437 genes that are over-expressed in both natural and
399 synthetic *A. suecica* relative to *A. thaliana*. These expression changes are thus likely an
400 immediate consequence of hybridization and must reflect trans-regulation. Genes in this
401 cluster are enriched for "response to cadmium", "double-strand break repair" and "pre-
402 replicative complex assembly". *A. arenosa* is known to be somewhat metal-tolerant[62], and it is
403 possible that the *A. arenosa* subgenome simply upregulates related genes on the *A. thaliana*
404 subgenome. Upregulation of double-strand break repair and pre-replicative complex assembly
405 could be an immediate response to genome-doubling, given the degree of aneuploidy we
406 detect in the synthetic *A. suecica* lines (Supplementary Fig. 16). Notably, the synthetic lines
407 used in the expression analysis were selected to be healthy-looking, and did not show signs
408 of aneuploidy (Supplementary Fig. 17).
409 For the *A. arenosa* subgenome, we also find three clusters of DEGs (Fig. 5d) with GO
410 enrichment for two (Fig. 5e, Supplementary Table 2). Cluster 1 consists of 1,165 genes that
411 show increased expression in natural *A. suecica* compared to *A. arenosa* and synthetic *A.*
412 *suecica,* and are enriched for plastid-related functions, including oxidation-reduction, the
413 reductive pentose-phosphate pathway and photosynthesis. We attribute this to selection on
414 the *A. arenosa* subgenome to restore communication with the new plastid environment as
415 plastid genomes were maternally inherited from *A. thaliana*. Cluster 2 consists of 2,849 genes
416 that show decreased gene expression in *A. suecica* compared to *A. arenosa* and synthetic *A.*
417 *suecica*. These genes were primarily enriched for mRNA processing and cell division. Cluster

3 (122 genes) did not have a GO overrepresentation and showed an intriguing pattern discussed in the next section.

6. 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 just mentioned (Fig. 5d, Cluster 3). 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 further observation that 97 of the 122 genes (Supplementary Figure 20a) 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 97 genes and the doubled coverage for their homeologs on the *A. thaliana* subgenome, strongly suggested a homeologous exchange (HE) event resulting in AS530 carrying four copies of the *A. thaliana* subgenome and zero copies of the *A. arenosa* genome with respect to this this, roughly 2.5 Mb region of the genome (Fig. 6c). This explanation was further supported by HiC data, which showed clear evidence for interchromosomal contacts between *A. thaliana* subgenome chromosome 1 and *A. arenosa* subgenome chromosome 6 around the breakpoints of the putative HE in AS530 (Fig. 6 d,e), and by multiple discordant Illumina paired-end reads at the breakpoints between the homeologous chromosomes, which independently support the HE event (Supplementary Fig. 19a-d).

Based on this we examined the two outlier *A. suecica* accessions for the *A. thaliana* subgenome (Fig. 5a; “AS150” and “ASÖ5”), and found that they likely 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 roughly 1Mb in size, see Supplementary Figure 18). This demonstrates that HE occurs in *A. suecica* and contributes to the intraspecific variation we observed in gene expression (Fig 5a, b). 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 (Supplementary Figure 19) 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/ASÖ5 are missing 53 genes that were only present on the *A. thaliana* subgenome.

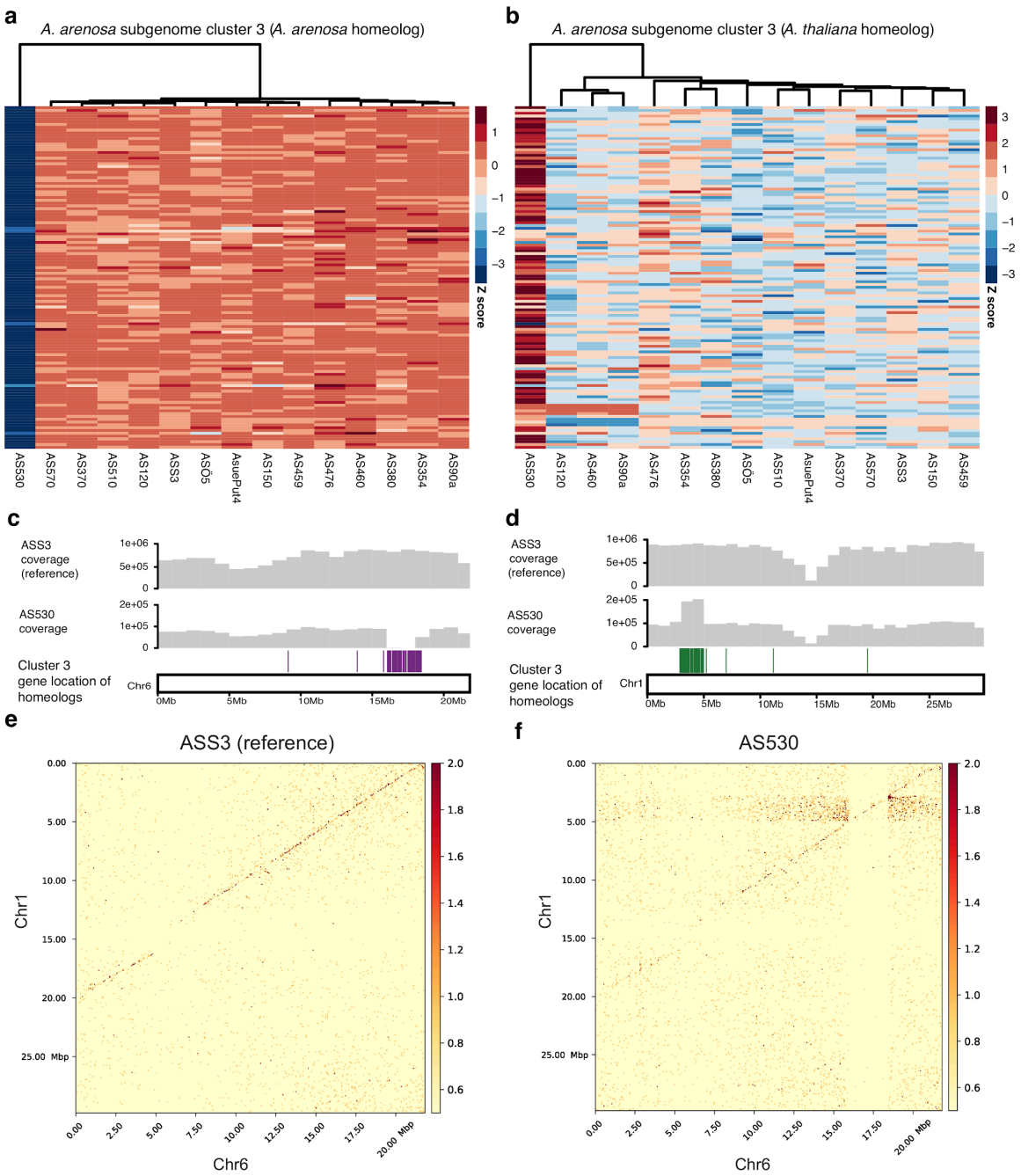


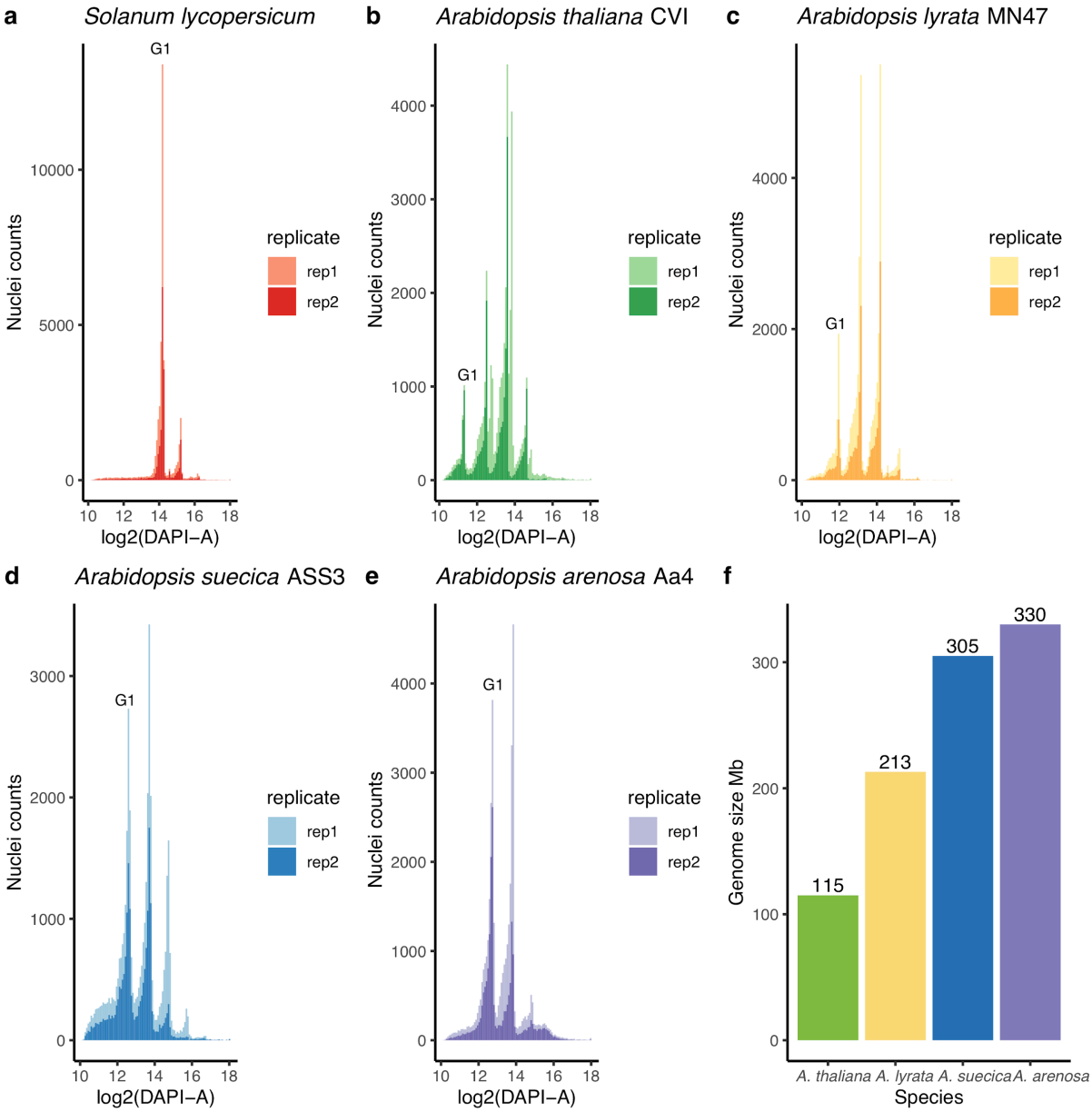
Figure 6. Homeologous exchange contributes to expression variance within *A. suecica*. a Cluster 3 of Fig. 5d explains the outlier accession AS530 which is not expressing a cluster of genes on the *A. arenosa* subgenome. **b** Homeologous genes of this cluster on the *A. thaliana* subgenome of *A. suecica* show the opposite pattern and are more highly expressed in AS530 compared to the rest of the population. **c** 97 of the 122 genes from cluster 3 are located in close proximity to each other on the reference genome but appear to be deleted in AS530 based on sequencing coverage. **d** The *A. thaliana* subgenome homeologs have twice the DNA coverage, suggesting they are duplicated. **e** HiC data show (spurious) interchromosomal contacts at 25 Kb resolution between chromosome 1 and chromosome 6 around the breakpoint of the cluster of 97 genes in AS530 but not in reference accession ASS3.

Conclusion

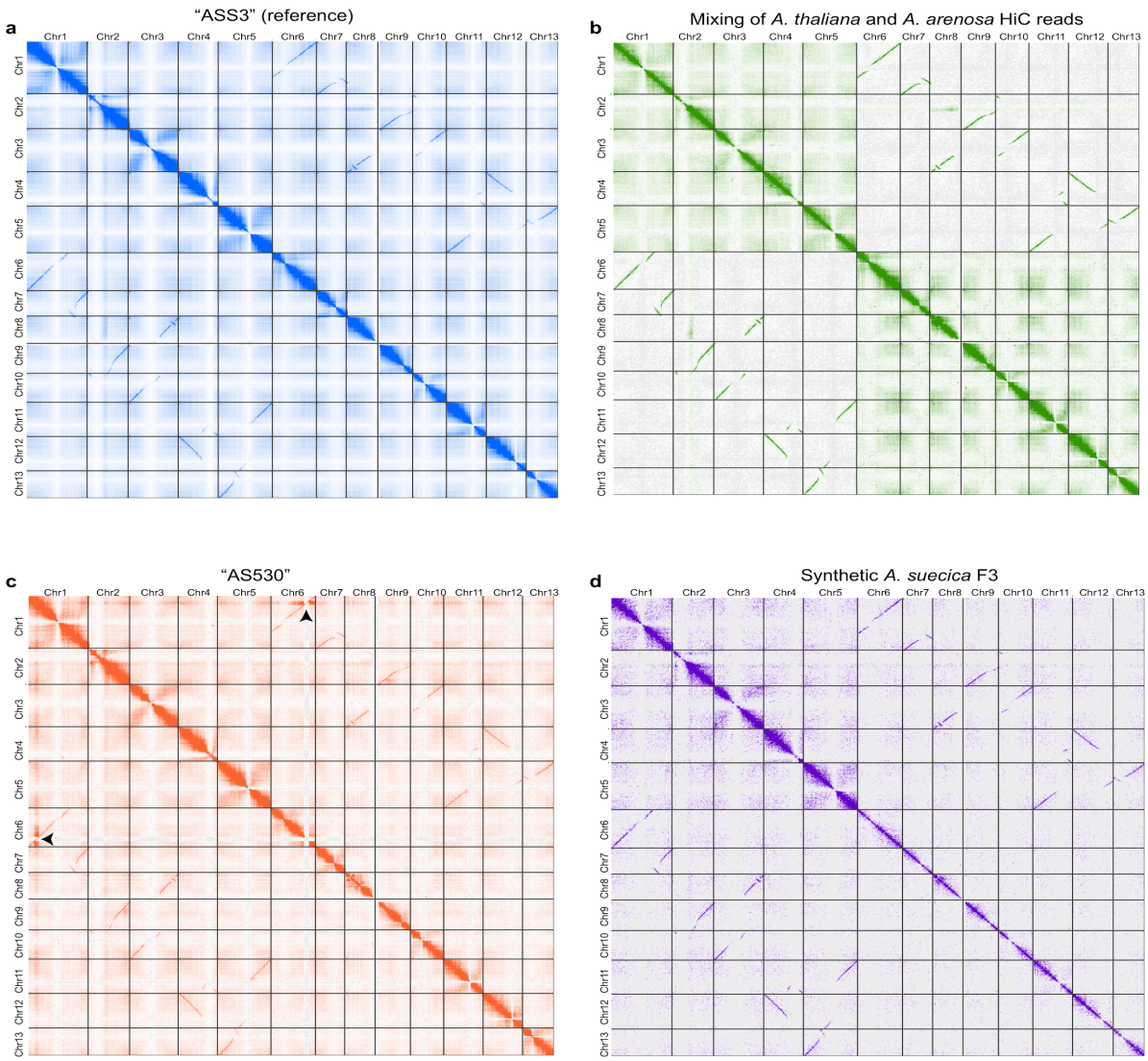
This study has focused on the process of polyploidization in a young allotetraploid species, *A. suecica*, that was generated roughly 16 kya through the hybridization of two species, *A. thaliana* and *A. arenosa*, which differ substantially in everything from genome size and chromosome number to mating system and ecology. Our study is one of relatively few to focus on a natural polyploid rather than a domesticated species, and is unparalleled in its resolution thanks to one of the parents being a major model species.

The main conclusion of our study is that polyploid speciation, at least in this case, was clearly 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 crosses[16]. We also find no evidence for the kind of dramatic "genome shock" that has often been suggested to accompany polyploidization and hybridization. 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, coupled with the observation that experimentally generated *A. suecica* are often unviable and do exhibit evidence of genome rearrangements, suggest that the most important bottleneck in polyploid speciation may be selective, and that domesticated polyploids may not be representative of natural polyploidization. Darwin famously argued that "Natura non facit saltum"[63] — we suggest that polyploidization is no exception from this.

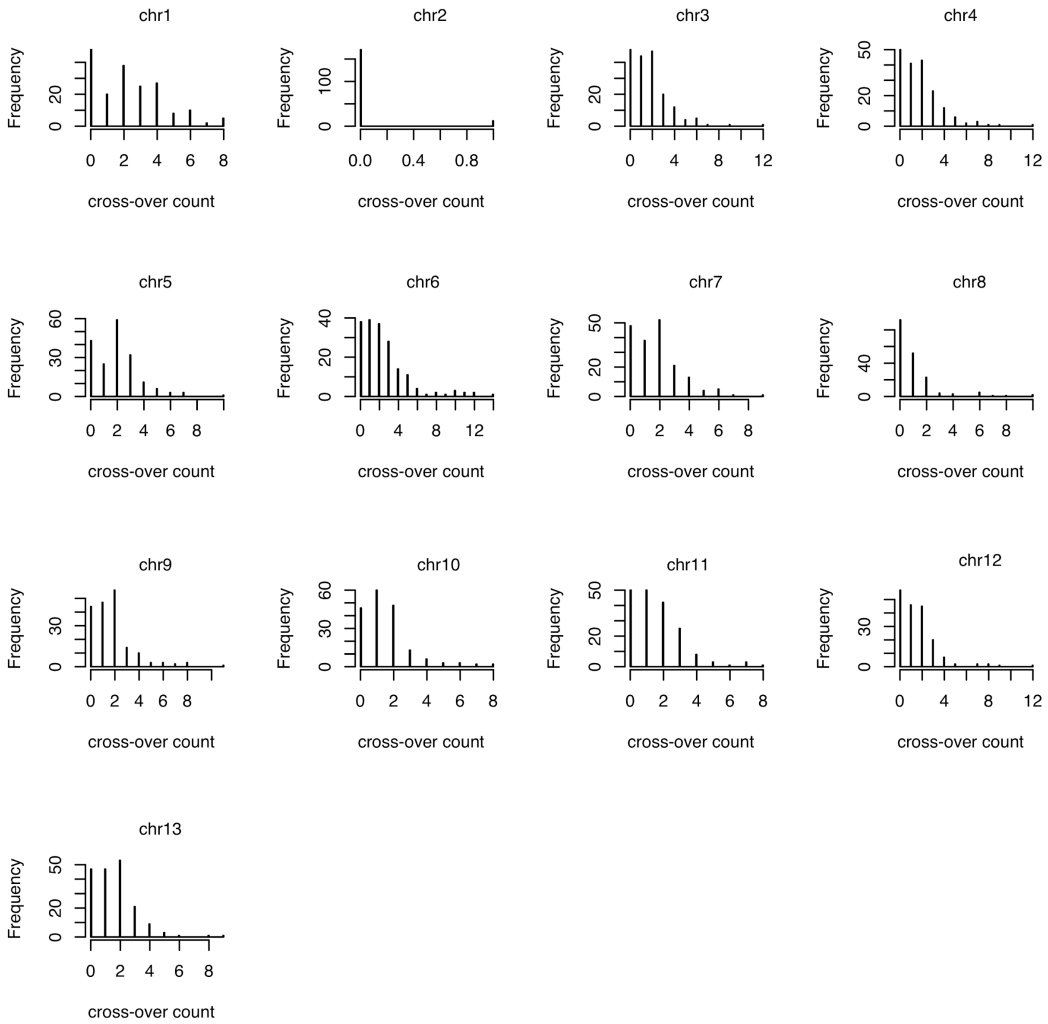
Supplementary figures



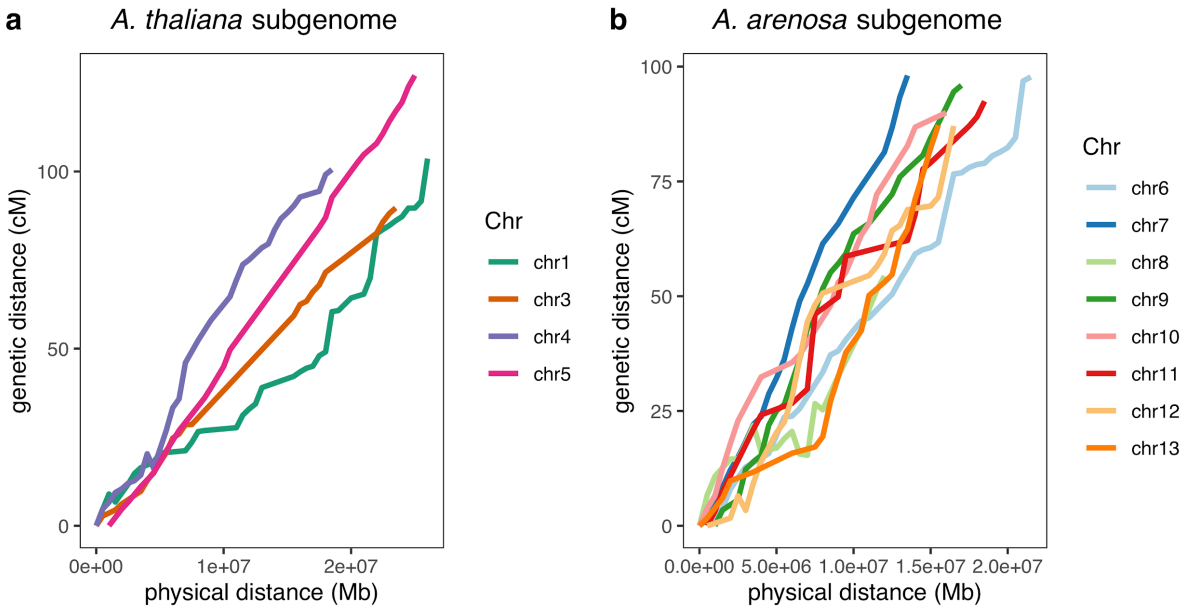
Supplementary Figure 1. Measuring genome sizes of *Arabidopsis* species using flow cytometry. a FACs sorting of *Solanum lycopersicum* cells from 3 week old leaf tissue for two replicates. G1 represents the peak denoting the G1 phase of the cell cycle. Cells in the G1 phase have 2C DNA content (i.e. a 2N genome). **b** *A. thaliana* "CVI" accession **c** *A. lyrata* "MN47" (the reference accession) **d** *A. suecica* "ASS3" (the reference accession) **e** autopolyploid *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 .



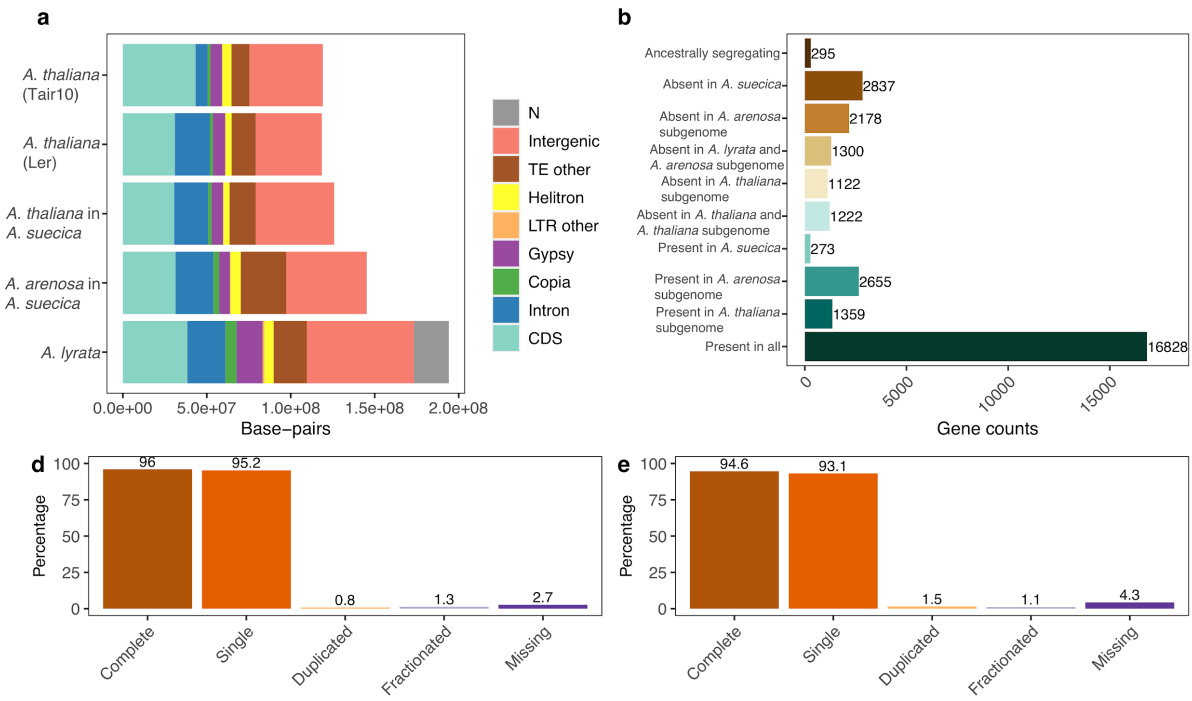
Supplementary Figure 2. HiC as a tool to investigate structural rearrangements. **a** HiC contact map for the full chromosome-level genome assembly of *A. suecica*. **b** Mixing of *A. thaliana* and *A. arenosa* HiC reads suggest interchromosomal contacts between homeologous chromosomes is a result of mis-mapping for HiC reads. Such mis-mapping is typically filtered out in short read DNA and RNA datasets using insert size and proper pairs mapping filters, however in HiC long range chromosomal contacts are not filtered out. **c** Accession "AS530" with the region of homeologous exchange highlighted with an arrow (figure 6), no other rearrangements were observed. **d** HiC of synthetic *A. suecica* (F3).



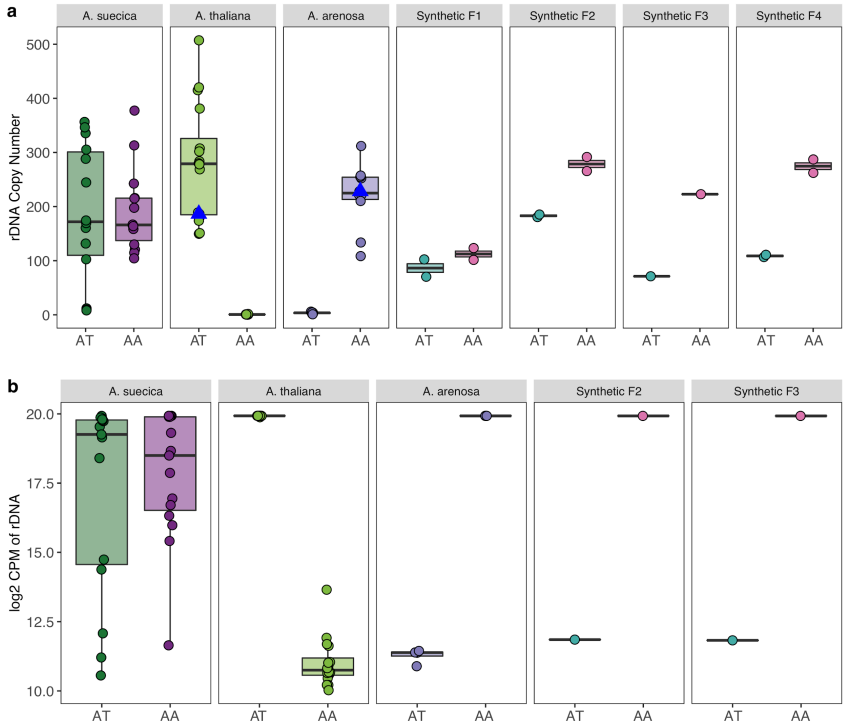
522
523 **Supplementary Figure 3. Crossover counts in an *A. suecica* F2 population.** Per
524 chromosome crossover counts in our F2 population (N=185). Chromosome 2 had too few
525 SNPs to be analysed in our cross due to the recent bottleneck in *A. suecica*[16].
526
527



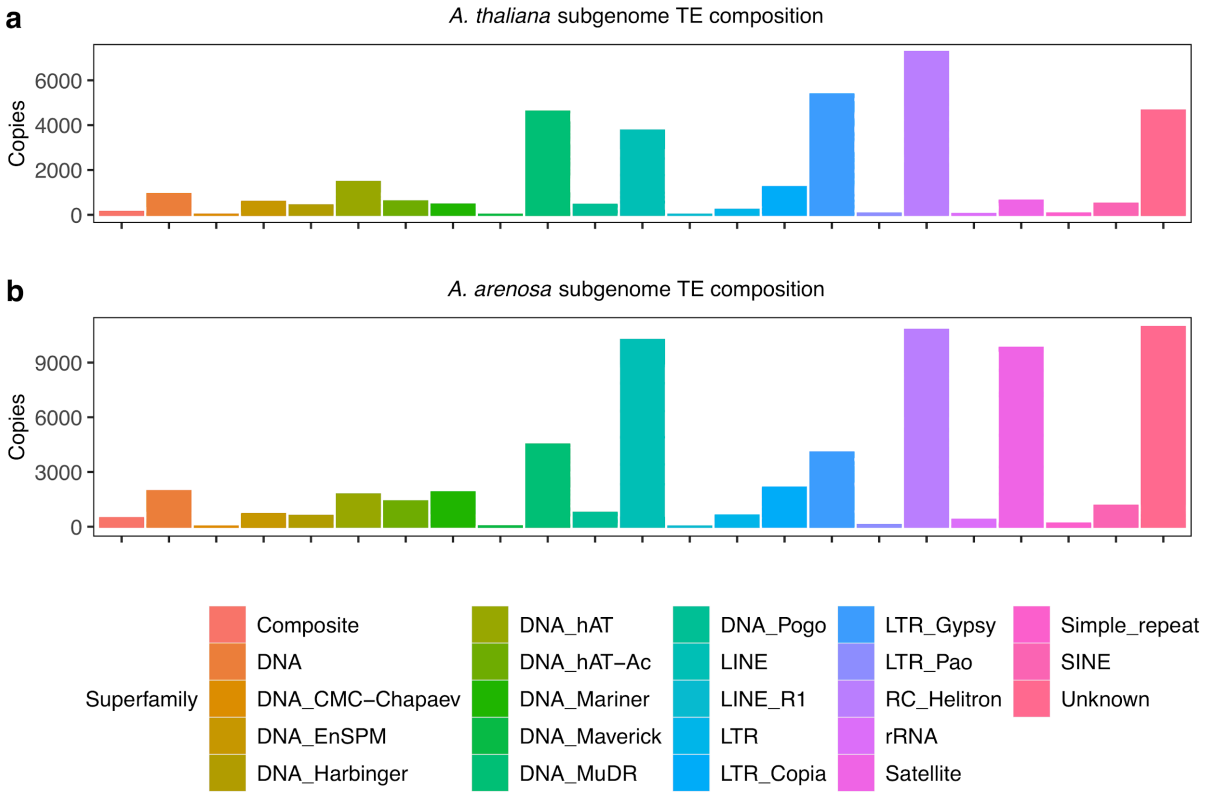
Supplementary Figure 4. A genetic map for *A. suecica*. Physical distance (Mb) vs genetic distance (cM) is plotted for each: **a** *A. thaliana* subgenome and; **b** *A. arenosa* subgenome chromosome. Chromosome 2 is not plotted as there are too few SNPs on this chromosome in our cross, due to the recent bottleneck in *A. suecica*[16]



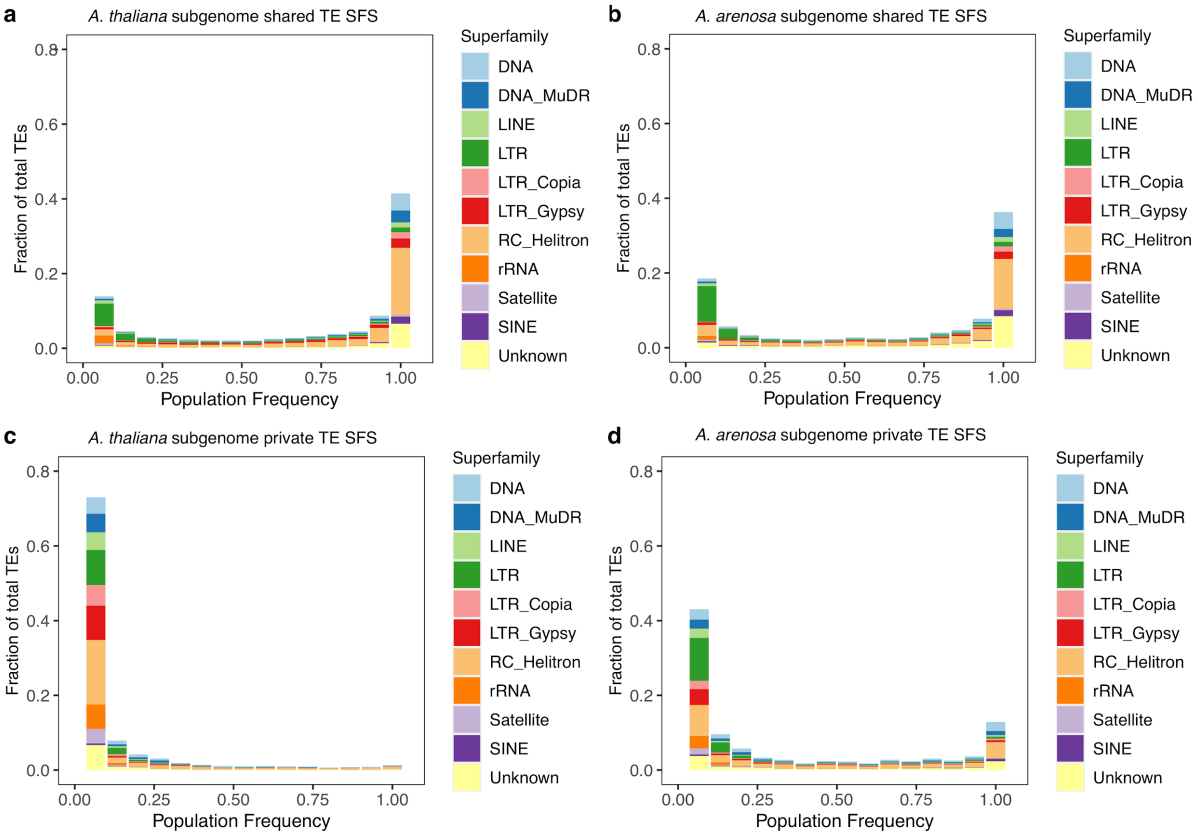
Supplementary Figure 5. Genome composition and orthologous gene relationships in *A. suecica*. a Genome composition of the *A. suecica* subgenomes and the ancestral genomes of *A. thaliana* and *A.lyrata* (here a substitute reference for *A. arenosa* because it is annotated). Ler = Landsberg *Erecta* (*A. thaliana*)[64]. The difference in CDS between the Tair10 annotation with that of and Ler and the *A. thaliana* subgenome is most likely due to the manual curation of Tair10 and the de novo gene prediction pipeline used in this paper. **b** Counts of orthologous relationships between the subgenomes of the reference *A. suecica* genome and the reference *A. thaliana* and *A. lyrata* genome. Ancestrally segregating genes are genes shared between the *A. thaliana* reference and the *A. arenosa* subgenome or shared between the *A. lyrata* reference and the *A. thaliana* subgenome. Therefore they most likely represent genes ancestrally segregating in the ancestor of *A. thaliana* and *A. lyrata*. BUSCO analysis of *A. suecica* using the BUSCO set for eudicots for the **c** *A. thaliana* and **d** *A. arenosa* subgenome.



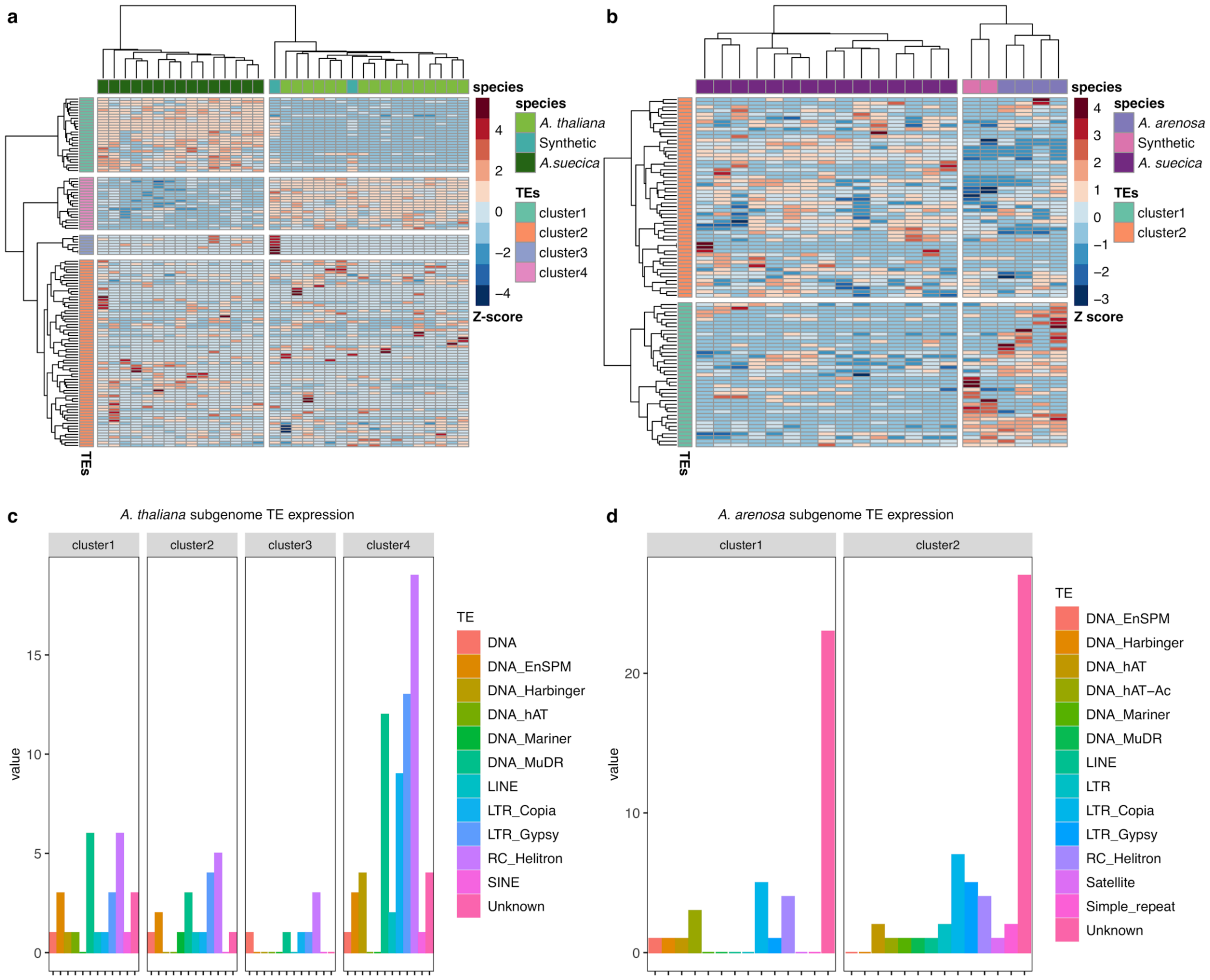
Supplementary figure 6. **rDNA copy number variation and expression. a** Copy number of *A. thaliana* and *A. arenosa* rDNA in natural *A. suecica*, ancestral species and synthetic lines. Blue triangles represent the *A. thaliana* and *A.arenosa* parent lines of the synthetic *A. suecica* cross. AT represents results when mapping to the *A. thaliana* consensus sequence and AA to the *A. arenosa* consensus sequences for the 45S rRNA **b** Expression (log2 CPM) of *A. thaliana* and *A. arenosa* rDNA in natural *A. suecica*, ancestral species and synthetic lines. Accessions with log2 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).



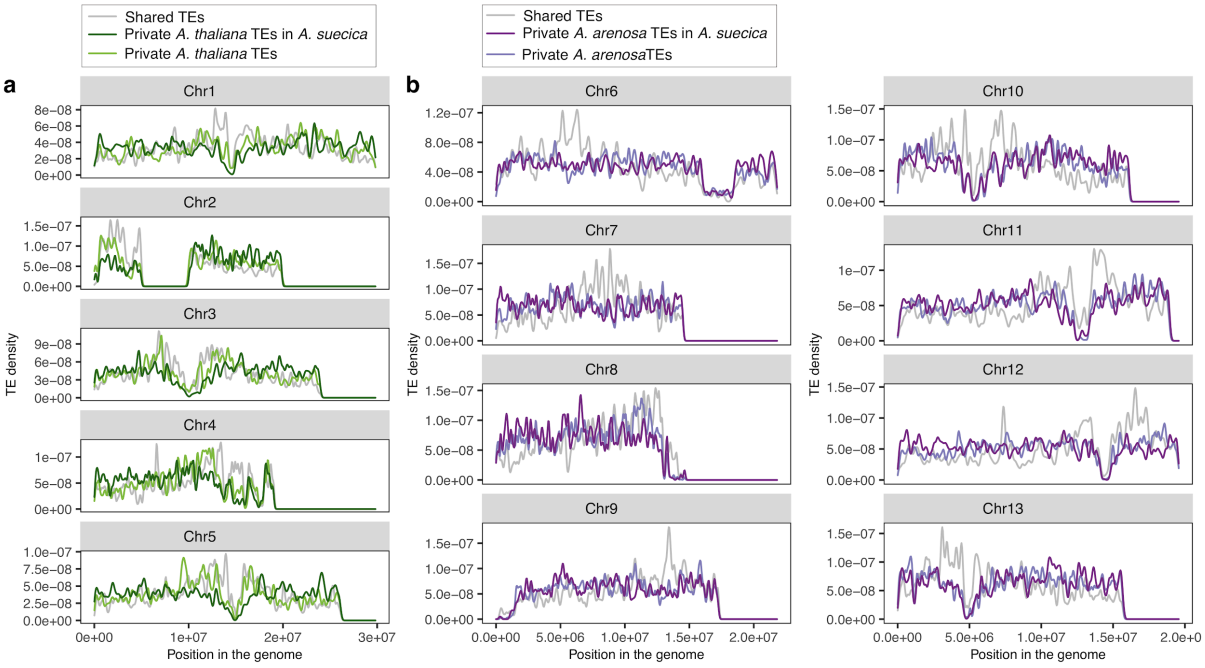
Supplementary Figure 7. TE-composition of the *A. suecica* reference genome. TE composition of the **a** *A. thaliana* and **b** *A. arenosa* subgenome of *A. suecica*.



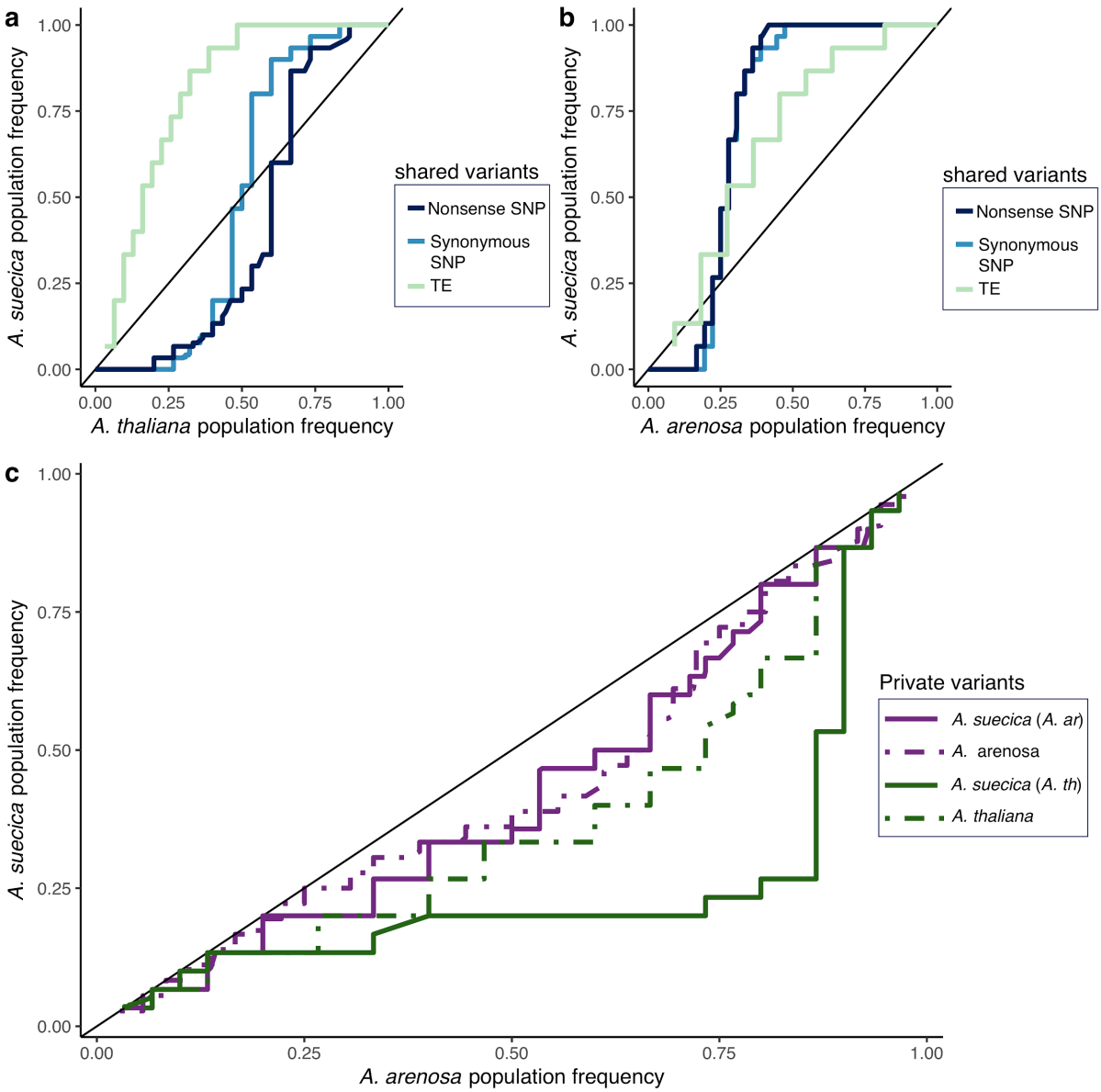
Supplementary Figure 8. Site frequency spectrum (SFS) of shared TEs and private TEs in *A. suecica* broken down by TE family. Shared TE SFS for the **a** *A. thaliana* and **b** *A. arenosa* subgenome. Private TE SFS for the **c** *A. thaliana* and **d** *A. arenosa* subgenome.



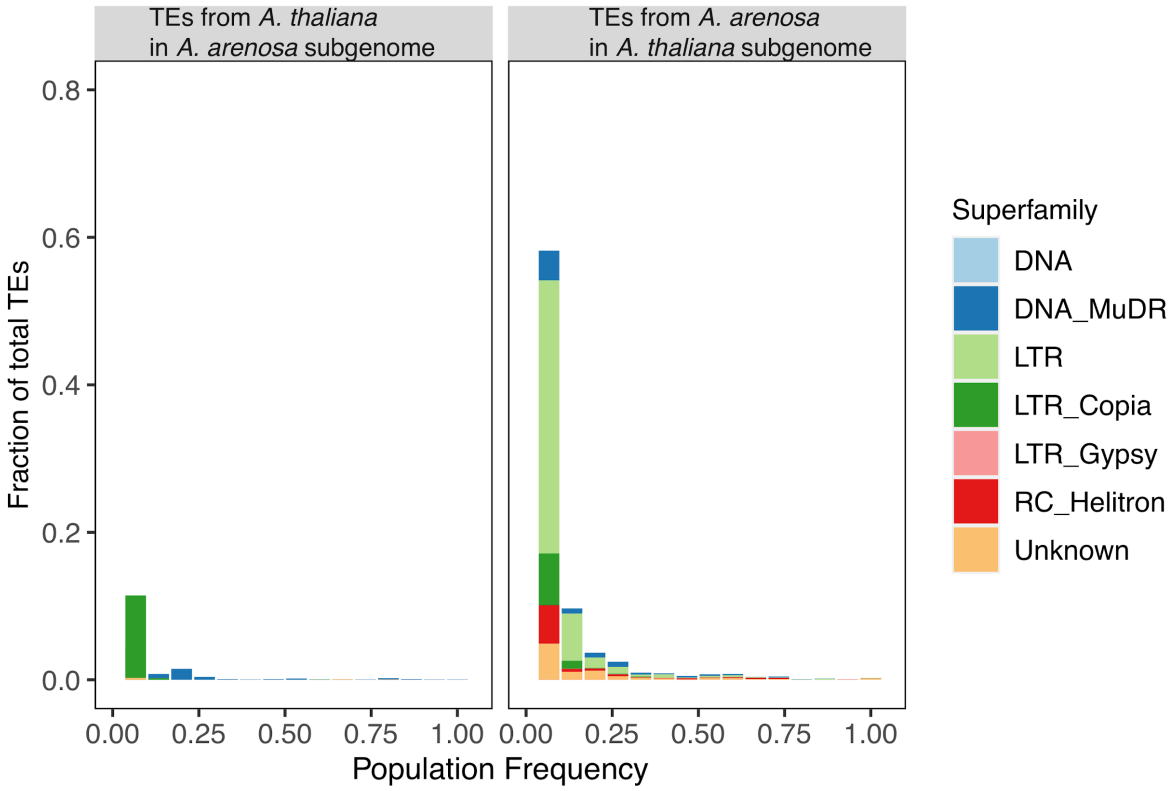
Supplementary Figure 9. Analysis of TE expression in *A. suecica*. Patterns of TE expression in natural and synthetic *A. suecica* show that allopolyploidy is not accompanied by an overall up-regulation in TE expression as predicted by the "genome shock" hypothesis. **a** Heatmap of TE expression for the *A. thaliana* subgenome of *A. suecica* (dark green) synthetic *A. suecica* (cyan) and *A. thaliana* (light green). **b** Heatmap 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.



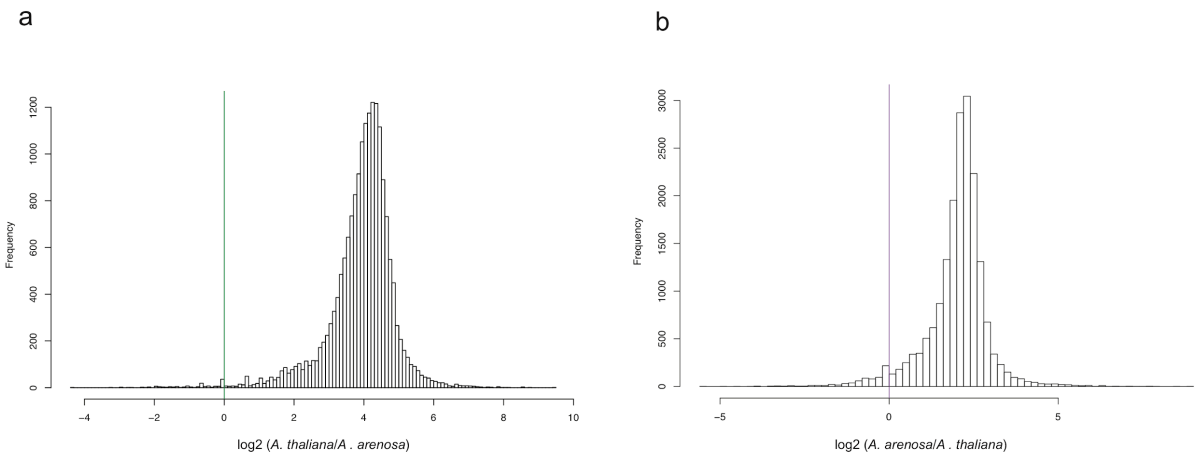
Supplementary Figure 10. Genomic distribution of TEs in the *A. suecica* genome. a Shared TEs in the population between *A. thaliana* and the *A. thaliana* subgenome of *A. suecica.* Shared TEs are likely older than private TEs and are enriched around the pericentromeric regions in the *A. thaliana* subgenome. Private TEs are enriched in the chromosomal arms for both species, where protein coding gene density is higher (Fig. 1b). **b** as in **a** 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.



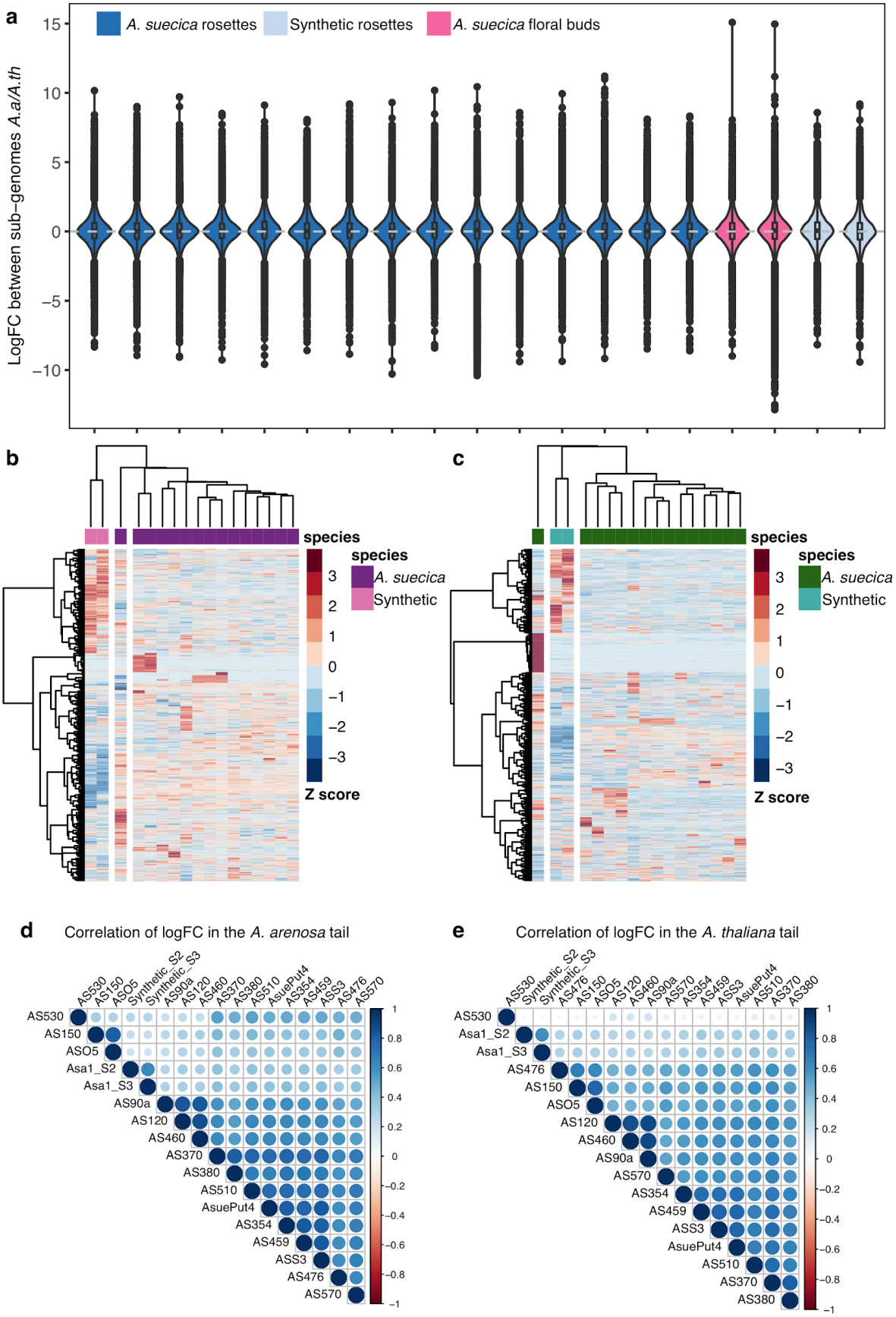
Supplementary Fig 11. Patterns of selection in *A. suecica.* **a** Comparison of shared variation (Nonsense SNPs, synonymous SNPs, and TEs) population frequencies in the *A. thaliana* subgenome of 15 natural *A. suecica* accessions and the closest 31 *A. thaliana* accessions. **b** Comparison of shared variation (Nonsense SNPs, synonymous SNPs, and TEs) frequencies in *A. arenosa* subgenome of 15 *A. suecica* accessions and 11 Swedish *A. arenosa* lines. Although results may be affected by the sampling and potential misidentification of the ancestral populations, the current data suggests strong purifying selection on the *A. thaliana* subgenome of *A. suecica* for shared nonsense SNPs, due to the lower population frequency of nonsense SNPs (the similar pattern observed for synonymous SNPs could be explained by background selection in this case). The similar pattern on both of the subgenomes for TEs shows a bottleneck effect. **c** Plotting quantile pairs of the population frequencies of private nonsynonymous and synonymous SNPs in *A. suecica* and ancestral populations against each other, each species shows evidence of evolution under purifying selection, since population frequency quantiles of nonsynonymous SNPs are skewed to lower values than population frequency quantiles of synonymous SNPS.



Supplementary Figure 12. Population frequencies of presence-absence calls for TEs that have mobilized between the subgenomes in *A. suecica.* **a** TEs ancestrally from *A. thaliana* present in the *A. arenosa* subgenome of *A. suecica* and **b** TEs ancestrally from *A. arenosa* present in the *A. thaliana* subgenome of *A. suecica*. Y axis represents the fraction of total TEs that have mobilized between the subgenomes.

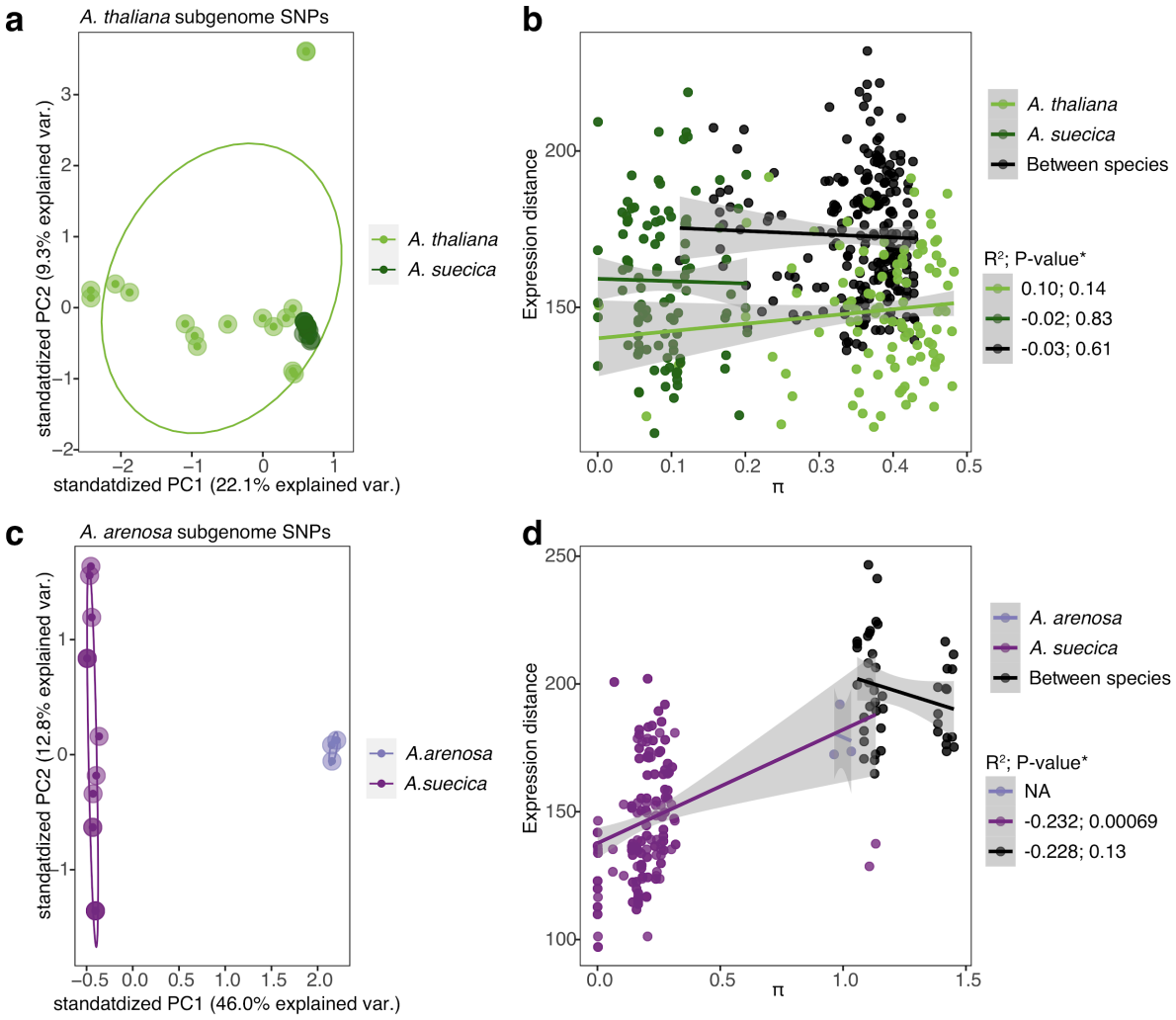


Supplementary Figure 13. Gene filtered for cross mapping in RNA-seq. **a** A histogram of log2 read counts (*A. thaliana*/*A. arenosa*) of genes on the *A. thaliana* subgenome in *A. suecica.* The green vertical line represents a log fold-change of 0 **b** A histogram of log2 read counts (*A. arenosa*/*A. thaliana*) of genes on the *A. arenosa* subgenome in *A. suecica.* The purple vertical line represents a log fold-change of 0

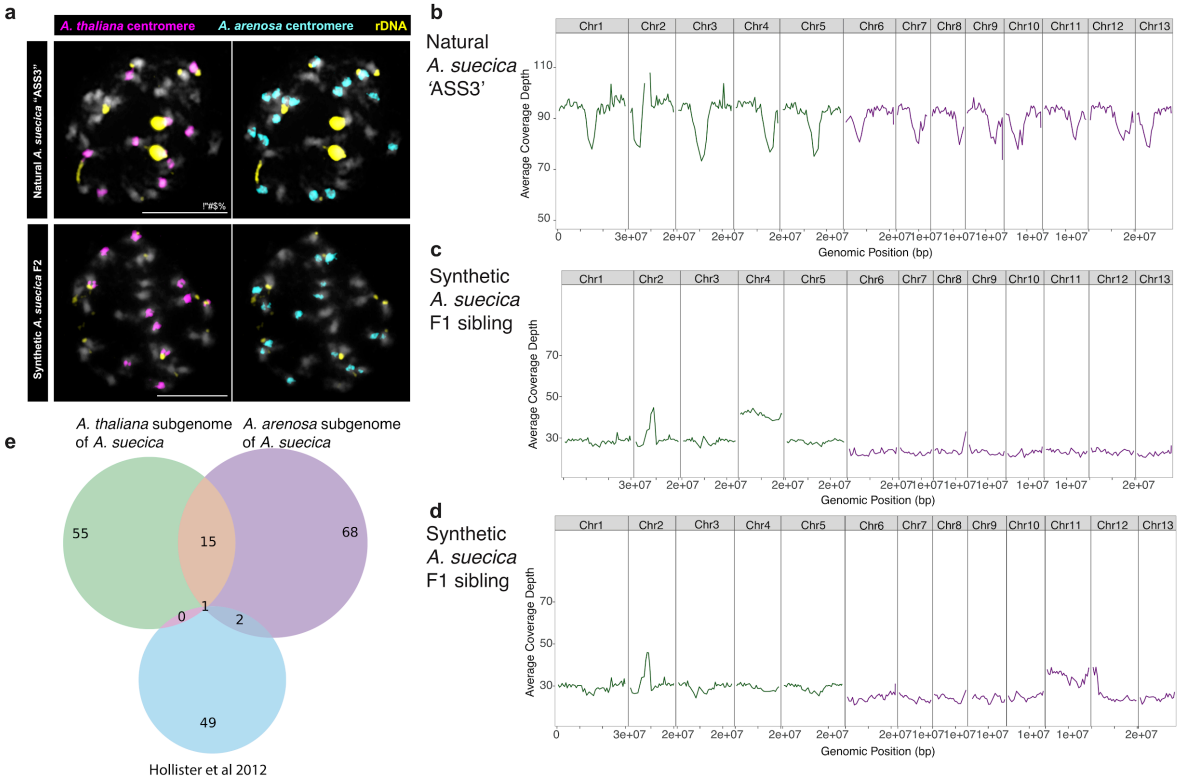

a
A. suecica rosettes
Synthetic rosettes
A. suecica floral buds
LogFC between sub–genomes A.a/A.th
15
10
5
0
−5
−10
b
species
species
A. suecica
Synthetic
3
2
1
0
−1
−2
−3
Z score
c
species
species
A. suecica
Synthetic
3
2
1
0
−1
−2
−3
Z score
d
Correlation of logFC in the A. arenosa tail
AS530
AS150
ASO5
Synthetic_S2
Synthetic_S3
AS90a
AS120
AS460
AS370
AS380
AS510
AsuePut4
AS354
AS459
ASS3
AS476
AS570
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
e
Correlation of logFC in the A. thaliana tail
AS530
Synthetic_S2
Synthetic_S3
AS476
AS150
ASO5
AS120
AS460
AS90a
AS570
AS354
AS459
ASS3
AsuePut4
AS510
AS370
AS380
Asa1_S2
Asa1_S3
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

635
636

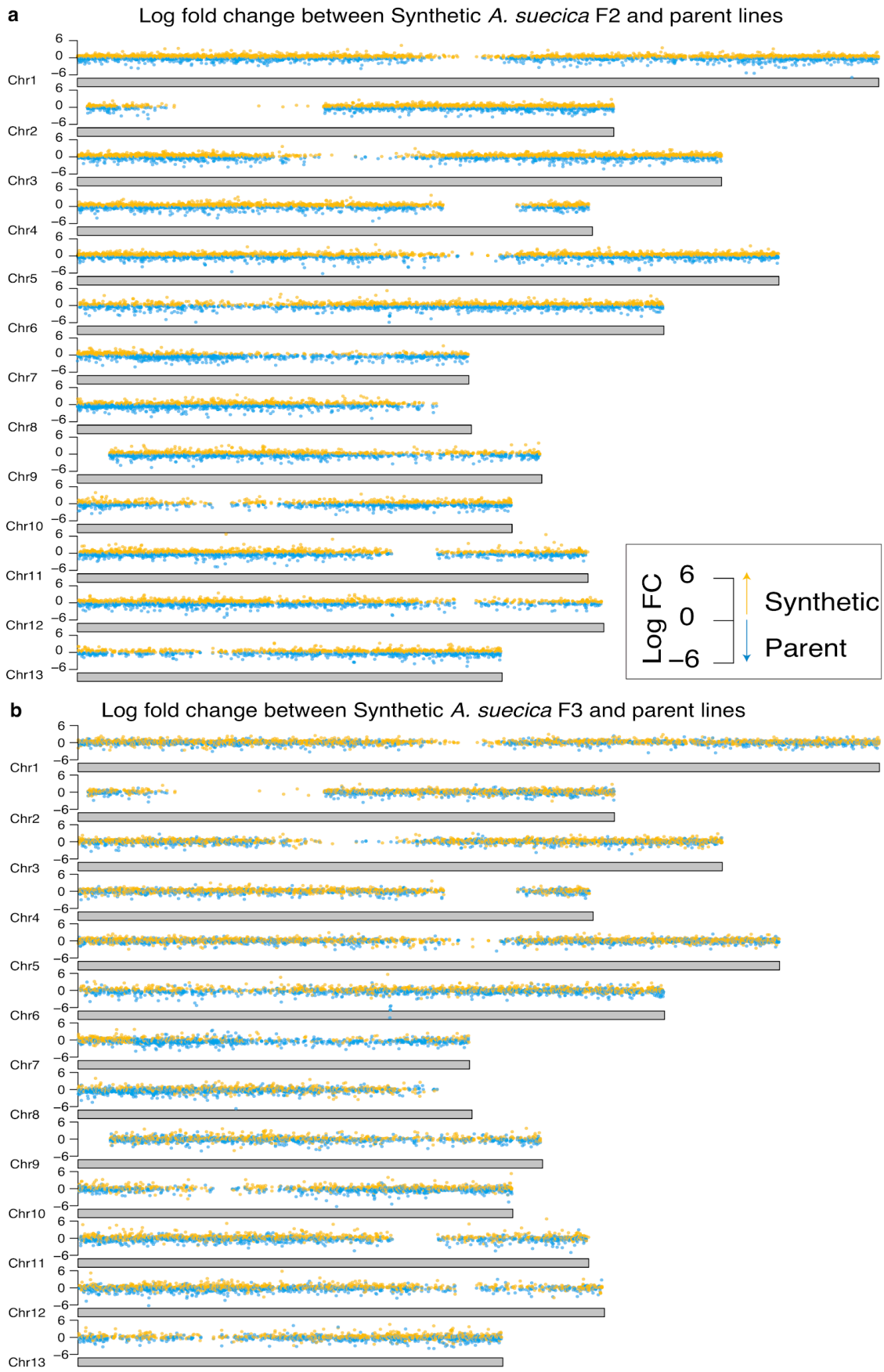
Supplementary figure 14. Expression differences between subgenomes in natural and synthetic *A. suecica*. a The distribution of expression differences across homeologous gene pairs in natural and synthetic *A. suecica*. **b** A heatmap of log fold change for genes in the top 5% biased toward the *A. 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**



Supplementary Figure 15. 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=280,000 biallelic SNPs), of the analyzed 13,647 genes in gene expression in addition to 500bp up and downstream of each gene sequence **b** Correlation of π (pairwise genetic differences) and expression distance (i.e. euclidean distance) for 13,647 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= 233,070 biallelic SNPs), of the analyzed 13,647 genes in gene expression in addition to 500bp up and downstream of each gene sequence **d** Correlation of Pi (pairwise genetic differences) and expression distance (i.e. euclidean distance) for 13,647 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.

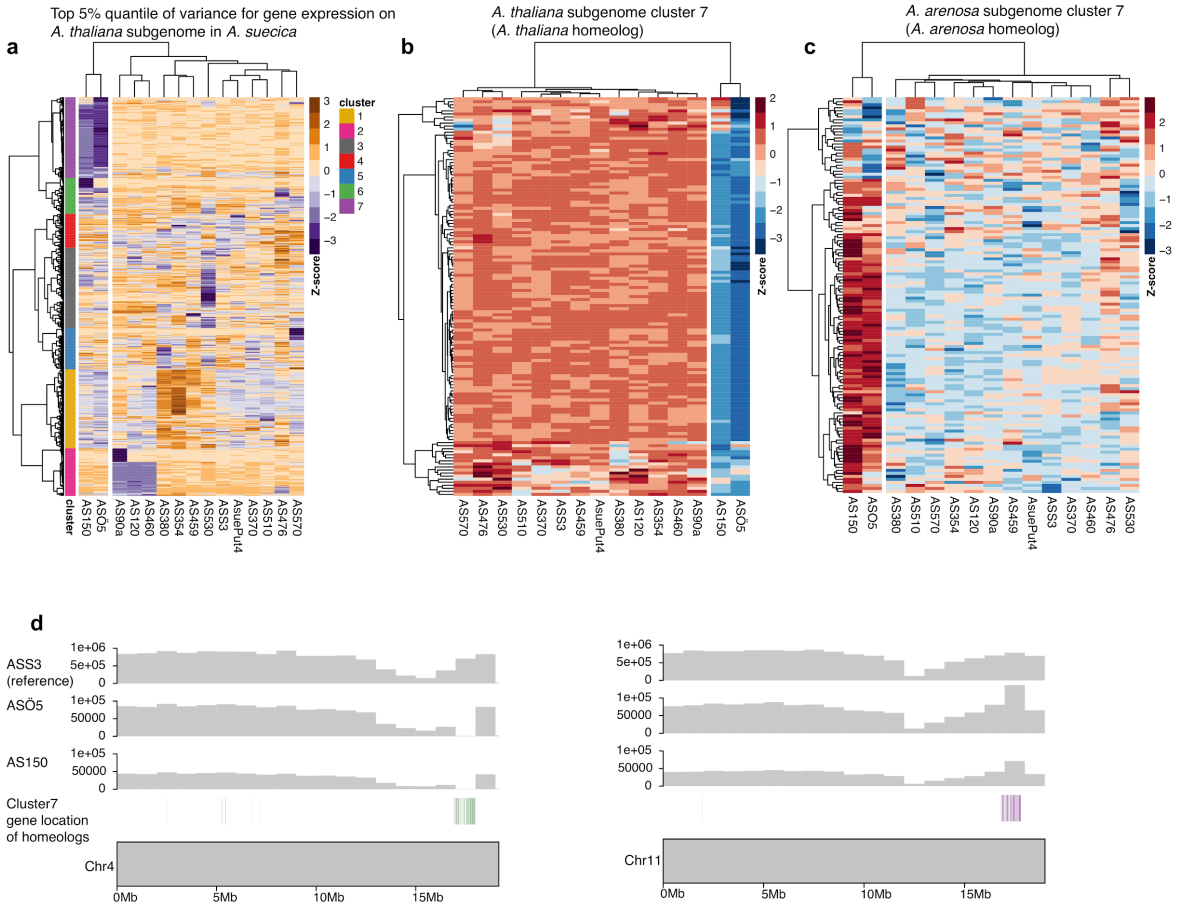


Supplementary Figure 16. **Aneuploidy is frequent in synthetic *A. suecica***. **a** Comparison of FISH analyses of the reference natural *A. suecica* "ASS3 "and synthetic *A. suecica.* Synthetic *A. suecica* shows aneuploidy in both subgenomes in the F2 generation (gain of one chromosome on the *A. 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 in synthetic *A. suecica*, chromosome 4 in **c** and chromosome 11 in **d e** overlap of genes involved in cell division from figure 5e and genes previously shown to play a role in the adaptation to autopolyploidy in *A. arenosa*[59]. The little overlap in genes between *A. suecica* and *A. arenosa* highlights that successful meiosis in polyploids is likely a complex trait.

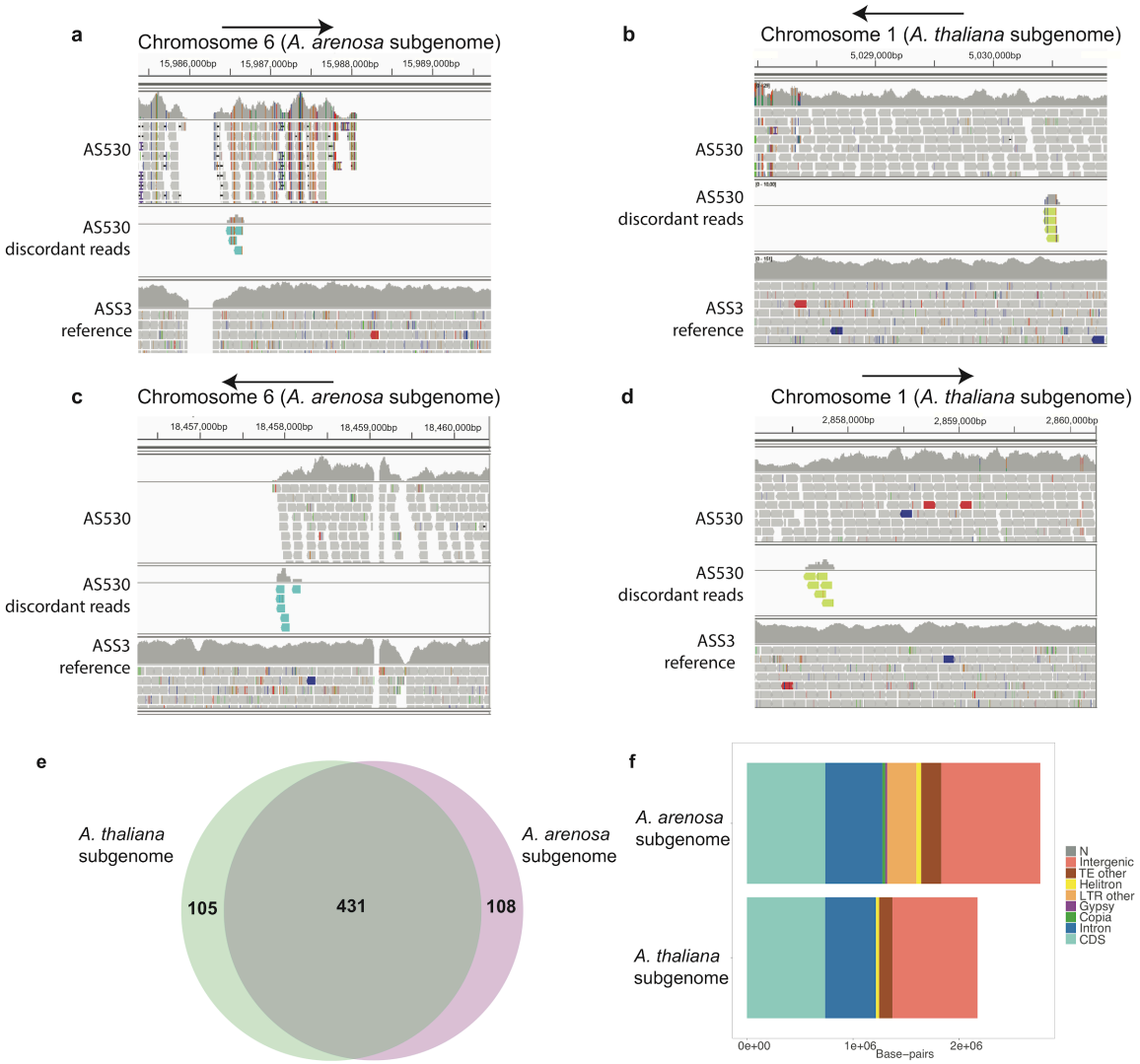

a
Log fold change between Synthetic A. suecica F2 and parent lines
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
6
0
−6
Log FC
Synthetic
Parent
b
Log fold change between Synthetic A. suecica F3 and parent lines

688

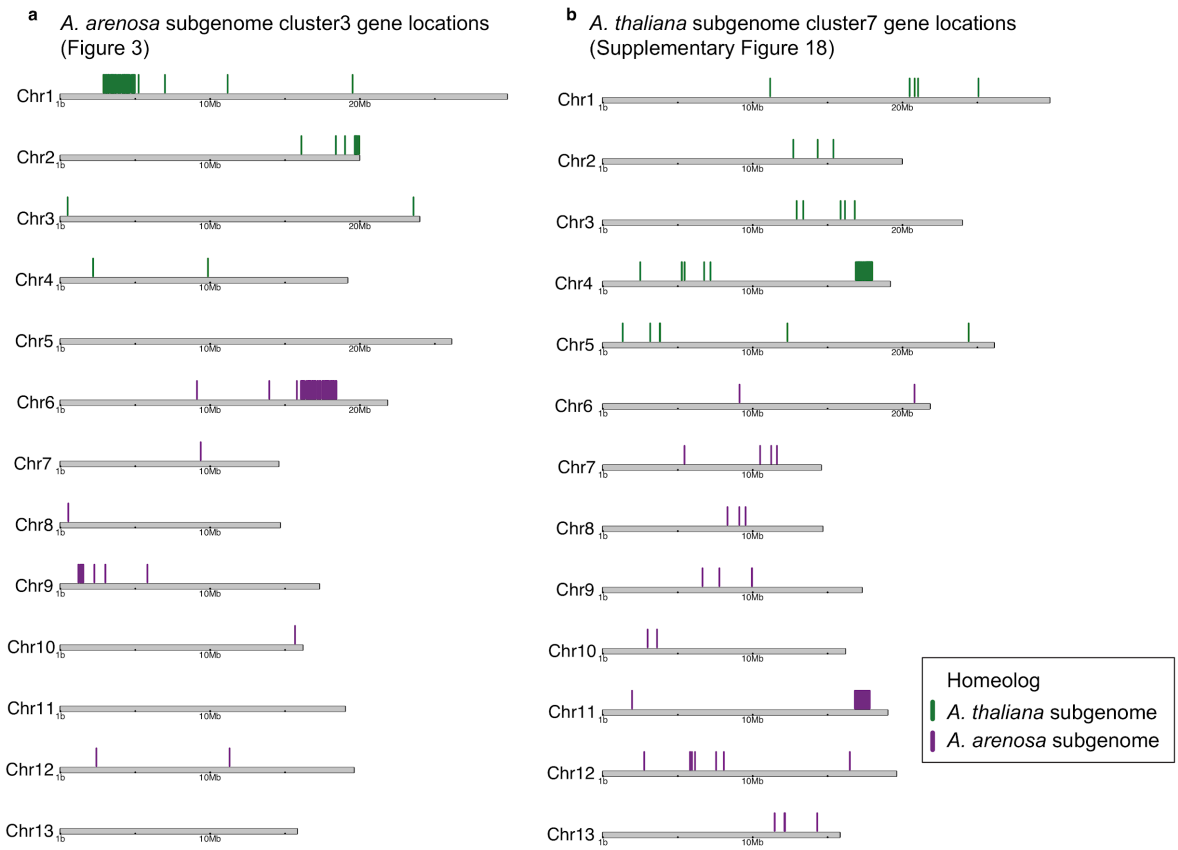
Supplementary Figure 17. No aneuploidy in synthetic *A. suecica* lines used for RNA seq based on log fold change to parent lines. Log fold change for gene expression in **a** the 2nd and **b** the 3rd generation of synthetic *A. suecica* compared to the parent lines (Chr1-5 represent comparisons with the *A. thaliana* parents and Chr6-13 represent comparisons with the *A. arenosa* parent). No clear signal of aneuploidy (i.e. an elevated increase in expression for a full chromosome) is evident.



Supplementary Figure 18. Homeologous exchange contributes to expression variance within *A. suecica* on the *A. thaliana* subgenome. A Taking the top 5% quantiles (N=646) for variation in gene expression for the *A. thaliana* subgenome we find a large cluster 7 (N=131) where the two outlier accessions in our PCA ("AS150" and "ASÖ5") are expressing these genes differently to the rest of the population. **b** 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 **c** shows the opposite pattern and are higher expressed in "AS150" and "ASÖ5" compared to the rest of the population. **d** 103 of the 131 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" as they do not have DNA sequencing coverage. The *A. arenosa* subgenome homeologs (located on chromosome 11) have twice the DNA coverage, suggesting they are duplicated, in agreement with the expectations of HE event.



Supplementary Figure 19 Discordant read analysis supports HE in *A. suecica* **a** IGV screen grab of reads mapped to the beginning of the likely HE event in chromosome 6 (at ~ 15.9Mb) before coverage depth decreases to 0 in "AS530". Arrows point to the direction of the break along the chromosome. Discordant read pairs (cyan) map between the *A. arenosa* subgenome on chromosome 6 and the read pair maps (green) maps to the homeologous chromosome 1 on the *A. thaliana* subgenome (at ~5Mb) in **b**. The end of the likely HE event in chromosome 6 (at ~18.4Mb). Discordant read pairs map (cyan) between the *A. arenosa* subgenome in **c** and the read pair (green) maps to chromosome 1 (at ~2.8Mb) 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 in this region and 105 genes are specific to the *A. thaliana* subgenome. **f** Composition of the syntenic regions between the two subgenomes



Supplementary Figure 20 Genomic locations of genes investigated for HE signatures in *A. suecica*.

a **Genes biased in expression in rosettes for both sub-genomes in *A. suecica***

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0009768	photosynthesis, light harvesting in phot...	19	17	0.59	2.5e−24
2	GO:0018298	protein−chromophore linkage	34	17	1.05	2.3e−17
3	GO:0015979	photosynthesis	181	72	5.61	2.1e−16
4	GO:0015995	chlorophyll biosynthetic process	52	18	1.61	3.0e−14
5	GO:0009735	response to cytokinin	174	28	5.39	6.7e−13
6	GO:0042742	defense response to bacterium	278	34	8.61	2.0e−12
7	GO:0055114	oxidation−reduction process	960	80	29.74	6.0e−10
8	GO:0019253	reductive pentose−phosphate cycle	15	9	0.46	2.0e−09
9	GO:0009773	photosynthetic electron transport in pho...	10	7	0.31	2.9e−09
10	GO:0009409	response to cold	281	32	8.71	1.8e−08
11	GO:0009645	response to low light intensity stimulus	14	7	0.43	2.2e−08
12	GO:0009416	response to light stimulus	533	58	16.51	5.6e−08
13	GO:0010114	response to red light	56	12	1.74	1.1e−07
14	GO:0032544	plastid translation	11	6	0.34	3.4e−07
15	GO:0019761	glucosinolate biosynthetic process	27	10	0.84	5.8e−07
16	GO:0010218	response to far red light	44	10	1.36	7.0e−07
17	GO:0010258	NADH dehydrogenase complex (plastoquinon...	4	4	0.12	9.1e−07
18	GO:0010206	photosystem II repair	13	6	0.40	1.2e−06
19	GO:0009767	photosynthetic electron transport chain	39	17	1.21	1.8e−06
20	GO:0009644	response to high light intensity	56	12	1.74	6.1e−06
21	GO:1901259	chloroplast rRNA processing	17	6	0.53	7.9e−06
22	GO:0009769	photosynthesis, light harvesting in phot...	3	3	0.09	3.0e−05
23	GO:0010207	photosystem II assembly	19	7	0.59	8.9e−05
24	GO:0009637	response to blue light	79	11	2.45	9.7e−05
25	GO:0009817	defense response to fungus, incompatible...	36	7	1.12	1.0e−04
26	GO:0010196	nonphotochemical quenching	11	4	0.34	0.00025
27	GO:0019464	glycine decarboxylation via glycine clea...	5	3	0.15	0.00028
28	GO:0006782	protoporphyrinogen IX biosynthetic proce...	12	4	0.37	0.00037
29	GO:0006412	translation	510	33	15.80	0.00051
30	GO:0009098	leucine biosynthetic process	13	4	0.40	0.00052
31	GO:0043489	RNA stabilization	7	3	0.22	0.00094

b **Genes biased in expression in floral buds for both sub-genomes in *A. suecica***

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0009733	response to auxin	250	47	18.63	1.2e−08
2	GO:0009753	response to jasmonic acid	181	35	13.49	3.5e−07
3	GO:0009751	response to salicylic acid	141	26	10.51	1.9e−06
4	GO:0080167	response to karrikin	87	20	6.48	4.5e−06
5	GO:0009737	response to abscisic acid	430	61	32.05	1.5e−05
6	GO:0010224	response to UV−B	49	13	3.65	4.2e−05
7	GO:0055085	transmembrane transport	785	93	58.50	6.2e−05
8	GO:0009739	response to gibberellin	98	19	7.30	6.5e−05
9	GO:0071456	cellular response to hypoxia	148	25	11.03	0.00017
10	GO:0006355	regulation of transcription, DNA−templat...	1356	137	101.06	0.00022
11	GO:1905582	response to mannose	6	4	0.45	0.00041
12	GO:0043496	regulation of protein homodimerization a...	3	3	0.22	0.00041
13	GO:0071492	cellular response to UV−A	3	3	0.22	0.00041
14	GO:0051603	proteolysis involved in cellular protein...	326	26	24.30	0.00059
15	GO:0010345	suberin biosynthetic process	11	5	0.82	0.00072
16	GO:0009723	response to ethylene	184	20	13.71	0.00076
17	GO:0043086	negative regulation of catalytic activit...	91	17	6.78	0.00080

c **Genes biased in expression in rosettes for the *A. arenosa* sub-genome of *A. suecica***

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0019253	reductive pentose−phosphate cycle	15	4	0.23	6.9e−05
2	GO:0009409	response to cold	281	13	4.39	0.00047
3	GO:0015700	arsenite transport	3	2	0.05	0.00072

d **Genes biased in expression in rosettes for the *A. thaliana* sub-genome of *A. suecica***

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0071456	cellular response to hypoxia	148	9	2.38	0.00065

e **Genes biased in expression in floral buds for the *A. arenosa* sub-genome of *A. suecica***

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0006032	chitin catabolic process	10	4	0.41	0.00049

754

755 **Supplementary Table 1. Gene ontology (GO) analysis for gene expression comparison**
756 **between whole rosettes and floral buds in *A. suecica*. a** GO enrichment for genes that
757 overlap in expression bias (log fold-change > 2 in whole rosette compared to floral buds) in
758 the *A. thaliana* and *A. arenosa* subgenome of *A. suecica*. No significant GO was found for
759 genes biased towards the *A. thaliana* subgenome of *A. suecica* for floral buds.

a

***A. thaliana* cluster1**

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0006355	regulation of transcription, DNA–templat...	1356	264	205.59	4.8e–08
2	GO:0010167	response to nitrate	26	9	3.94	9.4e–06
3	GO:0007623	circadian rhythm	113	34	17.13	2.3e–05
4	GO:0035672	oligopeptide transmembrane transport	24	12	3.64	6.5e–05
5	GO:0046323	glucose import	28	13	4.25	8.5e–05
6	GO:0009741	response to brassinosteroid	80	26	12.13	0.00015
7	GO:0009638	phototropism	13	8	1.97	0.00017
8	GO:0009723	response to ethylene	184	52	27.90	0.00029
9	GO:0009739	response to gibberellin	98	31	14.86	0.00029
10	GO:1902600	proton transmembrane transport	125	28	18.95	0.00041
11	GO:0071577	zinc ion transmembrane transport	15	8	2.27	0.00065

b

***A. thaliana* cluster2**

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0007018	microtubule–based movement	61	31	6.27	7.6e–16
2	GO:0009735	response to cytokinin	174	40	17.88	9.3e–09
3	GO:0009658	chloroplast organization	190	47	19.53	1.1e–07
4	GO:0000911	cytokinesis by cell plate formation	53	20	5.45	6.5e–07
5	GO:1901259	chloroplast rRNA processing	17	10	1.75	1.3e–06
6	GO:0006412	translation	510	78	52.42	1.8e–06
7	GO:0032544	plastid translation	11	7	1.13	2.7e–05
8	GO:0007088	regulation of mitotic nuclear division	49	14	5.04	7.2e–05
9	GO:0006268	DNA unwinding involved in DNA replicatio...	17	8	1.75	0.00013
10	GO:0051301	cell division	327	71	33.61	0.00017
11	GO:0006880	intracellular sequestering of iron ion	7	5	0.72	0.00020
12	GO:0009409	response to cold	281	49	28.88	0.00021
13	GO:0000413	protein peptidyl–prolyl isomerization	44	13	4.52	0.00033

c

***A. thaliana* cluster3**

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0000727	double–strand break repair via break–ind...	9	5	0.29	3.9e–06
2	GO:0006267	pre–replicative complex assembly involve...	8	4	0.26	6.8e–05
3	GO:0046686	response to cadmium ion	256	21	8.27	8.6e–05

d

***A. arenosa* cluster1**

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0055114	oxidation–reduction process	960	153	81.94	3.8e–13
2	GO:0098869	cellular oxidant detoxification	89	23	7.60	1.0e–06
3	GO:0019253	reductive pentose–phosphate cycle	15	8	1.28	1.0e–05
4	GO:0009854	oxidative photosynthetic carbon pathway	7	5	0.60	8.1e–05
5	GO:0051186	cofactor metabolic process	437	74	37.30	0.00010
6	GO:0009247	glycolipid biosynthetic process	43	12	3.67	0.00018
7	GO:0019682	glyceraldehyde–3–phosphate metabolic pro...	34	10	2.90	0.00038
8	GO:0009407	toxin catabolic process	23	8	1.96	0.00042
9	GO:0050665	hydrogen peroxide biosynthetic process	10	5	0.85	0.00078
10	GO:0006749	glutathione metabolic process	31	9	2.65	0.00083
11	GO:0015979	photosynthesis	181	41	15.45	0.00096

e

***A. arenosa* cluster2**

	GO.ID	Term	Annotated	Significant	Expected	classic
1	GO:0031145	anaphase–promoting complex–dependent cat...	8	7	1.68	0.00012
2	GO:0006397	mRNA processing	311	119	65.22	0.00018
3	GO:0000245	spliceosomal complex assembly	19	11	3.98	0.00033
4	GO:0006606	protein import into nucleus	46	22	9.65	0.00060
5	GO:0051301	cell division	327	86	68.57	0.00087

Supplementary Table 2. List of overrepresented gene ontologies on the Fig. 5e

762 # Materials & Methods

763 ## PacBio sequencing of *A. suecica*

764 We used genomic DNA from whole rosettes of one *A. suecica* ("ASS3") accession to generate
765 PacBio sequencing data. DNA was extracted using a modified PacBio protocol for preparing
766 *Arabidopsis* genomic DNA for size-selected ~20kb SMRTbell libraries. Briefly, whole genomic
767 DNA was extracted from 32g of 3-4 week old plants, grown at 16°C and subjected to a 2-day
768 dark treatment. This generated 23 micrograms of purified genomic DNA with a fragment length
769 of >40Kb for *A. suecica*. We assessed DNA quality with a Qubit fluorometer and a Nanodrop
770 analysis, and ran the DNA on a gel to visualize fragmentation. Genomic libraries and single-
771 molecule real-time (SMRT) sequence data were generated at the Functional Genomics Center
772 Zurich (FGCZ), in Switzerland. The Pacbio RSII instrument was used with P6/C4 chemistry
773 and an average movie length of 6 hours. A total of 12 SMRT cells were processed generating
774 16.3Gb of DNA bases with an N50 read length of 20 Kbp and median read length of 14 Kbp.
775 Using the same genomic library, an additional 3.3 Gbp of data was generated by a Pacbio
776 Sequel instrument at the Vienna Biocenter Core Facilities (VBCF), in Austria, with a median
777 read length of 10Kbp.

778 ## *A. suecica* genome assembly

779 To generate the *A. suecica* assembly we first used FALCON[65] (version 0.3.0) with a length
780 cutoff for seed reads set to 1 Kb in size. The assembly produced 828 contigs with an N50 of
781 5.81 Mb and a total assembly size of 271 Mb. Additionally, we generated a Canu[66] (v.1.3.0)
782 assembly using default settings, which resulted in 260 contigs with an N50 of 6.65 Mb and a
783 total assembly size of 267 Mb. Then we merged the two assemblies using the software
784 quickmerge[67]. The resulting merged assembly consisted of 929 contigs with an N50 of 9.02
785 Mb and a total draft assembly size of 276 Mb. We polished the assembly using Arrow[68]
786 (smrtlink release 5.0.0.6792) and Pilon (version 1.22). For Pilon[69], 100bp (with PCR duplicates
787 removed), and a second PCR-free 250bp, Illumina paired end reads were used that had been
788 generated from the reference *A. suecica* accession "ASS3".

789 ## Pacbio sequencing of *A. arenosa*

790 A natural Swedish autotetraploid *A. arenosa* accession "Aa4" was inbred in a lab for two
791 generations in order to reduce heterozygosity. We extracted 64 g of whole genomic DNA from
792 three week old seedlings in the same way as described for *A. suecica* (above), generating 50
793 µg of purified genomic DNA with a fragment sizes longer than 40 Kb in length. The *A. arenosa*
794 genomic libraries and SMRT sequence data were generated at the Vienna Biocenter Core
795 Facilities (VBCF), in Austria. A Pacbio Sequel instrument was used to generate a total of 22
796 Gbp of data from five SMRT cells, with an N50 of 13 Kbp and median read length 10 Kbp. In
797 addition, two runs of Oxford Nanopore sequencing were carried out at the VBCF producing
798 750 Mbp in 180,000 reads (median 5 Kbp and 2.6 Kbp; N50 8.7 and 6.7 Kbp, respectively).

Assembly of autotetraploid *A. arenosa*

We assembled a draft contig assembly for the autotetraploid *A. arenosa* accession "Aa4" using FALCON (version 0.3.0) as for *A. suecica*. The assembly produced 3,629 contigs with an N50 of 331 Kb, 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 FACs (see Supplementary Figure 2) probably because of the high levels of heterozygosity in *A. arenosa*. The resulting assembly was polished as described for *A. suecica*.

HiC tissue fixation and library preparation

To generate physical scaffolds for the *A. suecica* assembly we generated proximity-ligation HiC sequencing data. We collected approximately 0.5 gram of tissue from 3-week old seedlings of the same reference *A. suecica* accession. Freshly collected plant tissue was fixed in 1% formaldehyde. Cross-linking was stopped by the addition of 0.15 M Glycine. The fixed tissue was ground to a powder in liquid nitrogen and suspended in 10 ml of nuclei isolation buffer. Nuclei was digested by adding 50 U DpnII and the digested chromatin was blunt-ended by incubation with 25 µL of 0.4 mM biotin-14-dCTP and 40 U of Klenow enzyme, as described in [ref]. 20 U of T4 DNA ligase was then added to start proximity ligation. The extracted DNA was sheared by sonication with a Covaris S220 to produce 250-500bp fragments. This was followed by size fractionation using AMPure XP beads. Biotin was then removed from unligated ends. DNA fragments were blunt-end repaired and adaptors were ligated to the DNA products following the NEBNext Ultra II RNA Library Prep Kit for Illumina.

To analyse structural rearrangements we collected tissue for 1 other natural *A. suecica* “AS530”, 1 *A. thaliana* accession ”6978”, 1 *A. arenosa* “Aa6” and 1 synthetic *A. suecica* (F3). Each sample had two replicates. We collected tissue and prepared libraries in the same manner as described above. 125bp paired-end Illumina reads were mapped using HiCUP[70] (version 0.6.1).

Reference-guided scaffolding of the *A. suecica* genome with *LACHESIS*

We sequenced 207 million pairs of 125bp paired-end Illumina reads from the HiC library of the reference accession "ASS3". We mapped reads using HiCUP (version 0.6.1) to the draft *A. suecica* contig assembly. This resulted in ~137 million read pairs with a unique alignment.

Setting an assembly threshold of >= 1 Kb in size, contigs of the draft *A. suecica* assembly were first assigned to the *A. thaliana* or *A. arenosa* subgenome. To do this, we used nucmer from the software MUMmer[71] (version 3.23) to perform whole-genome alignments. We aligned the draft *A. suecica* assembly to the *A. thaliana* TAIR10 reference and to our *A. arenosa* draft contig assembly, simultaneously. We used the MUMer command dnadiff to produce 1-to-1 alignments. As the subgenomes are only ~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 based on percentage identity were examined manually, and the length of the alignment was used to determine subgenome assignment.

Finally, we used the software LACHESIS[72] (version 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

841 the contigs (we used *A. lyrata* here instead of our draft *A. arenosa* contig assembly, as *A.*
842 *lyrata* is a chromosome-level assembly). This produced a 13-scaffold chromosome-level
843 assembly for *A. suecica*.

844 Construction of the *A. suecica* genetic map

845 We crossed natural *A. suecica* accession "AS150" with the reference accession "ASS3". The
846 cross was uni-directional with "AS150" as the maternal and "ASS3" as the paternal plant. F1
847 plants were grown, and F2 seeds were collected, from which we grew and collected 192 F2
848 plants. We multiplexed the samples on 96 well plates using 75bp paired end reads and
849 generated data of 1-2x coverage per sample. Samples were mapped to the repeat-masked
850 scaffolds of the reference *A. suecica* genome using BWA-MEM[73] (version 0.7.15). Samtools[74]
851 (version 0.1.19) was used to filter reads for proper pairs and a minimum mapping quality of 5
852 (-F 256 -f 3 -q 5). We called variants directly from samtools mpileup output on the sequenced
853 F2 individuals at known biallelic sites between the two accessions used to generate the cross
854 (a total of 590,537 SNPs). We required sites to have non-zero coverage in a minimum of 20
855 individuals and filtered SNPs to have frequency between 0.45-0.55 in our F2 population (as
856 the expectation is 50:50),. We removed F2 individuals that did not have genotype calls for
857 more than 90% of the data. This resulted in 183 individuals with genotype calls for 334,257
858 SNPs.
859 Since sequencing coverage for the F2s was low this meant we had a low probability of
860 calling heterozygous SNPs, and a higher probability of calling a SNP as homozygous.
861 Therefore, we applied a Hidden Markov Model implemented in R package HMM[75] to classify
862 SNPs as homozygous or heterozygous for each of our F2 lines. We then divided the genome
863 into 500Kb non-overlapping windows, and classified each window as homozygous (here 0 or
864 1, for the reference or alternate SNP) or heterozygous (here 0.5). If the frequency of 1, 0 or
865 0.5 represented more than 50% of the SNPs in a given window, and exceeded missing calls
866 (NA), the window was designated as 1, 0 or 0.5 (otherwise it was NA). This was done per
867 chromosome and the resulting file for each chromosome and their markers were processed in
868 the R package qtl[76], in order to generate a genetic map. Markers genotyped in less than 100
869 F2s were excluded from the analysis. Linkage groups were assigned with a minimum LOD
870 score of 8 and a maximum recombination fraction of 0.35. Each chromosome was assigned
871 to one linkage group. We defined the final marker order by the best LOD score and the lowest
872 number of crossover events.
873 Notably, the assistance of a genetic map corrected the erroneous placement of a contig
874 at the beginning of chromosome 1 of the *A. arenosa* subgenome. The misplaced contig was
875 relocated from chromosome 1 to the pericentromeric region of chromosome 2 of the *A.*
876 *arenosa* subgenome in *A. suecica*. This error was a result of a mis-assembly of chromosome
877 1 in the *A. lyrata* reference, as was previously pointed out [27]. Also of note, chromosome 2 of
878 the *A. thaliana* subgenome of *A. suecica* was previously shown to be largely devoid of
879 intraspecific variation, thus we had sparse marker information for this chromosome in the
880 genetic map. Therefore, this chromosome-scale scaffold was largely assembled by the
881 manual inspection of 3D-proximity information based on our HiC sequencing and reviewing
882 contig order using the software Juicebox[77].

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 AUGUSTUS[78] on the set of conserved single copy genes using BUSCO[79] 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 TAIR10 *A. thaliana* assembly to the *A. thaliana* subgenome of *A. suecica*, while the peptides from *A. lyrata* from the second version of *A. lyrata* annotation[80] (Alyrata_384_v2.1) were aligned to the *A. arenosa* subgenome of *A. suecica* using GenomeThreader[81] (1.7.0). We mapped the RNAseq reads from the reference accession of *A. suecica* (ASS3) from the rosettes and flower buds tissues (see above) to the reference genome using tophat[82] and generated intron hints from the split reads using bam2hints extension of AUGUSTUS. 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 Trinity[83] (2.6.6). Separately for each of the subgenomes, we filtered the assembled transcripts using tpm cutoff set to 1, collapsed similar transcripts using CD-HIT[84,85] with sequence identity set to 90 percent, and chose the longest open reading frame from the six-frame translation. We then aligned the proteins from *A. thaliana* and *A. arenosa* parts of *A. suecica* to the corresponding subgenomes using GenomeThreader (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* RNAseq, (2) homology hints from ancestral proteins and (3) hints from *A. suecica* proteins.

RepeatModeler[86] (version 1.0.11) was used in order to build a *de novo* TE consensus library for *A. suecica* and identify repetitive elements based on the genome sequence. Genome locations for the identified TE repeats were determined by using RepeatMasker[87] (version 4.0.7) and filtered for full length matches using a code described in Bailly-Bechet et. al[88]. Helitrons are the most abundant TE family in both subgenomes (Supplementary Fig. 7).

In order to compare differences between the manual curation of Tair10 and our de novo pipeline, we *de novo* annotated the long read reference genome of *Ler* [64] with RNA-seq from the same tissue and growth conditions[89].

Synthetic *A. suecica* lines

To generate synthetic *A. suecica* we crossed a natural tetraploid *A. thaliana* accession (6978 aka “Wa-1”) to a natural Swedish autotetraploid *A. arenosa* (“Aa4”) accession. Similar to the natural *A. suecica*, *A. thaliana* was the maternal and *A. arenosa* was the paternal plant in this cross. Crosses in the opposite direction were unsuccessful. We managed to obtain very few F1 hybrid plants, which after one round of selfing set higher levels of seed formation. The resulting synthetic line was able to self-fertilize. F2 seeds were descended from a common F1 and were similar to natural *A. suecica* in appearance. We further continued the synthetic line to F3 (selfed 3rd generation).

Synteny analysis

We performed all-against-all BLASTP search using CDS sequences for the reference *A. suecica* genome and the ancestral genomes, *A. thaliana* and *A. lyrata* (here the closest substitute reference genome for *A. arenosa*, with annotation). We used the SynMap tool[90] from the online CoGe portal[91]. We examined synteny using the default parameters for DAGChainer

926 (maximum distance between two matches = 20 genes; minimum number of aligned pairs = 5
927 genes).

928 Estimating copy number of rDNA repeats using short DNA 929 reads

930 To measure copy number of 45S rRNA repeats in our populations of different species, we
931 aligned short DNA reads to a single reference 45S consensus sequence of *A. thaliana*[92]. An
932 *A. arenosa* 45S rRNA consensus sequence was constructed by finding the best hit using
933 BLAST in our draft *A. arenosa* contig assembly. This hit matched position 1571-8232 bp of
934 the *A. thaliana* consensus sequence, was 6,647 bp in length and is 97% identical to the *A.*
935 *thaliana* 45s rRNA consensus sequence. The aligned regions of these two 45S rRNA
936 consensus sequences, determined by BLAST, were used in copy number estimates, to ensure
937 that the size of the sequences were equal. The relative increase in sequence coverage of
938 these loci, when compared to the mean coverage for the reference genome, was used to
939 estimate copy number.

940 Plant material for RNA sequencing

941 Transcriptomic data generated in this study included 15 accessions of *A. suecica*, 16
942 accessions of *A. thaliana*, 4 accessions of *A. arenosa* and 2 generations of an artificial *A.*
943 *suecica* line (the 2nd and 3rd selfed-generation). The sibling of a paternal *A. arenosa* parent
944 (Aa4) and the maternal tetraploid *A. thaliana* parent (6978 aka "Wa-1") of our artificial *A.*
945 *suecica* line were included as part of our samples (Supplementary Data 2). Each accession
946 was replicated 3 times. Seeds were stratified in the dark for 4 days at 4°C in 1 ml of sterilised
947 water. Seeds were then transferred to pots in a controlled growth chamber at 21°C. Humidity
948 was kept constant at 60%. Pots were thinned to 2-3 seedlings after 1 week. Pots were re-
949 randomized each week in their trays. Whole rosettes were collected when plants reached the
950 7-9 true-leaf stage of development. Samples were collected between 14:00-17:00h and flash-
951 frozen in liquid nitrogen.

952 RNA extraction and library preparation

953 For each accession, 2-3 whole rosettes in each pot were pooled and total RNA was extracted
954 using the ZR Plant RNA MiniPrepTM kit. We treated the samples with DNAse, and performed
955 purification of mRNA and polyA selection using the AMPure XP magnetic beads and the
956 Poly(A) RNA Selection Kit from Lexogen. RNA quality and degradation were assessed using
957 the RNA Fragment Analyzer (DNF-471 stranded sensitivity RNA analysis kit, 15nt).
958 Concentration of RNA per sample was measured using the Qubit fluorometer. Library
959 preparation was carried out following the NEBNext Ultra II RNA Library Prep Kit for Illumina.
960 Barcoded adaptors were ligated using NEBNext Multiplex Oligos for Illumina (Index Primers
961 Set 1 and 2). The libraries were PCR amplified for 7 cycles. 125bp paired-end sequencing
962 was carried out at the VBCF on Illumina (HiSeq 2500) using multiplexing.

RNA-seq mapping and gene expression analysis

We mapped 125bp paired-end reads to the *de novo* assembled *A. suecica* reference using STAR[93] (version 2.7), we filtered for primary and uniquely aligned reads using the parameters --outfilterMultimapNmax 1 --outSamprimaryFlag OneBestScore. We quantified reads mapped to genes using --quantMode GeneCounts.

In order to reduce signals that are the result of cross mapping between the subgenomes of *A. suecica* we used *A. thaliana* and *A. arenosa* as a control. For each gene in the *A. thaliana* subgenome we compared log fold change of gene counts in our *A. thaliana* population to those in our *A. arenosa* population. We filtered for genes with a $\log_2$(*A. thaliana*/*A. arenosa*) below 0. We applied the same filters for genes on the *A. arenosa* subgenome, here a $\log_2$(*A. arenosa*/*A. thaliana*) below 0. (see Supplementary Figure 13). This reduced the number of genes analyzed from 22,383 to 22,193 on the *A. thaliana* subgenome, and 22,665 to 21,886 on the *A. arenosa* subgenome

Expression analysis was then further restricted to unique homeologous gene pairs between the subgenomes of *A. suecica* and that are 1:1 (16,742 gene pairs). Gene counts were normalized for gene size by calculating Transcripts Per Million (TPM). The effective library sizes were calculated by computing a scaling factor based on the trimmed mean of M-values (TMM) in edgeR[94], separately for each subgenome. Lowly expressed genes were removed from the analysis by keeping genes that were expressed in at least five accessions in *A. thaliana* and *A. suecica*, at least three accessions of *A. arenosa* and at least one accession of synthetic *A. suecica*. 13,647 homeologous gene pairs satisfied our expression criteria. Since *A. suecica* is expressing both subgenomes, in order to correctly normalize the effective library size in *A. suecica* accessions, the effective library size was calculated as a mean of TPM counts in both subgenomes. The effective library size of *A. thaliana* accessions was calculated for TPM counts using the *A. thaliana* subgenome of the reference genome, as genes from this subgenome will be expressed in *A. thaliana*, and the effective library size of *A. arenosa* lines using the *A. arenosa* subgenome of the reference *A. suecica* genome. Gene counts were transformed to count per million (CPM) with a prior count of 1, and were $\log_2$-transformed. We used the mean of replicates per accession for downstream analyses.

To compare homeologous genes between the subgenomes in *A. suecica* we computed a log-fold change using $\log_2$(*A. arenosa* homeolog/*A. thaliana* homeolog). For tissue-specific genes we took genes that showed a log-fold change >=2 in expression between two tissues.

For comparing homologous genes between the (sub-)genomes of *A. suecica* and the ancestral species *A. thaliana* and *A. arenosa*, we performed a Wilcoxon test independently for each of the 13,647 homeologous gene-pairs. Using the normalised CPM values, we compared the relative expression level of a gene on the *A. thaliana* subgenome between our population of *A. thaliana* and *A. suecica*. We performed the same test on the *A. arenosa* subgenome comparing relative expression of a gene between our population of *A. arenosa* and *A. suecica*. We filtered for genes with an adjusted p-value below <0.05 (using FDR correction). This amounted to 3,854 and 4,136 DEGs for the *A. thaliana* and *A. arenosa* subgenomes, respectively.

1004

1005 Expression analysis of rRNA

1006 RNA reads were mapped in a similar manner as DNA reads for the analysis of rDNA copy
1007 number (above). Expression analysis was performed in a similar manner to protein coding
1008 genes, in edgeR. We defined the exclusive expression of a particular 45S rRNA gene by taking
1009 a cut-off of 15 for $\log_2$(CPM) as this was the maximum level of cross-mapping we observed
1010 for the ancestral species (see Supplementary Fig. 6).

1011 Expression analysis of transposable elements

1012 To analyse the expression of transposable elements between species, the annotated TE
1013 consensus sequences in *A. suecica* were aligned using BLAST all vs all. Highly similar TE
1014 sequences (more than 85% similar for more than 85% percent of the TE sequence length),
1015 were removed, leaving 813 TE families out of 1213. Filtered *A. suecica* TEs were aligned to
1016 annotated *A. thaliana* (TAIR10) and *A. arenosa* (the PacBio contig assembly presented in
1017 this study) TE sequences to assign each family to an ancestral species using BLAST. 208
1018 TE families were assigned to the *A. thaliana* parent and 171 TE families were assigned to
1019 the *A. arenosa* parent.
1020 RNA reads were mapped to TE sequences using a similar approach as for gene
1021 expression analysis using edgeR. TEs that showed expression using a cut-off of $\log_2$CPM >
1022 2 were kept. 121 *A. thaliana* TE sequences and 93 *A. arenosa* TE sequences passed this
1023 threshold. We took the mean of replicates per accession for further downstream analyses.

1024 Gene ontology (GO) enrichment analysis

1025 We used the R package TopGO[95] to conduct gene ontology enrichment analysis. We used
1026 the “weight01” algorithm when running TopGO which accounts for the hierarchical structure
1027 of GO terms and thus implicitly corrects for multiple testing. GO annotations were based on
1028 the *A. thaliana* ortholog of *A. suecica* genes. Gene annotations for *A. thaliana* were obtained
1029 using the R package biomaRt[96] from Ensembl ‘biomaRt::useMart(biomart = "plants_mart",
1030 dataset = "athaliana_eg_gene", host = 'plants.ensembl.org').

1031 Genome sizes measurements

1032 We measured genome size for the reference *A. suecica* accession “ASS3” and the *A. arenosa*
1033 accession used for PacBio “Aa4”, using *Solanum lycopersicum* cv. Stupicke (2C = 1.96 pg
1034 DNA) as the standard. The reference *A. lyrata* accession “MN47” and the *A. thaliana*
1035 accession “CVI” were used as additional controls. Each sample had 2 replicates.
1036 In brief, the leaves from three week old fresh tissue were chopped using a razor blade in 500
1037 µl of UV Precise P extraction buffer + 10 µl mercaptoethanol per ml (kit PARTEC CyStain PI
1038 Absolute P no. 05- 5022) to isolate nuclei. Instead of the Partec UV Precise P staining buffer,
1039 however, 1 ml of a 5 mg DAPI solution was used, as DAPI provides DNA content histograms
1040 with high resolution. The suspension was then passed through a 30 µm filter (Partec CellTrics
1041 no. 04-0042-2316) and incubated for 15 minutes on ice before FACs.

1042 Genome size was measured using flow cytometry and a FACS Aria III sorter with near UV
1043 375nm laser for DAPI. Debris was excluded by selecting peaks when plotting DAPI-W against
1044 DAPI-A for 20,000 events.
1045 The data were analyzed using the flowCore[97] package in R. Genome size was estimated
1046 by comparing the mean G1 of the standard *Solanum lycopersicum* to that of each sample to
1047 calculate the 2C DNA content of that sample using the equation:

1048 $$Sample\ 2C\ DNA\ content\ =\ [(sample\ G1\ peak\ mean)/(standard\ G1\ peak\ mean)]$$
1049 $$*\ standard\ 2C\ DNA\ content$$

1050 We also measured genome size for the reference *A. suecica* accession "ASS3" using the
1051 software jellyfish[98] and findGSE[99] using kmers (21mers). The genome size estimated was
1052 312Mb, compared to the 305Mb estimated using FACs (see Supplementary Fig 1).

1053 Mapping of TE insertions

1054 We used PopoolationTE2[40] (version v1.10.04) to identify TE insertions. The advantage of this
1055 TE-calling software to others is that it avoids a reference bias by treating all TEs as *de-novo*
1056 insertions. Briefly, it works by using discordant read pairs to calculate the location and
1057 abundance of a TE in the genome for an accession of interest.
1058 We mapped 100 bp Illumina DNA reads from [16,26,100], in addition to our newly generated
1059 synthetic *A. suecica* using BWA MEM[73] (version 0.7.15) to a repeat-masked version of the *A.*
1060 *suecica* reference genome, concatenated with our annotated repeat sequences (see 'Genome
1061 annotation'), as this is the data format required by PopoolationTE2. Reads were given an
1062 increased penalty of 15 for being unpaired. Reads were de-duplicated using Samtools[74] rmdup
1063 (version 1.9). The resulting bam files were then provided to PopoolationTE2 to identify TE
1064 insertions in the genome of each of our *A. suecica*, *A. thaliana* and *A. arenosa* accessions.
1065 We used a mapping quality of 10 for the read in the discordant read pair mapping to the
1066 genome. We used the 'separate' mode in the 'identify TE signatures' step and a '--min-
1067 distance -200 --max-distance 500' in the 'pairupsignatues' step of the pipeline. TE counts
1068 within each accession were merged if they fell within 400 bp of each other and if they mapped
1069 to the same TE sequence. All TE counts (i.e. the processed TE counts for each accession)
1070 were then combined to produce a population-wide count estimate. Population wide TE
1071 insertions were merged if they mapped to the same TE sequence and fell within 400 bp of
1072 each other. Coverage of each TE insertion in the population was also calculated for each
1073 accession. The final file was a list TE insertions present in the population and the presence or
1074 absence (or "NA" if there was no coverage to support the presence or absence of a TE
1075 insertion) in each accession analyzed (Supplementary Data 1a,b). To avoid insertions that
1076 could be potential mismapping, we excluded the region between 5 and 10Mb on chromosome
1077 2 of the *A. thaliana* subgenome as this area shows synteny with an unplaced contig.

1078 Assigning ancestry to TE sequences

1079 In order to examine TE consensus sequences that have mobilized between the subgenomes
1080 of *A. suecica*, we first examined which of our TE consensus sequences (N=1152) have at
1081 least the potential to mobilize (i.e. have full length TE copies in the genome of *A. suecica)*.
1082 We filtered for TE consensus sequences that had TE copies in the genome of *A. suecica*
1083 that are more than 80% similar in identity for more than 80% of the consensus sequence
1084 length (N=936). Of these, 188 consensus sequences were private to the *A. thaliana*

1085 subgenome, 460 were private to the *A. arenosa* subgenome, and 288 TE consensus
1086 sequences were present in both subgenomes of *A. suecica*. To determine if TEs have
1087 jumped from the *A. thaliana* subgenome to the *A. arenosa* subgenome and vice versa we
1088 next needed to assign ancestry to these 288 TE consensus sequences. To do this we used
1089 BLAST to search for these consensus sequences in the ancestral genomes of *A. suecica*,
1090 using the TAIR10 *A. thaliana* reference and our *A. arenosa* PacBio contig assembly. Using
1091 the same 80%-80% rule we assigned 55 TEs to *A. arenosa* and 15 TEs to *A. thaliana*
1092 ancestry.

1093 Read mapping and SNP calling

1094 To call biallelic SNPs we mapped reads to the *A. suecica* reference genome using the same
1095 filtering parameters described in "Mapping of TE insertions". Biallelic SNPs were called using
1096 HaplotypeCaller from GATK[101] (version 3.8) using default quality thresholds. SNPs were
1097 annotated using SnpEff[102]. Biallelic SNPs on the *A. thaliana* subgenome were polarized
1098 using 38 diploid *A. lyrata* lines[26] and biallelic SNPS on the *A. arenosa* subgenome were
1099 polarized using 30 *A. thaliana* accessions[100] closely related to *A. suecica*[16].

1100 Chromosome preparation and FISH

1101 Whole inflorescences of *A. arenosa*, *A. suecica* and *A. thaliana* were fixed in freshly
1102 prepared ethanol:acetic acid fixative (3:1) overnight, transferred into 70% ethanol and stored
1103 at -20°C until use. Selected inflorescences were rinsed in distilled water and citrate buffer
1104 (10 mM sodium citrate, pH 4.8), and digested by a 0.3% mix of pectolytic enzymes
1105 (cellulase, cytohelicase, pectolyase; all from Sigma-Aldrich) in citrate buffer for c. 3 hrs.
1106 Mitotic chromosome spreads were prepared from pistils as previously described[103] by
1107 Mandáková and Lysak and suitable slides pretreated by RNase (100 µg/ml, AppliChem) and
1108 pepsin (0.1 mg/ml, Sigma-Aldrich).
1109 For identification of *A. thaliana* and *A. arenosa* subgenomes in the allotetraploid
1110 genome of *A. suecica*, FISH probes were made from plasmids pARR20–1 or pAaCEN
1111 containing 180 bp of *A. thaliana* (pAL; Vongs et al. 1993) or ~250 bp of *A. arenosa* (pAa;
1112 Kamm et al. 1995) pericentromeric repeats, respectively. The *A. thaliana* BAC clone T15P10
1113 (AF167571) bearing 45S rRNA gene repeats was used for in situ localization of NORs.
1114 Individual probes were labeled with biotin-dUTP, digoxigenin-dUTP and Cy3-dUTP by nick
1115 translation, pooled, precipitated, and resuspended in 20 µl of hybridization mixture [50%
1116 formamide and 10% dextran sulfate in 2× saline sodium citrate (2× SSC)] per slide as
1117 previously described[96].
1118 Probes and chromosomes were denatured together on a hot plate at 80°C for 2 min
1119 and incubated in a moist chamber at 37°C overnight. Post hybridization washing was
1120 performed in 20% formamide in 2× SSC at 42°C. Fluorescent detection was as follows:
1121 biotin-dUTP was detected by avidin–Texas Red (Vector Laboratories) and amplified by goat
1122 anti-avidin–biotin (Vector Laboratories) and avidin–Texas Red; digoxigenin-dUTP was
1123 detected by mouse anti-digoxigenin (Jackson ImmunoResearch) and goat anti-mouse Alexa
1124 Fluor 488 (Molecular Probes). Chromosomes were counterstained with DAPI (4',6-
1125 diamidino-2-phenylindole; 2 µg/ml) in Vectashield (Vector Laboratories). Fluorescent signals
1126 were analyzed and photographed using a Zeiss Axioimager epifluorescence microscope and
1127 a CoolCube camera (MetaSystems). Images were acquired separately for the four

fluorochromes using appropriate excitation and emission filters (AHF Analysentechnik). The monochromatic images were pseudo colored 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[104]. To test for enrichment of a gene set (for example the genes in *A. thaliana* cluster 2 on Fig. 3) for target genes of a particular transcription factor, we performed a hyper-geometric test in R. As a background we used the total 13,647 genes used in our gene expression analysis. We then performed FDR correction for multiple testing to calculate an accurate p-value of the enrichment.

Data Availability

The raw sequencing for genome assembly ####. Genome assembly and annotation ####. The raw RNA-seq reads SRA ####. TE presence/absence calls for *A. suecica* and ancestral species can be found in the Supplementary Data 1a,b.

Acknowledgments

This work was supported, in part, by DFG SPP 1529 to M.N. and Detlef Weigel. T.M. and M.A.L. were supported by the Czech Science Foundation (grant no. 19-03442S) and the CEITEC 2020 project (grant no. LQ1601). P.Y.N. acknowledges postdoctoral fellowship of the Research Foundation–Flanders (12S9618N). We thank the Next Generation Sequencing Unit of the Vienna Biocenter Core Facilities (VBCF) for assistance. We thank Svante Holm and Torbjörn Säll for material collections and helpful discussions. We also thank Yves Van de Peer for providing useful feedback on the manuscript.

References

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

2 Soltis PS, Soltis DE. Ancient WGD events as drivers of key innovations in angiosperms. *Curr Opin Plant Biol* 2016; **30**: 159–165.

3 Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 2005; **3**: e314.

4 Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ *et al.* Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc Natl Acad Sci U S A* 2018; **115**: 4713–4718.

1163 5 Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM *et*
1164 *al.* Genomic diversifications of five Gossypium allopolyploid species and their
1165 impact on cotton improvement. *Nat Genet* 2020; **52**: 525–533.

1166 6 Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR *et al.*
1167 Origin and evolution of the octoploid strawberry genome. *Nat Genet* 2019; **51**:
1168 541–547.

1169 7 Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L
1170 *et al.* The transcriptional landscape of polyploid wheat. *Science* 2018; **361**.
1171 doi:10.1126/science.aar6089.

1172 8 Zhuang W, Chen H, Yang M, Wang J, Pandey MK, Zhang C *et al.* The genome
1173 of cultivated peanut provides insight into legume karyotypes, polyploid evolution
1174 and crop domestication. Nature Genetics. 2019; **51**: 865–876.

1175 9 Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G *et al.* The
1176 genome sequence of segmental allotetraploid peanut Arachis hypogaea. *Nat*
1177 *Genet* 2019; **51**: 877–884.

1178 10 Kasianov AS, Klepikova AV, Kulakovskiy IV, Gerasimov ES, Fedotova AV,
1179 Besedina EG *et al.* High-quality genome assembly of Capsella bursa-pastoris
1180 reveals asymmetry of regulatory elements at early stages of polyploid genome
1181 evolution. *Plant J* 2017; **91**: 278–291.

1182 11 Kryvokhyzha D, Milesi P, Duan T, Orsucci M, Wright SI, Glémin S *et al.* Towards
1183 the new normal: Transcriptomic convergence and genomic legacy of the two
1184 subgenomes of an allopolyploid weed (Capsella bursa-pastoris). *PLoS Genet*
1185 2019; **15**: e1008131.

1186 12 Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB *et al.* Hybrid
1187 origins and the earliest stages of diploidization in the highly successful recent
1188 polyploid Capsella bursa-pastoris. *Proc Natl Acad Sci U S A* 2015; **112**: 2806–
1189 2811.

1190 13 Griffiths AG, Moraga R, Tausen M, Gupta V, Bilton TP, Campbell MA *et al.*
1191 Breaking Free: The Genomics of Allopolyploidy-Facilitated Niche Expansion in
1192 White Clover. *Plant Cell* 2019; **31**: 1466–1487.

1193 14 Gordon SP, Contreras-Moreira B, Levy JJ, Djamei A, Czedik-Eysenberg A,
1194 Tartaglio VS *et al.* Gradual polyploid genome evolution revealed by pan-
1195 genomic analysis of Brachypodium hybridum and its diploid progenitors. *Nat*
1196 *Commun* 2020; **11**: 3670.

1197 15 Catalán P, López-Álvarez D, Bellosta C, Villar L. Updated taxonomic
1198 descriptions, iconography, and habitat preferences of Brachypodium distachyon,
1199 B. stacei , and B. hybridum (Poaceae). *An Jard Bot Madr* 2016; **73**: 028.

1200 16 Novikova PY, Tsuchimatsu T, Simon S, Nizhynska V, Voronin V, Burns R *et al.*
1201 Genome Sequencing Reveals the Origin of the Allotetraploid Arabidopsis
1202 suecica. *Mol Biol Evol* 2017; **34**: 957–968.

1203 17 Fowler NL, Levin DA. Ecological Constraints on the Establishment of a Novel
1204 Polyploid in Competition with Its Diploid Progenitor. *Am Nat* 1984; **124**: 703–
1205 711.

1206 18 Bomblies K, Madlung A. Polyploidy in the Arabidopsis genus. *Chromosome Res*
1207 2014; **22**: 117–134.

1208 19 Hollister JD, Arnold BJ, Svedin E, Xue KS, Dilkes BP, Bomblies K. Genetic
1209 adaptation associated with genome-doubling in autotetraploid Arabidopsis
1210 arenosa. *PLoS Genet* 2012; **8**: e1003093.

1211 20 Bomblies K, Jones G, Franklin C, Zickler D, Kleckner N. The challenge of
1212 evolving stable polyploidy: could an increase in 'crossover interference distance'
1213 play a central role? *Chromosoma* 2016; **125**: 287–300.

1214 21 Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C *et al.* Impact of transposable
1215 elements on the organization and function of allopolyploid genomes. *New Phytol*
1216 2010; **186**: 37–45.

1217 22 McClintock B. The significance of responses of the genome to challenge.
1218 Science. 1984; **226**: 792–801.

1219 23 Hohmann N, Wolf EM, Lysak MA, Koch MA. A Time-Calibrated Road Map of
1220 Brassicaceae Species Radiation and Evolutionary History. *Plant Cell* 2015; **27**:
1221 2770–2784.

1222 24 O'Kane SL, Schaal BA, Al-Shehbaz IA. The Origins of Arabidopsis suecica
1223 (Brassicaceae) as Indicated by Nuclear rDNA Sequences. *Syst Bot* 1996; **21**:
1224 559–566.

1225 25 Jakobsson M, Hagenblad J, Tavaré S, Säll T, Halldén C, Lind-Halldén C *et al.* A
1226 unique recent origin of the allotetraploid species Arabidopsis suecica: Evidence
1227 from nuclear DNA markers. *Mol Biol Evol* 2006; **23**: 1217–1231.

1228 26 Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G *et al.*
1229 Sequencing of the genus Arabidopsis identifies a complex history of
1230 nonbifurcating speciation and abundant trans-specific polymorphism. *Nat Genet*
1231 2016; **48**: 1077–1082.

1232 27 Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L *et al.* The
1233 Capsella rubella genome and the genomic consequences of rapid mating
1234 system evolution. *Nat Genet* 2013; **45**: 831–835.

1235 28 Ozkan H, Levy AA, Feldman M. Allopolyploidy-Induced Rapid Genome
1236 Evolution in the Wheat (Aegilops–Triticum) Group. *Plant Cell* 2001; **13**: 1735–
1237 1747.

1238 29 Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. Sequence elimination
1239 and cytosine methylation are rapid and reproducible responses of the genome
1240 to wide hybridization and allopolyploidy in wheat. *Plant Cell* 2001; **13**: 1749–
1241 1759.

1242 30 Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X *et al.* Plant
1243 genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus
1244 oilseed genome. *Science* 2014; **345**: 950–953.

1245 31 Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, Doerge RW *et al.*
1246 Genomic changes in synthetic Arabidopsis polyploids. *Plant J* 2005; **41**: 221–
1247 230.

1248 32 Copenhaver GP, Pikaard CS. Two-dimensional RFLP analyses reveal
1249 megabase-sized clusters of rRNA gene variants in Arabidopsis thaliana,
1250 suggesting local spreading of variants as the mode for gene homogenization
1251 during concerted evolution. The Plant Journal. 1996; **9**: 273–282.

1252 33 Navashin M. Chromosome Alterations Caused by Hybridization and Their
1253 Bearing upon Certain General Genetic Problems. *Cytologia* 1934; **5**: 169–203.

1254 34 Tucker S, Vitins A, Pikaard CS. Nucleolar dominance and ribosomal RNA gene
1255 silencing. *Curr Opin Cell Biol* 2010; **22**: 351–356.

1256 35 Pontes O, Neves N, Silva M, Lewis MS, Madlung A, Comai L *et al.*
1257 Chromosomal locus rearrangements are a rapid response to formation of the
1258 allotetraploid Arabidopsis suecica genome. Proceedings of the National
1259 Academy of Sciences. 2004; **101**: 18240–18245.

1260 36 Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A *et al.* Massive
1261 genomic variation and strong selection in Arabidopsis thaliana lines from
1262 Sweden. *Nat Genet* 2013; **45**: 884–890.

1263 37 Rabanal FA, Mandáková T, Soto-Jiménez LM, Greenhalgh R, Parrott DL,
1264 Lutzmayer S *et al.* Epistatic and allelic interactions control expression of
1265 ribosomal RNA gene clusters in Arabidopsis thaliana. *Genome Biol* 2017; **18**:
1266 75.

1267 38 Steige KA, Slotte T. Genomic legacies of the progenitors and the evolutionary
1268 consequences of allopolyploidy. *Curr Opin Plant Biol* 2016; **30**: 88–93.

1269 39 Vicient CM, Casacuberta JM. Impact of transposable elements on polyploid
1270 plant genomes. *Ann Bot* 2017; **120**: 195–207.

1271 40 Kofler R, Gomez-Sanchez D, Schlotterer C. PoPoolationTE2: Comparative
1272 Population Genomics of Transposable Elements Using Pool-Seq. *Mol Biol Evol*
1273 2016; **33**: 2759–2764.

1274 41 Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA,
1275 Jeddeloh JA *et al.* The Arabidopsis thaliana mobilome and its impact at the
1276 species level. *Elife* 2016; **5**. doi:10.7554/eLife.15716.

1277 42 Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. Population
1278 scale mapping of transposable element diversity reveals links to gene regulation
1279 and epigenomic variation. *Elife* 2016; **5**. doi:10.7554/eLife.20777.

1280 43 Wolfe KH. Yesterday's polyploids and the mystery of diploidization. *Nat Rev*

1281 *Genet* 2001; **2**: 333–341.

1282 44 Conant GC, Birchler JA, Pires JC. Dosage, duplication, and diploidization:
1283 clarifying the interplay of multiple models for duplicate gene evolution over time.
1284 *Curr Opin Plant Biol* 2014; **19**: 91–98.

1285 45 Aköz G, Nordborg M. The Aquilegia genome reveals a hybrid origin of core
1286 eudicots. *Genome Biol* 2019; **20**: 256.

1287 46 Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE *et*
1288 *al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* 2011; **473**: 97–
1289 100.

1290 47 Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. Polyploidy and genome
1291 evolution in plants. *Curr Opin Genet Dev* 2015; **35**: 119–125.

1292 48 Thomas BC, Pedersen B, Freeling M. Following tetraploidy in an Arabidopsis
1293 ancestor, genes were removed preferentially from one homeolog leaving
1294 clusters enriched in dose-sensitive genes. *Genome Res* 2006; **16**: 934–946.

1295 49 Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes
1296 by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad*
1297 *Sci U S A* 2011; **108**: 4069–4074.

1298 50 Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K *et al.* Biased gene fractionation and
1299 dominant gene expression among the subgenomes of Brassica rapa. *PLoS One*
1300 2012; **7**: e36442.

1301 51 Renny-Byfield S, Gong L, Gallagher JP, Wendel JF. Persistence of subgenomes
1302 in paleopolyploid cotton after 60 my of evolution. *Mol Biol Evol* 2015; **32**: 1063–
1303 1071.

1304 52 Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. Two
1305 evolutionarily distinct classes of paleopolyploidy. *Mol Biol Evol* 2014; **31**: 448–
1306 454.

1307 53 Li Q, Qiao X, Yin H, Zhou Y, Dong H, Qi K *et al.* Unbiased subgenome evolution
1308 following a recent whole-genome duplication in pear (Pyrus bretschneideri
1309 Rehd.). *Hortic Res* 2019; **6**: 34.

1310 54 Bird KA, Niederhuth C, Ou S, Gehan M, Chris Pires J, Xiong Z *et al.* Replaying
1311 the evolutionary tape to investigate subgenome dominance in allopolyploid
1312 Brassica napus. doi:10.1101/814491.

1313 55 Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y *et al.*
1314 Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a
1315 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. *Plant Cell*
1316 2017; **29**: 2150–2167.

1317 56 Alger EI, Edger PP. One subgenome to rule them all: underlying mechanisms of
1318 subgenome dominance. *Curr Opin Plant Biol* 2020; **54**: 108–113.

1319 57 Carlson KD, Fernandez-Pozo N, Bombarely A, Pisupati R, Mueller LA, Madlung
1320 A. Natural variation in stress response gene activity in the allopolyploid
1321 Arabidopsis suecica. *BMC Genomics* 2017; **18**: 653.

1322 58 Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV. Homoeolog-specific
1323 retention and use in allotetraploid Arabidopsis suecica depends on parent of
1324 origin and network partners. *Genome Biol* 2010; **11**: R125.

1325 59 Yant L, Hollister JD, Wright KM, Arnold BJ, Higgins JD, Franklin FC *et al.*
1326 Meiotic adaptation to genome duplication in Arabidopsis arenosa. *Curr Biol*
1327 2013; **23**: 2151–2156.

1328 60 Morgan C, Zhang H, Henry CE, Franklin FCH, Bomblies K. Derived alleles of
1329 two axis proteins affect meiotic traits in autotetraploid Arabidopsis arenosa. *Proc*
1330 *Natl Acad Sci U S A* 2020; **117**: 8980–8988.

1331 61 Haga N, Kobayashi K, Suzuki T, Maeo K, Kubo M, Ohtani M *et al.* Mutations in
1332 MYB3R1 and MYB3R4 cause pleiotropic developmental defects and preferential
1333 down-regulation of multiple G2/M-specific genes in Arabidopsis. *Plant Physiol*
1334 2011; **157**: 706–717.

1335 62 Preite V, Sailer C, Syllwasschy L, Bray S, Ahmadi H, Krämer U *et al.*
1336 Convergent evolution in Arabidopsis halleri and Arabidopsis arenosa on
1337 calamine metalliferous soils. *Philos Trans R Soc Lond B Biol Sci* 2019; **374**:
1338 20180243.

1339 63 Darwin C. The origin of species by means of natural selection : or The
1340 preservation of favored races in the struggle for life / by Charles Darwin. 1872.
1341 doi:10.5962/bhl.title.2106.

1342 64 Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B *et al.*
1343 Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of
1344 translocation and inversion polymorphisms. *Proc Natl Acad Sci U S A* 2016;
1345 **113**: E4052–60.

1346 65 Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A *et al.*
1347 Phased diploid genome assembly with single-molecule real-time sequencing.
1348 *Nat Methods* 2016; **13**: 1050–1054.

1349 66 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu:
1350 scalable and accurate long-read assembly via adaptive k-mer weighting and
1351 repeat separation. *Genome Res* 2017; **27**: 722–736.

1352 67 Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and
1353 accurate de novo assembly of metazoan genomes with modest long read
1354 coverage. *Nucleic Acids Res* 2016; **44**: e147.

1355 68 Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C *et al.*
1356 Nonhybrid, finished microbial genome assemblies from long-read SMRT
1357 sequencing data. Nature Methods. 2013; **10**: 563–569.

1358 69 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S *et al.* Pilon:

1359 an integrated tool for comprehensive microbial variant detection and genome
1360 assembly improvement. *PLoS One* 2014; **9**: e112963.

1361 70 Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P *et*
1362 *al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* 2015; **4**:
1363 1310.

1364 71 Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A.
1365 MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol*
1366 2018; **14**: e1005944.

1367 72 Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J.
1368 Chromosome-scale scaffolding of de novo genome assemblies based on
1369 chromatin interactions. *Nat Biotechnol* 2013; **31**: 1119–1125.

1370 73 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
1371 transform. *Bioinformatics* 2009; **25**: 1754–1760.

1372 74 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* The
1373 Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–
1374 2079.

1375 75 Himmelmann L. HMM: Hidden Markov Models. *R package version* 2010; **1**.

1376 76 Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental
1377 crosses. *Bioinformatics* 2003; **19**: 889–890.

1378 77 Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES *et al.*
1379 Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C
1380 Experiments. *Cell Syst* 2016; **3**: 95–98.

1381 78 Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in
1382 eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 2005; **33**:
1383 W465–7.

1384 79 Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and
1385 Annotation Completeness. In: Kollmar M (ed). *Gene Prediction: Methods and*
1386 *Protocols*. Springer New York: New York, NY, 2019, pp 227–245.

1387 80 Rawat V, Abdelsamad A, Pietzenuk B, Seymour DK, Koenig D, Weigel D *et al.*
1388 Improving the Annotation of Arabidopsis lyrata Using RNA-Seq Data. *PLoS One*
1389 2015; **10**: e0137391.

1390 81 Gremme G, Brendel V, Sparks ME, Kurtz S. Engineering a software tool for
1391 gene structure prediction in higher organisms. *Information and Software*
1392 *Technology* 2005; **47**: 965–978.

1393 82 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2:
1394 accurate alignment of transcriptomes in the presence of insertions, deletions
1395 and gene fusions. *Genome Biol* 2013; **14**: R36.

1396 83 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I *et al.* Full-

1397 length transcriptome assembly from RNA-Seq data without a reference genome.
1398 *Nat Biotechnol* 2011; **29**: 644–652.

1399 84 Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-
1400 generation sequencing data. *Bioinformatics* 2012; **28**: 3150–3152.

1401 85 Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of
1402 protein or nucleotide sequences. *Bioinformatics* 2006; **22**: 1658–1659.

1403 86 Smit AFA, Hubley R. RepeatModeler Open-1.0 http://www.repeatmasker.org.
1404 2008-2015.

1405 87 Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. . 2013-2015.

1406 88 Bailly-Bechet M, Haudry A, Lerat E. 'One code to find them all': a perl tool to
1407 conveniently parse RepeatMasker output files. *Mob DNA* 2014; **5**: 13.

1408 89 Lei L, Steffen JG, Osborne EJ, Toomajian C. Plant organ evolution revealed by
1409 phylotranscriptomics in Arabidopsis thaliana. *Sci Rep* 2017; **7**: 7567.

1410 90 Lyons E, Pedersen B, Kane J, Freeling M. The value of nonmodel genomes and
1411 an example using SynMap within CoGe to dissect the hexaploidy that predates
1412 the rosids. *Trop Plant Biol* 2008; **1**: 181–190.

1413 91 Lyons E, Freeling M. How to usefully compare homologous plant genes and
1414 chromosomes as DNA sequences. *Plant J* 2008; **53**: 661–673.

1415 92 Rabanal FA, Nizhynska V, Mandáková T, Novikova PY, Lysak MA, Mott R *et al.*
1416 Unstable Inheritance of 45S rRNA Genes in Arabidopsis thaliana. *G3* 2017; **7**:
1417 1201–1209.

1418 93 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S *et al.* STAR:
1419 ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29**: 15–21.

1420 94 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for
1421 differential expression analysis of digital gene expression data. *Bioinformatics*
1422 2010; **26**: 139–140.

1423 95 Alexa A, Rahnenfuhrer J. topGO: enrichment analysis for gene ontology. *R*
1424 *package version* 2010; **2**: 2010.

1425 96 Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the
1426 integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat*
1427 *Protoc* 2009; **4**: 1184–1191.

1428 97 Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, Sarkar D *et al.* flowCore:
1429 a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*
1430 2009; **10**: 106.

1431 98 Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting
1432 of occurrences of k-mers. *Bioinformatics* 2011; **27**: 764–770.

1433 99 Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size
1434 variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics*
1435 2018; **34**: 550–557.

1436 100 Genomes Consortium. Electronic address, magnus nordborg gmi oeaw ac at,
1437 Genomes, Consortium. 1,135 Genomes Reveal the Global Pattern of
1438 Polymorphism in Arabidopsis thaliana. *Cell* 2016; **166**: 481–491.

1439 101 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A *et al.*
1440 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
1441 generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.

1442 102 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L *et al.* A program for
1443 annotating and predicting the effects of single nucleotide polymorphisms,
1444 SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2;
1445 iso-3. *Fly* 2012; **6**: 80–92.

1446 103 Mandáková T, Lysak MA. Chromosome Preparation for Cytogenetic Analyses in
1447 Arabidopsis. *Curr Protoc Plant Biol* 2016; **1**: 43–51.

1448 104 O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, Nery JR *et al.*
1449 Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*
1450 2016; **165**: 1280–1292.

1451