ARTICLES
https://doi.org/10.1038/s41559-021-01523-y

nature ecology & evolution

Check for updates

OPEN

Concerted genomic and epigenomic changes accompany stabilization of *Arabidopsis* allopolyploids

Xinyu Jiang[1], Qingxin Song[1,2], Wenxue Ye[1] and Z. Jeffrey Chen[2]✉

During evolution successful allopolyploids must overcome 'genome shock' between hybridizing species but the underlying process remains elusive. Here, we report concerted genomic and epigenomic changes in resynthesized and natural *Arabidopsis suecica* (TTAA) allotetraploids derived from *Arabidopsis thaliana* (TT) and *Arabidopsis arenosa* (AA). *A. suecica* shows conserved gene synteny and content with more gene family gain and loss in the A and T subgenomes than respective progenitors, although *A. arenosa*-derived subgenome has more structural variation and transposon distributions than *A. thaliana*-derived subgenome. These balanced genomic variations are accompanied by pervasive convergent and concerted changes in DNA methylation and gene expression among allotetraploids. The A subgenome is hypomethylated rapidly from F_1 to resynthesized allotetraploids and convergently to the T-subgenome level in natural *A. suecica*, despite many other methylated loci being inherited from F_1 to all allotetraploids. These changes in DNA methylation, including small RNAs, in allotetraploids may affect gene expression and phenotypic variation, including flowering, silencing of self-incompatibility and upregulation of meiosis- and mitosis-related genes. In conclusion, concerted genomic and epigenomic changes may improve stability and adaptation during polyploid evolution.

Polyploidy or whole-genome duplication (WGD) is a pervasive feature of genome evolution in animals and flowering plants[1–6]. Many important crops are allopolyploids, such as wheat, cotton and canola and autopolyploids including alfalfa and potato. Many other plants, such as *Arabidopsis thaliana* and maize, are palaeopolyploids that underwent one or more rounds of WGD during evolution. The common occurrence of polyploidy suggests advantages for polyploids to possess genomic diversity, gene expression and epigenetic changes in response to selection, adaptation and domestication[1,2,6,7]. Notably, many newly resynthesized or naturally formed allotetraploids have experienced 'genome shock'[8], including rapid genomic reshuffling as observed in *Brassica napus*[9] and *Tragopogon miscellus*[10], while others, such as *A. suecica*[11–13] and cotton (*Gossypium*) allotetraploids[14], show genomic stability and conservation. The basis for this paradox between rapid genomic reshuffling and relatively stable genomes among different allopolyploids is unknown.

Arabidopsis is a powerful model for studying plant biology and polyploid evolution, consisting of diploids (for example, *A. thaliana*, Ath), autotetraploids (*A. arenosa*, Aar and *A. lyrata*, Aly) and allotetraploids such as *A. suecica* (Asu)[15] and *A. kamchatica* (Aka); the latter was formed between *A. lyrata* and *A. halleri* (Aha)[16]. Asu (AATT, $2n=4x=26$) was formed naturally[15] and can also be resynthesized by pollinating tetraploid Ath Ler4 ecotype (TTTT, $2n=4x=20$) with Aar (AAAA, $2n=4x=32$) pollen, generating two independent and genetically stable strains (Allo733 and Allo738)[11,12]. Consistently, A subgenome of natural *A. suecica* is reported to be more closely related to tetraploid than diploid *A. arenosa*[17]. Resynthesized and natural *A. suecica* provides a powerful model for studying genetic and epigenetic changes in morphological evolution, non-additive gene expression, nucleolar dominance and hybrid vigour[7,11–13,18–22]. However, despite genomes of 1,135 *A. thaliana*[23] and several related species, including Aly[24], Aha[25] and Aka[26], having been sequenced, *A. arenosa* and *A. suecica* genomes are unavailable, except for a draft sequence of Asu[13].

Here, we report high-quality sequences of both Ath Ler and Aar genomes in resynthesized allotetraploids and two subgenomes of natural *A. suecica*. Using these sequences, we studied genomic variation, DNA methylation and gene expression changes between the progenitors and their related subgenomes in resynthesized and natural *A. suecica*. Our findings indicate that balanced genomic diversifications in allotetraploids are accompanied by convergent and concerted changes in DNA methylation and gene expression between two subgenomes. This example of genomic and epigenomic reconciliation may provide a basis for stabilizing subgenomic structure and function to improve adaptation during polyploid evolution.

Results

Sequences, assemblies and annotation of *A. suecica* and *A. arenosa* genomes. *A. arenosa* is obligately outcrossing and highly heterozygous[12]. To overcome the heterozygosity issue, we sequenced the genome of a resynthesized allotetraploid, Allo738, that had been maintained by self-pollination for ten or more generations (Fig. 1a)[11,19,27]. In addition, we sequenced natural allotetraploid *A. suecica* (Asu) that was formed 14,000–300,000 years ago[13,28]. Here, we adopted chromosome nomenclatures, T1–T5 (T subgenome) and A1–A8 (A subgenome) for resynthesized allotetraploids (Allo733 and Allo738) and sT1–sT5 and sA1–sA8 for natural *A. suecica*, while Col, Ler2 (diploid) and Ler4 (tetraploid), Aar and Asu were used to specify individual genomes. The genomes were assembled de novo using integrated sequencing approaches of single-molecule real-time (PacBio Sequel, ~130×), paired-end (Illumina HiSeq, ~80×) and chromosome conformation capture (Hi-C, ~80×) (Methods). Genome sizes of *A. suecica* and Allo738

[1]State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China. [2]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX, USA. ✉e-mail: zjchen@austin.utexas.edu

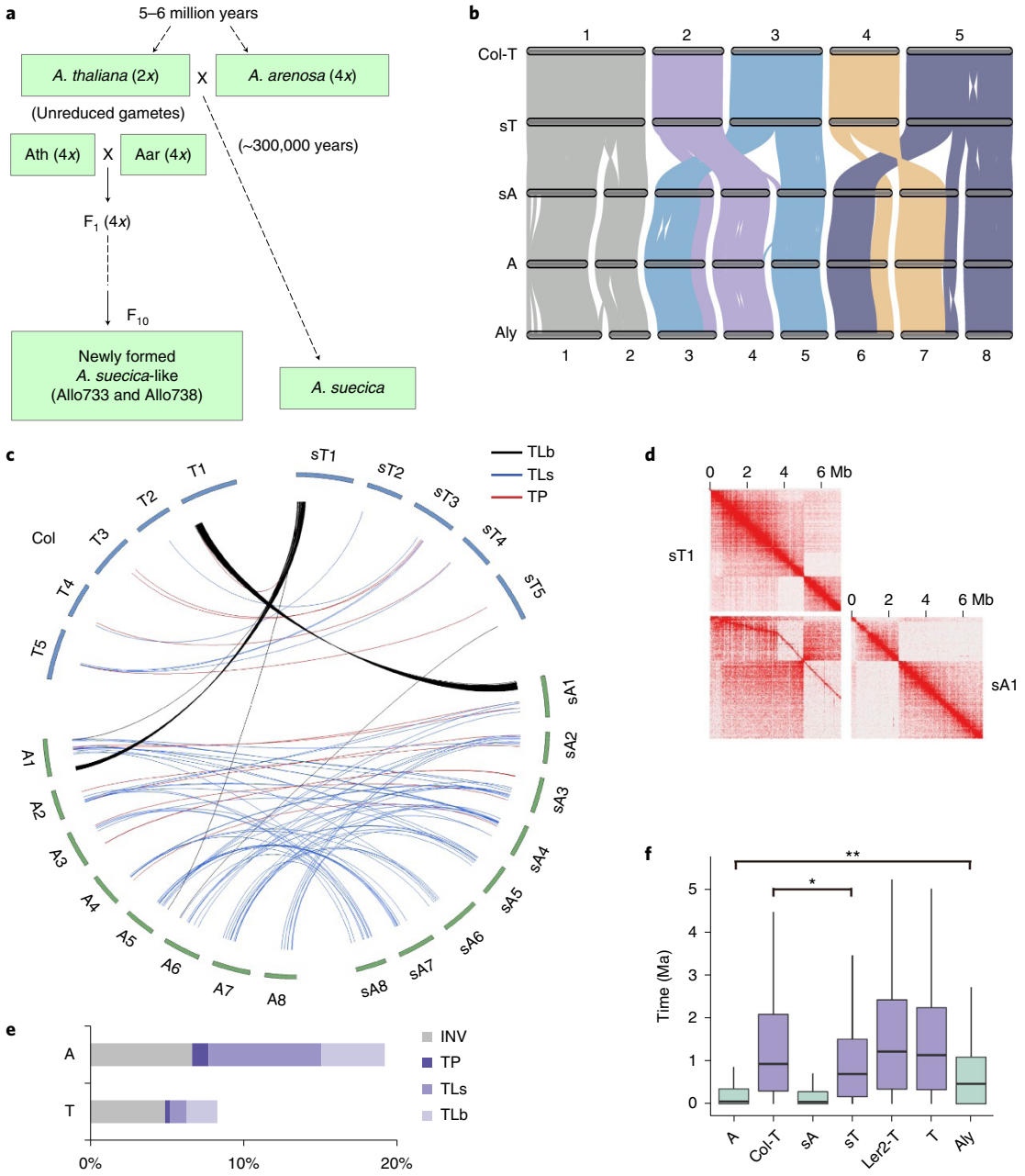


Fig. 1 | Conservation and diversification of *A. suecica* genome. a, Diagram of resynthesized allotetraploids and natural *A. suecica* (Asu). Allo733 and Allo738 are two stable *A. suecica*-like allotetraploids derived from tetraploid *A. thaliana* (Ath, Ler4) and *A. arenosa* (Aar, Care-1)[11]. Natural Asu was formed ~300,000 years ago. **b**, Genomic synteny of *A. thaliana* Col genome (Col-T), A (Aar-related) subgenome of Allo738, sT and sA subgenomes of Asu and *A. lyrata* (Aly). Syntenic blocks with 30 or more genes are shown. **c**, Rearrangements between sT (1–5) and sA (1–8) subgenomes of natural Asu and extant progenitors, *A. thaliana* (Col, T1–T5) and *A. arenosa* (A subgenome of Allo738, A1–A8). Ribbons indicate translocation between subgenomes (TLb, black), within a subgenome (TLs, blue) and transposition in the same chromosomes (TP, red). **d**, A large (~5 Mb) translocation is present between sT1 and sA1 relative to T1 and A1 chromosomes, which was validated by chromatin contact (Hi-C) maps. **e**, Proportion of sequence variation in sA and sT subgenomes of Asu relative to extant progenitors. INV, inversion; TP, transpositions; TLs, translocations within a subgenome; TLb, translocations between subgenomes. **f**, Boxplots of the estimated time for intact LTR insertion (million years ago, Ma) in A (Aar-related) subgenome of 738, Col genome (Col-T), sT and sA subgenomes of Asu, diploid Ler genome (Ler2-T), T (Ler4) subgenome of Allo738 and *A. lyrata* (Aly) genome. Single and double asterisks indicate statistical significance levels of $P < 0.05$ and 0.01, respectively (permutation test using 1,000 permutations).

were estimated to be 272.4 and 269.2 megabases (Mb), respectively, of which 96.9 and 98.6% were represented in the 13 largest scaffolds, including 120.9–121.1 Mb among five chromosomes (T1–T5) in T or sT subgenome and 147.4–150.6 Mb among eight chromosomes (A1–A8) in A or sA subgenome (Table 1 and Extended Data Fig. 1). Completeness and continuity of these genomes were supported by BUSCO scores[29] of 95.9–99.2%, although *A. suecica* genome was estimated to be ~345 Mb by flow cytometry[30,31]. *A. suecica* subgenomes were aligned colinearly with gold-standard genomes of *A. thaliana*[23,32] and *A. lyrata*[24], respectively, except for several known inversions on chromosomes sT4, sA3, sA7 and a new inversion on chromosome sT5, all of which were confirmed by Hi-C contact

Table 1 | Genome assembly and annotation statistics of two *Arabidopsis* allotetraploids

Sequence statistics	*A. suecica* (T + A)	Allo738 (T + A)
Total length of contigs (bp)	272,218,784	268,958,675
Total length of assemblies (bp)	272,391,284	269,147,175
Length of largest 13 super-scaffolds	263,860,340	265,394,178
Percentage of anchored (bp)	96.90%	98.60%
Number of contigs	380	470
Contig L50 (bp)	6,555,646	6,799,294
Number of scaffolds	269	218
Scaffold L50 (bp)	19,847,963	19,689,293
Total length of assemblies (A) (bp)	150,632,036	147,419,868
Total length of assemblies (T) (bp)	120,857,189	121,174,287
Percentage of repeat sequences (A)	26.1	25.0
Percentage of repeat sequences (T)	22.9	23.0
Number (%) of TEs (A)[a]	60,716 (20.9)	68,541 (23.8)
Number (%) of TEs (T)[a]	35,893 (21.4)	36,669 (21.6)
Number of genes (A)[a]	28,945 + 341	27,939 + 288
Number of genes (T)[a]	25,834 + 316	26,553 + 73
Complete BUSCOs (%) (A)	1,602 (99.2)	1,548 (95.9)
Complete BUSCOs (%) (T)	1,589 (98.5)	1,601 (99.2)

Note: Allo738 has *A. thaliana* (Ler) and *A. arenosa* genomes for ten or more generations[11,27], while the T (*A. thaliana* equivalent) and A (*A. arenosa* equivalent) subgenomes have evolved in Asu for 14,000–300,000 yr (refs. [13,28]). [a]Excludes those TEs and genes in unassembled scaffolds.

matrix analysis (Extended Data Fig. 2a,b). Approximately 50% of the genome is in genic regions including 54,861–55,534 annotated genes and ~20% consists of repetitive sequences including a variety of transposable elements (TEs) (Table 1 and Extended Data Fig. 1a). Many TEs were closely associated with genes and the nearest TEs from genes were closer in A-related than in T-related genomes (Extended Data Fig. 1b).

To test genome stability, we sequenced another neo-allotetraploid, Allo733 (Supplementary Fig. 1), and compared Allo733 and Allo738 genomes with Ler[33] and other *Arabidopsis* species[13,17] including *A. arenosa* (Aar) accession[34]. These data suggest that the newly assembled T subgenome of neo-allotetraploids is almost identical to the published Ler sequence and A subgenome is closely related to the *A. arenosa* sequence (Supplementary Information). For this study, we used A and T subgenomes of Allo738 (and Allo733) as *A. arenosa* (Aar) and *A. thaliana* (Ath, Ler) genomes, respectively, for further analysis.

Genomic diversity between progenitors and subgenomes. Two subgenomes in *A. suecica* have maintained high levels of colinearity and synteny compared to *A. lyrata* and extant progenitors (Aar and Ath, Col), respectively (Fig. 1b and Extended Data Fig. 1c,d). There were some large-scale sequence rearrangements, including a large translocation between sA1 and sT1 in Asu (Fig. 1c), which were confirmed by a Hi-C contact matrix analysis (Fig. 1d). Inversions and translocations occurred more frequently between *A. arenosa* and sA subgenome of *A. suecica* than between *A. thaliana* and sT subgenome (Fig. 1c and Extended Data Fig. 3a). This may suggest an increased rate of genetic diversity in outcrossing *A. arenosa* or a different *A. arenosa* strain involved in the formation of natural *A. suecica*. Whole-genome pairwise alignment analysis also showed more colinear regions in T (81.2%) than in A (56.2%) subgenome ($P = 0$, Fisher's exact test) (Fig. 1e and Extended Data Fig. 3b). While indel distributions were similar among these structural variants, single-nucleotide polymorphism (SNP) frequency in Asu A/T subgenomic translocation regions was twofold higher in sT than in sA subgenome (Extended Data Fig. 3c), suggesting stable maintenance of high SNP frequency in the A segment and low SNP frequency in the T segment of these homologous exchange (HE) regions. Notably, the total amount of HEs between two subgenomes in Allo738 was relatively small, only ~21.5 kilobases (kb) of Ath origin in A subgenome and ~1.4 Mb of Aar origin in T subgenome (Supplementary Dataset 1), suggesting a minor role of HEs in evolution of *A. suecica* allotetraploid genomes.

To assess nucleotide sequence evolution, we estimated synonymous (K_s) and non-synonymous (K_a) mutation rates using 14,668 single-copy orthogroups identified in Ath, Aar, Asu and Aly (Extended Data Fig. 3d and Methods). K_s value distribution was higher between *A. arenosa* and sA subgenome than between *A. thaliana* and sT subgenome. However, K_s value was similar between *A. arenosa* and *A. thaliana* and between two *A. suecica* subgenomes, suggesting concerted and independent evolution of subgenomes in allotetraploids. Considering that large structural variation affects genomes of evolutionary rate[35], genic sequences in rearranged regions between subgenomes, excluding small amounts of HEs, had lower neutral mutation rates than those in the syntenic regions ($P < 0.05$, Mann–Whitney U-test) (Extended Data Fig. 3e). Overall, K_a/K_s values were uniformly small among those species tested (Extended Data Fig. 3f,g), implying purifying selection. However, purifying selection is generally weaker due to redundancy of homologues in allopolyploids as reported in *A. kamchatica*[26] and *Capsella bursa*[36], and allopolyploidy might have weakened natural selection because of this bottleneck effect.

Among repetitive DNA, proportion of TEs in each subgenome was relatively similar (20.9–23.8%), although A subgenome had twice as many as T subgenome (Table 1). The order of TE insertion time was *A. thaliana* > *A. lyrata* > *A. arenosa*. (Fig. 1f), which tended to correlate with different mating systems and reduced from the transition of outcrossing in *A. lyrata* to selfing[37]. However, long terminal repeat (LTR) retrotransposons were more active (younger insertion events) in sT subgenome of *A. suecica* than in *A. thaliana* Ler and Col. Among 25 other *A. thaliana* ecotypes published previously[38,39], all except one had older LTR insertion events than sT. Kyo, an ecotype from Kyoto, Japan, had similar LTR insertion time to sT subgenome (Extended Data Fig. 3h). This result may suggest that T subgenome donor of Asu has more active LTRs.

Gene family expansion and contraction in allotetraploids. OrthoFinder identified 18,428 genes shared among *A. thaliana*, *A. arenosa* and A and T subgenomes of *A. suecica* (Fig. 2a). Among A-lineage orthogroups, gene families (744) from *A. arenosa*, *A. suecica*, *A. lyrata* and *A. halleri* were overrepresented in gene ontology (GO) terms of pollen–pistil interaction, multi-organism process, microbody and peroxisome (Fig. 2b), supporting their characteristics of outcrossing. GO enrichment terms of T-lineage orthogroups (1,415) from *A. thaliana* and *A. suecica* included endomembrane system and transfer RNA aminoacylation for translation.

Analysis of gene family contraction and expansion revealed uneven rates of gain or loss among allotetraploid species examined (Fig. 2c). Unlike similar numbers of gene family gain or loss in its diploid relatives, there was more gene family loss in T subgenome (gain/loss; 280/1,298) and more gene family gain in A subgenome (1,613/882) of *A. suecica* (Fig. 2c), which were unique to subgenomes and their respective extant species, respectively (Extended Data Fig. 4a). Note that *A. lyrata* and *A. arenosa* may be more closely related[17] and clustering between *A. lyrata* and *A. halleri* could result from the small number of species examined. Domain-based annotation showed a similar trend of gene family gain or loss between A- or T-lineage orthologues with a few exceptions (Extended Data

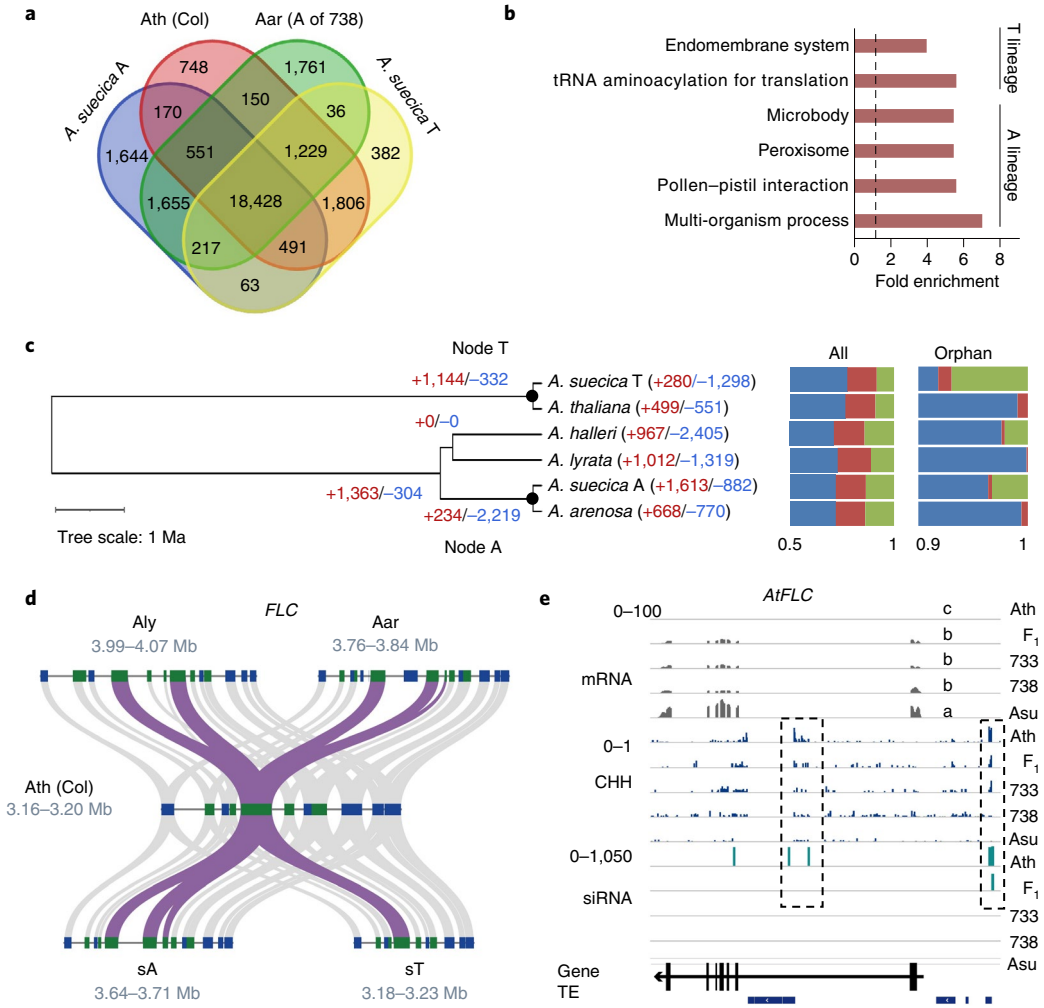


Fig. 2 | Gene family expansion and contraction. a, Venn diagram of orthologue clusters among *A. thaliana* (Ath, Col), *A. arenosa* (A subgenome of 738) and *A. suecica* (sT and sA subgenomes). **b**, GO enrichment terms of the genes specific to T lineage (Col, sT subgenome of *A. suecica*) and A lineage (*A. arenosa*, sA subgenome of *A. suecica*, *A. lyrata*, *A. halleri*). Dashed line indicates onefold enrichment. **c**, Expansion and contraction of gene families in *Arabidopsis*-related species with the numbers in parenthesis, indicating gene families subject to expansion (+red) and contraction (-blue), respectively. Black dots indicate node T (ancestor of *A. thaliana*) and node A (ancestor of *A. arenosa*), respectively. Bar graphs show the proportion of single-copy (blue), two-copy (red) and multi-copy (green) genes among all (All) and orphan (Orphan) genes in corresponding species. **d**, Micro-colinearity patterns between *FLC* and flanking genes in Col, Aar (A subgenome of 738), *A. suecica* (sT and sA subgenomes) and *A. lyrata* (Aly). Ribbons indicate colinearity of *FLC* genes (purple) and its flanking genes (grey). **e**, *FLC* expression is correlated with CHH methylation and siRNA levels in *A. thaliana* (Ath), F_1, resynthesized allotetraploids (Allo733 and Allo738) and natural *A. suecica* (Asu). Scales indicate mRNA (0–100), CHH methylation density (0–1) and 24-nucleotide siRNA (0–1,050) levels. Different letters in mRNA (transcripts per kilobase million) indicate statistical significance of $P < 0.05$ (analysis of variance (ANOVA) test, $n = 3$).

Fig. 4b). For example, F-box and CCHC-type zinc finger domain gene families shrank in T lineage but expanded in A lineage and the trend was opposite for the gene families with histone-fold associated domains and cytochrome P450 domains (Extended Data Fig. 4b). These differences in the gene family loss or gain between subgenomes may suggest pervasive lineage-specific evolutionary heterogeneities in allopolyploids, as observed among five *Gossypium* allotetraploid species[14].

Flowering time variation and S locus evolution in allopolyploids. Copy number variation has functional consequences[40]. *FLOWERING LOCUS C* (*FLC*), a MADS-box transcription factor, inhibits early flowering[41]. *FLC* has a copy number variation among *Arabidopsis* species, one in *A. thaliana*, two in *A. lyrata* and three in *A. arenosa* and A subgenome of *A. suecica*[42] (Fig. 2d). The first intron of *FLC* diverged dramatically (Extended Data Fig. 5a), for its role in *FLC* expression in response to vernalization via long non-coding RNAs[43,44]. Interestingly, *AaFLC1* and *AaFLC2* in *A. suecica* were clustered in one clade (Extended Data Fig. 5b), suggesting concerted evolution. The *FLC* copy number variation correlated with flowering time among these species[20], earliest in *A. thaliana*, followed by *A. arenosa*, resynthesized allotetraploid F_1 and stable Allo738 and Allo733 and the latest in natural *A. suecica*, which was consistent with higher *FLC* expression with lower DNA methylation levels in rosette leaves before bolting (Fig. 2e). Methylated regions are also target sites of small interfering RNAs[45], which may induce RNA-directed DNA methylation (RdDM)[46]. Similar results were observed for other A-lineage *FLC* homologues in *A. arenosa* and *A. suecica* (Extended Data Fig. 5c). Thus, RdDM may also regulate *FLC* expression and vernalization.

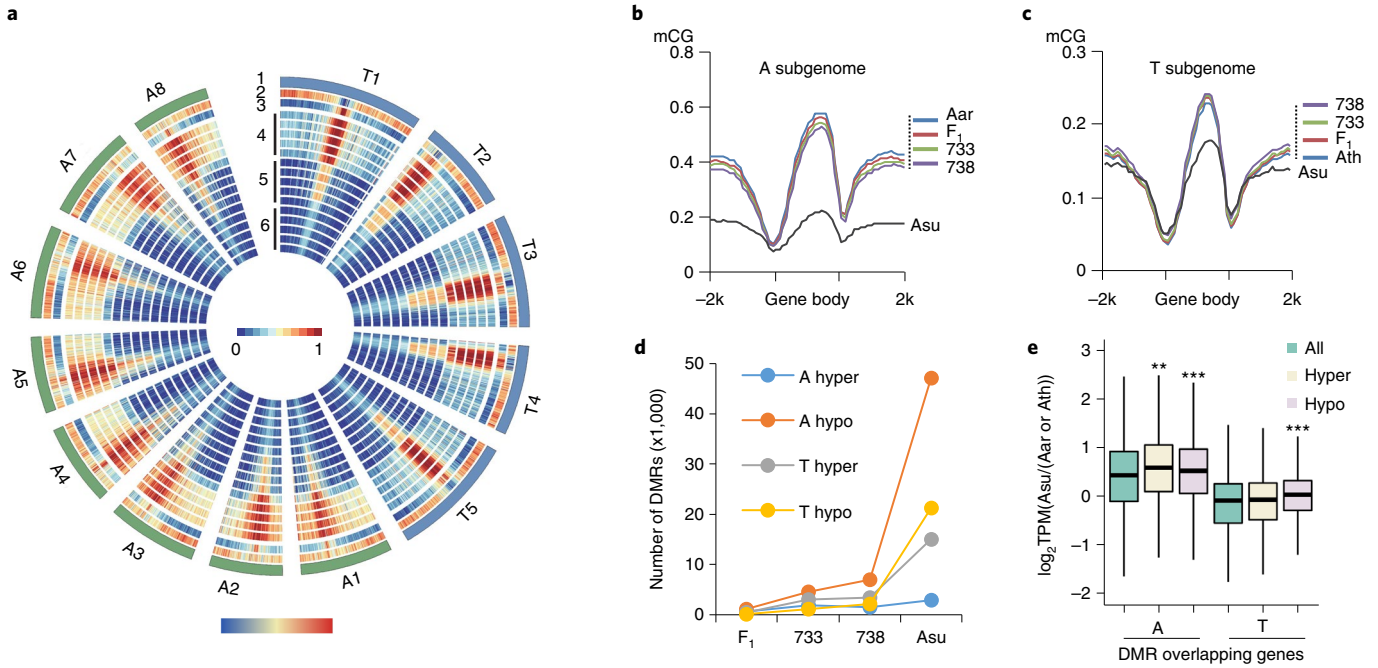


Fig. 3 | DNA methylation dynamics during the formation and evolution of allotetraploid *Arabidopsis*. **a**, Chromosome features and methylation distributions. Notes in circos plots: (1) chromosomes, (2) gene and (3) TE density and (4) CG, (5) CHG and (6) CHH methylation levels using 100-kb windows in Ath (Ler4) or Aar, F_1, Allo733, Allo738 and *A. suecica* (in that order from outside to inside in each methylation context). Note that strain identity is omitted in naming T and A chromosomes. **b**,**c**, CG methylation levels in the gene body and flanking (2-kb) sequences of the A (**b**) and T (**c**) subgenomes in F_1, Allo733 (733), Allo738 (738) and *A. suecica* (Asu), relative to *A. thaliana* (Ath) and *A. arenosa* (Aar), respectively. **d**, Numbers of DMRs between T subgenome and Ath (Col) or A subgenome and Aar in F_1, 733, 738 and Asu, respectively. **e**, Expression ratio $\log_2$TPM(Asu/(Aar or Ath)) of the genes flanking a 2-kb region of hypo- or hyper-DMRs between *A. suecica* (Asu) and *A. arenosa* (Aar) or *A. thaliana* (Ath, Ler4). Three asterisks indicate a statistical significance level of $P < 0.001$ (Mann–Whitney U-test). TPM, transcripts per kilobase per million.

Allopolyploids often become self-compatible, regardless of outcrossing behaviours in progenitors, suggesting silencing of self-incompatibility (S) locus from outcrossing *A. arenosa* in neo-allotetraploids and natural *A. suecica*[7,47,48]. S locus system comprises a combination of S locus cysteine-rich (SCR) protein in pollen coat and S locus receptor kinase (SRK) expressed on stigma surface[49]. *SRK* genes in A subgenome of resynthesized and natural *A. suecica* resembled *AhSRK1* and *AhSRK2* haplotypes[50], respectively, both of which are weak alleles in the S locus dominance hierarchy than *AhSRK4* haplotype in T subgenome[51] (Supplementary Fig. 2). These weak alleles that were immediately silenced by microRNA may contribute to a loss of self-incompatibility in early stages of allotetraploids and become non-functional in natural *A. suecica* (Supplementary Information).

Dynamic changes of DNA methylation in allotetraploids. Conserved genomic synteny between allotetraploids and related species may suggest a role for epigenetic modifications in non-additive gene expression in resynthesized and natural allopolyploids[7,22,52,53]. We examined methylome diversity in *A. thaliana* (Ath, Ler4), *A. arenosa* (Aar, 4*x*), F_1, Allo738 and Allo733[11,12,27] and natural *A. suecica* (Asu) (Extended Data Figs. 6a,b). To improve data comparability, we used shared methylation sites (35,853,727) and conserved cytosine with three or more reads among different lines for further analysis (Extended Data Fig. 6a,b). DNA methylation in plants occurs in CG, CHG and CHH (H = A, T or C) contexts[54]. Despite a similar proportion of repetitive DNA between *A. thaliana* and *A. arenosa* (Table 1), overall CG methylation levels were higher in *A. arenosa* than in *A. thaliana* (Fig. 3a and Extended Data Fig. 6a,b). Moreover, average methylation levels were highly correlated between parents (Aar/Ath) and F_1, Allo733, Allo738 or *A. suecica* (from the highest to the lowest) (Extended Data Fig. 6c). However, in *A. suecica*, A subgenome had lower methylation levels in all contexts especially the CG sites than *A. arenosa* ($P < 0.01$, Mann–Whitney U-test), while methylation levels in CHG sites were lower in T subgenome ($P < 0.01$, Mann–Whitney U-test) than in *A. thaliana* (Fig. 3a and Extended Data Fig. 6a,b,d,e). This CG hypomethylation between Asu and F_1, Allo733 or Allo738 was observed in all allotetraploids and more profound in A subgenome with a sharp reduction of methylation levels in the gene body and 5′ and 3′ sequences ($P < 0.001$, Asu versus Aar, Mann–Whitney U-test) (Fig. 3b), whereas in T subgenome hypomethylation might occur mainly in the gene body ($P > 0.05$, Asu versus Ath, Mann–Whitney U-test) (Fig. 3c). A similar trend was also observed in CHG methylation levels (Extended Data Fig. 6f) and to a lesser degree in CHH context (Extended Data Figs. 6g) of A subgenome. The data suggest that epigenomic modifications are dynamic, which occur largely in CG and CHG sites of natural *A. suecica* and throughout coding sequences, including 5′ and 3′ untranslated regions (UTRs) of A subgenome and in the gene body of T subgenome.

To track methylation changes during polyploid formation and evolution, we analysed differentially methylated regions (DMRs) between T subgenome and *A. thaliana* (Ath) or A subgenome and *A. arenosa* (Aar) in each allotetraploid. The majority of two DMR groups did not overlap (Extended Data Fig. 7a). Among 13,485 CG, 3,686 CHG and 2,785 CHH hypo-DMRs that overlapped with genes (within a 2-kb flanking region), 10,934 (81.8%), 612 (16.6%) and 272 (9.8%) were specific to CG, CHG and CHH DMRs, respectively (Extended Data Fig. 7b), suggesting association of most CG DMRs with genes. Some (14–62% in A subgenome and 14–44% in T subgenome) of these DMRs induced in F_1 were maintained in resynthesized and natural *A. suecica* (Extended Data Figs. 7c), as observed

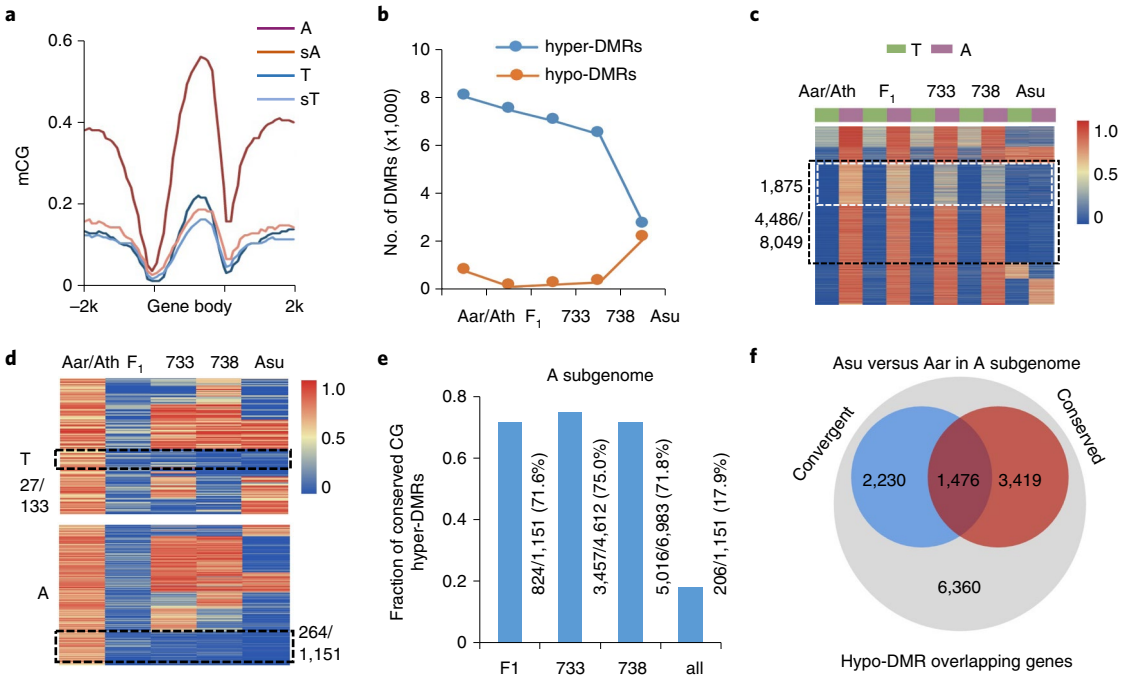


Fig. 4 | Convergence and inheritance of CG methylation levels between two subgenomes in allotetraploids. a, CG methylation levels of homologues in *A. thaliana* (Ler4, T), *A. arenosa* (Aar, A), sT and sA subgenomes of *A. suecica*. **b**, Numbers of DMRs between T subgenome and Ath (Col) or A subgenome and Aar in F_1, 733, 738 and Asu, respectively. Note that Allo733 and Allo738 may be treated as biological replicates of resynthesized allotetraploids. **c**, Clustering analysis of CG hyper-DMRs (A-T) in Aar/Ath and their changes in F_1, Allo733 (733), Allo738 (738) and *A. suecica* (Asu), respectively. Dashed black box indicates 4,486 convergent DMRs where hyper-DMRs between A and T (Ler4) were conserved in newly formed allotetraploids and reduced to the sT subgenome level in Asu. Note that the upper portion (white dashed box) indicates the overlap group (1,875) with conserved DMRs (also see **e**). **d**, Clustering of CG hypo-DMRs in F_1 and their changes in 733, 738 and Asu relative to Aar/Ath. Black dashed boxes indicate hypo-DMRs between T subgenome and Ath (upper panel) and between A subgenome and Aar (lower panel) in F_1 were conserved in Allo733, Allo738 and Asu. **e**, Fraction of conserved CG hypo-DMRs in F_1, Allo733, Allo738 and all three lines and their inheritance in Asu relative to Aar, with the numbers (conserved/total) shown to the left of each column. **f**, Venn diagram of the genes that overlapped with convergent (blue) and conserved (red) CG hypo-DMRs in Asu relative to Aar. Absolute values of CG methylation change thresholds were 0.5 in **c**,**d**.

in cotton allotetraploids[53]. Notably, hypo-DMRs in CG context were negatively associated with expression levels of DMR-associated genes in natural *A. suecica* (Extended Data Fig. 7d). Relative to DMRs between parents (Aar and Ath), the number of hypo-DMRs in A subgenome was substantially higher than that of hyper-DMRs in *A. suecica*, but hypo- and hyper-DMRs in T subgenome were similar and increased to a middle level in Asu (Fig. 3d). Moreover, expression levels of DMR-associated genes correlated negatively with hypo-DMRs but not with hyper-DMRs in both A and T subgenomes (Fig. 3e).

The number of CHG and CHH hypo-DMRs had similar changes in A subgenome, which increased slightly in F_1 and neo-allotetraploids and dramatically in *A. suecica*, while CHG hypo-DMRs in T subgenome increased dramatically only in *A. suecica* (Extended Data Fig. 7e,f). CHH hypo-DMRs displayed a similar trend to CHG hypo-DMRs, except that CHH hyper-DMRs had the highest number in T subgenome among all allotetraploids (Extended Data Fig. 7f). Considering that CG methylation is relatively abundant and stable and correlates with gene expression, we focused most analyses on CG methylation dynamics.

In plants, CG and CHG methylation is largely maintained by methyltransferase 1 (MET1)[55] and chromo methyltransferase 3 (CMT3)[56], respectively. CHH methylation is controlled by RdDM or RdDM-independent pathway[46]. *Repressor of Silencing1* (*ROS1*), encoding DNA glycosylase/AP lyase[57], is responsible for demethylation and maintains methylation homoeostasis by RdDM[58]. Consistent with genome-wide hypomethylation in *A. suecica*, *MET1* and *CMT3* were expressed at the lowest level in *A. suecica* and slightly higher levels in F_1 and two neo-allotetraploids, whereas *AtROS1-1* and *AaROS1–2* were expressed at high levels in *A. suecica* (Extended Data Fig. 8a). Upregulation of *ROS1* correlated with increased CHH methylation levels in promoter region (within a TE) and decreased CG methylation levels in gene body of *AtROS1* (Extended Data Fig. 8b) and *AaROS1–2* (Extended Data Fig. 8c) in neo-allotetraploids and *A. suecica*. *AaROS1–1* expression level was low in all lines tested. This type of allelic expression variation was also observed for *CCA1* and *FLC* homologous loci in allotetraploids[20,59], which may be controlled by a mechanism similar to nucleolar dominance[19,60]. Consistent with feedback regulation of *ROS1* expression by RdDM pathway[58], expression of several RdDM pathway genes examined was upregulated in *A. suecica* and neo-allotetraploids (Extended Data Fig. 8a), while CHH methylation levels were higher in the F_1, Allo733 and Allo738 than in *A. thaliana* or *A. arenosa* (Extended Data Fig. 6a,b and Extended Data Fig. 8b). We speculate that increased CHH methylation via RdDM pathway may lead to upregulation of *ROS1* expression, reducing overall methylation levels of A subgenome in natural *A. suecica*.

Homologous convergence of methylation changes in allotetraploids. Changes in DNA methylation levels between *A. arenosa* and A subgenome of allotetraploids can become convergent or conserved. Conserved DMRs were defined as hypo-DMRs in Asu and consistently present in F_1, Allo733 or Allo738, while convergent DMRs were identified as hyper-DMRs between Aar and Ath and

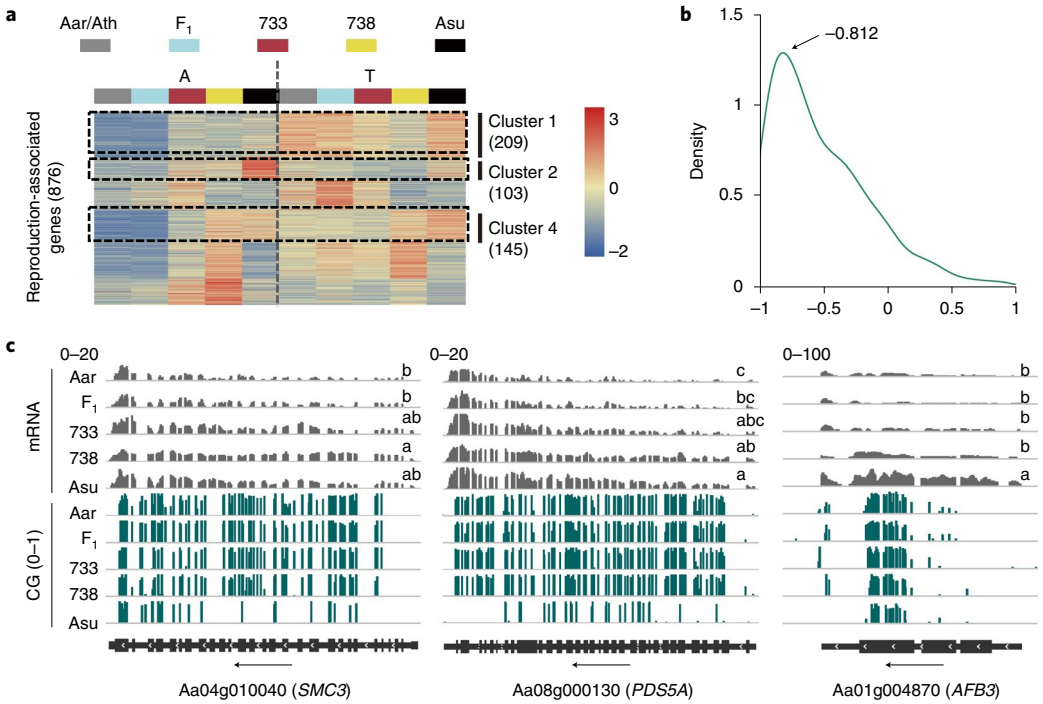


Fig. 5 | Association of CG methylation with expression of reproduction-related genes in *Arabidopsis* allotetraploids. a, Clustering analysis of expression levels of reproduction-related genes (GO:0000003) in *A. arenosa* (Aar, A) and *A.thaliana* (Ler4, T) (Aar/Ath), F_1, Allo733 (733), Allo738 (738) and *A. suecica* (Asu). **b**, Density plot of correlation coefficients between expression and CG methylation levels of the reproduction-related genes from clusters 1, 2 and 4 in the A subgenome. **c**, CG methylation near genic regions of *SMC3*, *PSD5A* and *AFB3* and their mRNA expression patterns in Aar, F_1, Allo733 (733), Allo738 (738) and Asu. Black arrows indicate the orientation of genes. SMC3, STRUCTURAL MAINTENANCE OF CHROMOSOMES3; PDS5A, PHYTOENE DESATURASE; AFB3, AUXIN SIGNALING F-BOX3. Scales indicate mRNA (0–20 and 0–100) and CG methylation density (0–1) levels. Different letters in mRNA (TPM) indicate statistical significance of $P < 0.05$ (ANOVA test, $n = 3$).

in F_1 and neo-allotetraploids and decreased to a similar level to T subgenome in Asu. We examined CG methylation levels of homologous gene pairs between two subgenomes in *A. suecica* and between Ath Ler4 (T) and Aar (A). In contrast to substantially overall higher methylation levels in *A. arenosa* than in *A. thaliana*, *A. suecica* had similar methylation levels between A and T homologues (Fig. 4a). This was accompanied by dramatic reduction of CG methylation levels in A homologues, which convergently reached a similar level to T homologues in *A. suecica*. Two subgenomes tend to maintain similar methylation levels during allopolyploid evolution.

We further analysed dynamics of hypo- and hyper-DMRs between Aar (A) and Ath (T) and between two subgenomes among different allotetraploids (Fig. 4b). The number of hyper-DMRs was reduced slightly from F_1 to Allo733 and Allo738 (~F_{10}) and dramatically in natural *A. suecica*, while the number of hypo-DMRs was relatively similar among F_1 and neo-allotetraploids but increased in *A. suecica*. Remarkably, 55.7% (4,486/8,049) of these hyper-DMRs (A versus T) were conserved in F_1, Allo733 and Allo738 and became hypomethylated at the same level in *A. suecica* (Fig. 4c), while smaller fractions of DMRs that remained hypermethylated in the A subgenome became hypermethylated in T subgenome or both. In cotton, it is the low methylated subgenome that has hypermethylated to reach a similar level in allotetraploids[53]. Although the mode of changes is different between *Arabidopsis* and cotton allotetraploids, most DMRs between two subgenomes reach similar methylation levels and evolve convergently during allotetraploid evolution.

In addition to convergent changes in DMRs, subsets of hypo-DMRs induced in the F_1 were maintained after ten or more generations in Allo733 and Allo738, some of which were also conserved in *A. suecica* (Fig. 4d). The overlap between convergent and conserved groups represented those DMRs convergent in neo-allotetraploids and maintained in Asu (Fig. 4c). Although methylated DMRs in CG, CHG and CHH contexts could be inherited across generations, more hypo-DMRs were inherited than hyper-DMRs (Fig. 4d,e and Extended Data Fig. 7c), consistent with global demethylation of the A subgenome. For example, CG hypo-DMRs in A subgenome of *A. suecica* overlapped ~71.6% (824/1,151) in F_1, ~75.0% in Allo733 and 71.8% in Allo738 (Fig. 4e). Among 13,485 genes that overlapped with CG hypo-DMRs in Asu A subgenome, 3,706 (27.5%) were convergent ($P < 8.01 \times 10^{-6}$) and 4,895 (36.3%) were conserved ($P < 0.29$), of which 1,476 (11.0%) overlapped ($P = 1$, all with Fisher's exact test) (Fig. 4f).

DNA methylation and expression correlation of reproduction-associated genes in *A. suecica*. These methylation changes affect homologue expression in *A. suecica*. Among 764 genes that were differentially methylated between Aar and Ath but have similar homologue methylation levels in *A. suecica*, 74.5% (569/764) showed decreased expression difference in two subgenomes of *A. suecica* relative to their parents (Extended Data Fig. 9a), suggesting that methylation may contribute to concerted expression level between homologues. This result may explain genome-wide non-additive gene expression in *A. suecica* as observed using microarrays[11]. However, the microarray data did not correlate well with allelic DNA methylation patterns (Extended Data Fig. 9b, c), probably because allelic expression cannot be discriminated in microarray experiments. Alternatively, DNA methylation might not explain non-additive gene expression in early generations of allotetraploids; other modifications such as histone H3K27me3 may be involved, as observed in an interspecific hybrid[61]. Over time, convergent and

concerted methylation changes between subgenomes may contribute to gene expression variation and stability in *A. suecica*.

To test consequences of convergent and conserved DMRs in *A. suecica*, we analysed GO enrichments of hypo-DMR-associated genes in natural *A. suecica*. Convergent CG hypo-DMR-associated genes were overrepresented in reproduction, seed development, system development and cell cycle, whereas the conserved hypo-DMR-associated genes were involved in transmembrane transport, pollen development and protein phosphorylation (Extended Data Fig. 9d). Those genes involved in several distinct pathways may suggest roles of DNA methylation in shaping plant growth, development and response to stresses and genome stability in allopolyploids.

Interestingly, GO term of reproduction (GO:0000003) was overrepresented for convergent DMR-associated genes (Extended Data Fig. 9d) and 52.2% (457/876) of reproduction-related genes were upregulated in *A. suecica* (Fig. 5a), including upregulation of 209 A and 248 both homologues. Expression levels of these three gene clusters correlated negatively with CG methylation levels (Fig. 5b). For example, *STRUCTURAL MAINTENANCE OF CHROMOSOMES3* (*SMC3*) is an essential gene for sister chromatid alignment and plant viability[62,63]. *PHYTOENE DESATURASE5A* (*PDS5A*) regulates mitotic sister chromatid cohesion[64] and *AUXIN SIGNALING F-BOX3 (AFB3)* is associated with pollen maturation and stamen development[65]. CG methylation levels of three genes (*SMC1*, *SMC6B* and *PDS5B*) in the same family of *SMC3* and *PDS5A* were reduced from Allo733 and Allo738 to *A. suecica* and their expression was upregulated in *A. suecica*, compared to that in Ath and F_1 (Fig. 5c and Extended Data Fig. 10a–d). Notably, downregulation of *PDS5* (Traes_7DS_0DA047A5F), a homologue of *PDS5A* and *SMC6B* (Traes_5DL_67A6B8CEB), a homologue of *SMC3*, in allohexaploid wheat led to meiotic instability[66]. Moreover, some of these genes, including *PDS5B* and *SMC3,* are highly diverged and under strong selection in *A. arenosa* tetraploids[64]. Meiotic instability is often associated with newly formed allotetraploids (F_1) and is gradually improved in resynthesized allotetraploids by self-pollination[52,67]. We predict that demethylation and upregulation of A homologues of reproduction-related genes may contribute to reproductive stability during evolution of *A. suecica* allotetraploids.

Discussion

In this study, we generated high-quality sequences of *A. suecica* natural and neo-allotetraploids including progenitors and interrogated genomic and epigenomic contributions to polyploid formation and evolution. *A. suecica* allotetraploids have maintained genomic synteny and gene content, which is another example of stable allopolyploids, following cotton allotetraploids[14]. The genomic stability is associated with subtle genomic, TE and gene family changes, including copy number and SNP variation in the genes related to flowering time and other adaptive traits. For example, *The Boy Named Sue* (*BYS*) is a fertility quantitative trait locus (QTL)[67] and spans ~240 kb on A4 chromosome, consisting of 56 annotated genes including a *FIS2* homologue[68]. *FIS2* is absent in *A. lyrata* and has variable sequences in *A. arenosa* and *A. suecica* (Supplementary Fig. 3). Function of candidate genes for the *BYS* locus remains to be investigated.

Newly formed allotetraploids have low fertility[12] due to self-incompatibility locus[7,47] (Supplementary Fig. 2), as well as meiotic instability[67]. Hypomethylation of the A subgenome may lead to upregulation of many genes involved in reproduction (meiosis, mitosis and pollination) and adaptation (stress responses), which can improve fertility and stability in allotetraploids. In wheat, downregulation of meiosis-related genes such as *PDS5* and *SMC6* is sufficient to confer unstable meiotic phenotypes[66]. Hypermethylation of reproduction-related genes may lead to gene loss, as some essential genes including meiotic genes can rapidly return to single copy following genome duplication[69,70]. For example, *ASY2* (asynaptic mutant2), a homologue of *ASY1* (ref. [71]), is heavily methylated and poorly expressed in Allo733, Allo738 and *A. suecica* and possesses a frameshift mutation, which are not observed in *A. thaliana* or *A. arenosa* (Supplementary Fig. 4). More transcriptome, methylome and resequencing data of *A. arenosa* and *A. suecica* populations in specific developmental stages such as meiosis are needed to elucidate this relationship between hypermethylation and retention of duplicate genes in allopolyploids.

Remarkably, balanced genomic diversifications in allotetraploids are accompanied by convergent and concerted changes in DNA methylation between two subgenomes. On one hand, DNA methylation of the A subgenome is reduced immediately in F_1, gradually during selfing in allotetraploids and convergently to the T-subgenome level in natural *A. suecica*. In cotton allotetraploids, it was the low methylated subgenome that became highly methylated to reach a similar level in the allotetraploids[53], resulting in convergent methylation levels of the two subgenomes. On the other hand, subsets of differentially methylated regions are conserved from F_1 to resynthesized allotetraploids and natural *A. suecica*, as observed in cotton allotetraploids[53]. These dual processes of convergent and conserved epigenomic modifications may provide a basis for allotetraploids to stabilize the two subgenomes derived from divergent hybridizing species. A combination of balanced genomic diversity and pervasive epigenomic modifications may be responsible for stabilizing subgenomes in cotton allotetraploids, which were formed ~1.5 million years ago[14,53], as well as in resynthesized hexaploid wheat[72] and tetraploid *A. suecica*. An obvious question is why other plant polyploids, including newly formed *B. napus*[9,73], *T. miscellus*[10] and resynthesized tetraploid wheat[74], display rapid genomic reshuffling. One possibility is that new species form at the right time and under suitable conditions. The species or strains used to form *B. napus* or wheat 8,000–10,000 years ago[75] may become extinct. Alternatively, homologous chromosomes from closely related progenitors may pair as in *Tragopogon*[10]. In *A. suecica*, sT and sA subgenomes are divergent enough to prevent homologous exchanges and subject to convergent and concerted changes in DNA methylation and gene expression including silencing of uniparental ribosomal DNA (rDNA) loci epigenetically via nucleolar dominance[19,60]. With advanced sequencing and epigenomic technologies, this paradox of rapid genomic reshuffling and genomic stability will be addressed to illuminate our understanding of polyploid genome evolution and to empower our efforts on editing genes and modifying epigenetic landscapes for crop improvement.

Methods

Plant materials. Plant materials included *A. thaliana* autotetraploid (Ler4, CS3900), *A. arenosa* (Care-1, CS3901), F_1 resynthesized allotetraploids, two F_{10} synthetic allotetraploids with verified chromosome compositions (Allo733, Allo738)[27] and a natural allotetraploid of *A. suecica* (As9502). All plants were grown in the growth chamber under the 16 h light/ 8 h dark cycle at 20 °C.

Genome sequencing and assembly. DNA was extracted from young leaves of Allo738 and *A. suecica* and sequenced on the PacBio Sequel platform using 11 and eight SMRT cells to produce 37.02 gigabases (Gb) (136X genome equivalent) and 35.52 Gb (132X) of raw data, respectively. The PacBio long clean reads were corrected and assembled to contigs by MECAT (v.1.0) with parameters (correctedErrorRate 0.02)[76]. Next, the clean subreads were mapped to the assembled contigs using BLASR of SMRTLINK and errors were corrected by ARROW of SMRTLINK (v.5.0.1.9585)[77]. The Illumina pair-end reads (~80X) were mapped to consensus contigs by BWA (v.0.7.15-r1140)[78] and further polished by Pillon (v.1.22) with the following parameters (--fix bases --changes --diploid)[79]. For Allo738 and *A. suecica*, chromatin conformation capture (3C or Hi-C) sequencing data consisting of 80–90 millions of effective read pairs were mapped to final contigs by Juicer (v.1.6.2)[80] with default parameters and scaffolded to the chromosome-scale assembly by a three-dimensional de novo DNA assembly (3D DNA) pipeline (v.180114) with parameters (-r 3 -m diploid)[81]. Finally, we manually modified the assembly error using Juicebox (v.1.8.8)[82] and generated the ultimate scaffolds, whose largest 13 scaffolds represented 13 chromosomes. The A subgenome of Allo738 represents the genome of *A. arenosa* (CS3901) and the T subgenome represents the genome of the autotetraploid *A. thaliana* (Ler), as Allo738 was generated by pollinating autotetraploid

A. thaliana with tetraploid *A. arenosa* and self-pollinated for more than ten generations to minimize heterozygosity[11,27].

Repeat identification. Repeats were de novo annotated and classified as repeat consensus database for Allo738 and *A. suecica* assemblies using RepeatModeler (v.1.0.11) (http://www.repeatmasker.org/). The *Arabidopsis* section of Repbase (v.20181026) and RepeatPeps (v.20181026) (https://www.girinst.org/) and MIPS (mipsREdat_9.3p)[83] repeat databases were used to correct de novo repeat database by BLASTN (v.2.5.0+) with criteria of more than 80% identity, 50% coverage and 80-base pair (bp) length[84]. The corrected repeat database with more than 80% identity and 50% coverage of protein-coding genes (without TE-associated genes) of *Arabidopsis* was removed. We then combined the corrected de novo database with the *Arabidopsis* section of Repbase and whole-genome repeat sequences of TAIR10 (https://www.arabidopsis.org/) to generate a final repeat database. In addition, intact LTR retrotransposons were de novo annotated using LTR-FINDER (v.1.0.7) with parameters (-D 20000 -d 1000 -L 3500 -l 100 -p 20 -C -M 0.9)[85] and LTR_retriever (v.2.0) with parameters (-similar 90 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1)[86]. Lastly, repeats were identified from the intact LTR-masked assembly by RepeatMasker (v.4.0.7) with parameters (-cutoff 250) (http://www.repeatmasker.org/) against the final repeat database. To estimate the insertion time of LTR, we used the Jukes–Cantor model to calculate the distance K (ref. [87]). Then the insertion time t was calculated as $t = K/2r$, where r is the rate of nucleotide substitution, which was 7×10^{-9} per site per generation (assumed to equal one year) by LTR_retriever.

Gene annotation. Genes were annotated by the integration of ab initio prediction, homology-based prediction and RNA sequencing (RNA-seq) data evidence for Allo738 and *A. suecica*. RNA-seq reads from different tissues were mapped to the assembly using HISAT2 (v.2.1.0)[88] to generate transcripts by StringTie (v.1.3.3b)[89]. Simultaneously, the genome-guided pipeline of Trinity (v.2.6.6) with parameters (-I 20000)[90] based on GSNAP (v.2018-07-04)[91] software was used to assemble transcripts which then aligned to the assembly by PASApipeline (v.2.3.3) with parameters (--ALIGNERS blat,gmap --MAX_INTRON_LENGTH 20000 --transcribed_is_aligned_orient --stringent_alignment_overlap 30.0)[92]. Next, we used TransDecoder (v.5.3.0) to identify candidate coding regions within transcript sequences generated by both Trinity and StringTie. AUGUSTUS (v.3.2.2)[93] was used for ab initio gene prediction on the basis of the hints of intron–exon boundaries from bam files of HISAT2 and repeat boundaries from RepeatMasker and model training was based on the transcripts assembled from Trinity. The homology-based prediction was conducted via Exonerate (v.2.2.0) with parameters (--percent 50 --maxintron 20000 -n 1)[94] on the basis of *Arabidopsis* protein sequences against the assembly. EVidenceModeler (v.1.1.1) with parameters (--segment size 500000 --overlapSize 10000)[95] was used to integrate the gene annotation files generated by these three methods with different weights: 1 for Augustus, 14 for Exonerate, 5 for PASA and 14 for TransDecoder. Finally, UTRs and alternatively spliced models were added by PASApipeline.

Genes were characterized for their putative function by performing InterProScan (v.5.32-71.0) with parameters (-appl ProDom, SMART,TIGRFAM, Pfam and SUPERFAMILY,PrositeProfiles -goterms -pa -iprlookup)[96]. Small RNAs were inferred by Infernal (v.1.1.2)[97] against the Rfam database (release 14.1)[98] and tRNAs were annotated by tRNAscan-SE (v.2.0)[99].

Assessment of assembly accuracy and integrity. We evaluated the integrity of the assembly by BUSCO (v.3.0.2)[29] and the accuracy of the assembly through whole-genome alignment against the reference genome of *A. thaliana* (TAIR10, Ler)[33] or *A. lyrata* (Alyrata_384_v2.1 from JGI)[100] by MUMmer (v.4.0.0beta2) with parameters (--mum -l 100 -c 1000 -d 10 --banded -D 5 && delta-filter -i 95 -o 95)[101], which identified one-to-one and multiple-to-multiple (M-to-M, including duplicates) alignment regions. Dotplots were constructed using mummerplot in MUMmer. For analysis of Allo738 genome stability, whole-genome alignments were performed between Allo738 and Allo733 or Aar4 (bioRxiv, https://doi.org/10.1101/2020.08.24.264432) and Asu and Aar4. Local variants (SNP and indel) were identified in one-to-one alignment region using the dnadiff function of MUMmer[101].

Variant calling and phylogenetic analysis. Paired-end resequence reads of 39 *A. arenosa* and 15 *A. suecica* were downloaded from NCBI Short Reads Archive (PRJNA309923 and PRJNA284572)[17]. Downloaded reads and the reads of Asu, Allo733 and Allo738 were filtered using Trimmomatic (v.0.39) with parameters (TruSeq3-PE.fa:2:30:10:8:true LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50)[102]. Clean reads of *A. arenosa* were mapped to the Aar assembly and reads of *A. suecica*, Allo733 and Allo738 were mapped to the combination of Aar and At Col (TAIR10) assembly by BWA program (v.0.7.17-r1188) with default parameters. Only uniquely mapped paired reads (-f 3 -q 10) were used for analysing sequence variation and polymerase chain reaction (PCR) duplicates were removed using Picard Toolkit (v.2.18.15) with default parameters (Broad Institute, GitHub Repository http://broadinstitute.github.io/picard/, 2019). Variant was called through the Genome Analysis Toolkit (GATK, v.4.1.3.0) with parameters (--min-base-quality-score 25 && "QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 4.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0"--filter-name "Fail" -G-filter "DP < 5" -G-filter-name "LowDP" -G-filter "GQ < 20" -G-filter-name "LowGQ" -G-filter "isHet == 1" -G-filter-name "isHetFilter" for SNP filter && "QD < 2.0 || FS > 200.0 || SOR > 10.0 || InbreedingCoeff < −0.8 || ReadPosRankSum < −20.0"–filter-name "Fail" -G-filter "DP < 5" -G-filter-name "LowDP" -G-filter "GQ < 20" -G-filter-name "LowGQ" -G-filter "isHet == 1" -G-filter-name "isHetFilter" for InDel filter). Finally, we generated variants of A genome and T genomes, respectively. Variants of 1,035 individuals[17] and of T subgenome of *A. suecica*, Allo733 and Allo738 were merged to the final variant file of T genome. Independent SNPs from A genome with minor allele frequency (MAF) < 0.05 and missing rate >0.05 were filtered by PLINK (v.1.9) with parameters (--geno 0.05 --maf 0.05 && --indep-pairwise 50 10 0.2)[103]. SNPs of the T genome were filtered using the same criteria except for missing rate >0.02. The filtered SNPs were used to construct phylogenetic trees by the neighbour-join method in TASSEL (v.5.0)[104] and visualized using iTOL[105].

Identification of rearrangements and local differences. We used MUMmer (v.4.0.0beta2)[101] with parameters (nucmer --mum -l 50 -c 100 -b 500 -g 100 && delta-filter -l 100 -i 90) for the whole-genome alignment of *A. suecica* and the combination of its assumed progenitors, *A. thaliana* and *A. arenosa*, to identify local and high-order variation. Local variants (SNP and indel) were identified in one-to-one alignment region using the dnadiff function of MUMmer[101]. High-order variation was analysed using SyRI (v.1.1)[106].

Syntenic analysis. Synthetic blocks were identified by MCscan (Python version) of jcvi (v.0.8.12) (https://doi.org/10.5281/zenodo.31631) (parameters: --cscore = .99) with 30 genes spanned per block[107].

Identification of orthologous genes for K_a/K_s calculations and phylogenetic inference. Orthologous gene clusters were recognized using OrthoFinder[108] (v.2.2.7) with parameters (-S diamond -M msa -T raxml)[109]. Single-copy genes of *A. thaliana*, *A. arenosa*, *A. suecica* and *A. lyrata* were used to calculate K_s, K_a and K_a/K_s values[110] by KaKs_Calculator (v.1.2)[111]. For gene family analysis, single-copy genes of *A. thaliana*, *A. arenosa*, *A. suecica*, *A. lyrata* and *A. halleri* were extracted using OrthoFinder[108] (v.2.2.7) and parameters (-S diamond -M msa -T raxml)[109] and r8s (v.1.81) were used to estimate divergence time to construct phylogenetic trees[112] with the constrained divergence time range following TimeTree[113]. Contraction and expansion of gene families were identified by CAFE (v.4.2.1) (parameters: -p 0.05 -filter)[114], which accounted for phylogenetic history and provided a statistical basis for evolutionary inference. *P* values were used to estimate the likelihood of the observed sizes given average rates of gain and loss and used to determine expansion or contraction for individual gene families in each node.

Small RNA-seq data analysis. Small RNA data were collected in rosette leaves before bolting for *A. thaliana*, *A. arenosa*, F_1, Allo733 and *A. suecica*[45] and downloaded from NCBI (GSE15443). Small RNA reads were mapped onto Allo738 genome using ShortStack (v.3.8.5)[115].

mRNA-seq data analysis. Total RNA was isolated from rosette leaves (6–7 weeks old), seedlings, flowers and fruit pods in Allo738 and *A. suecica* and was used for messenger RNA sequencing with three biological replicates with ~6.5 Gb per replicate on Illumina HiSeq X Ten platform. The mRNA-seq data were also collected for *A. thaliana*, *A. arenosa*, F_1, Allo733, *A. suecica* from (GSE29687)[116] and (GSE50715)[27]. Low-quality reads were filtered using Trimmomatic (v.0.39) with parameters (TruSeq3-PE.fa:2:30:10:8:true LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50)[102]. To exclude expression bias between *A. thaliana* and *A. arenosa* due to depth difference, reads of *A. thaliana* and *A. arenosa* were down-sampled to the same level and combined. Reads of *A. thaliana*, *A. arenosa*, F_1, Allo733 and *A. suecica* were mapped to the Allo738 genome along with the SNP table of Asu and Allo733 genomes, respectively, using HISAT2 and StringTie[117] with parameters (--score-min L, 0.0,−0.4). Reads of Allo738 were mapped to the Allo738 genome using HISAT2 and StringTie with default parameters. Only uniquely mapped reads were kept for further analysis. The expression level of each gene was calculated using StringTie. We selected homologous genes between Asu and Allo738 for expression between allotetraploid species and homologous gene pairs between A and T subgenomes for expression within an allotetraploid.

MethylC-seq data analysis. Total genomic DNA was extracted from rosette leaves before bolting (3–4 weeks for *A. thaliana* and 6–7 weeks for *A. arenosa*, F_1, Allo733, Allo738 and *A. suecica*). MethylC-seq libraries were constructed using a bisulfite method as previously described[53] and sequenced on Illumina HiSeq X Ten platform (~11 Gb per replicate). Low-quality reads were filtered using Trimmomatic (v.0.39) with parameters (TruSeq3-PE.fa:2:30:10:8:true LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50)[102]. MethylC-seq reads of *A. suecica* and Allo738 were mapped to the *A. suecica* and Allo738 genome using Bismark (v.0.15.1) with parameters (--score_min L,0,−0.2), respectively[118]. MethylC-seq

reads of *A. thaliana*, *A. arenosa*, F_1, Allo733 were mapped to the Allo738 genome using Bismark (v.0.15.1) with parameters (--score_min L,0,−0.4). Reads of Allo733 were mapped onto the Allo733-SNP-substituted Allo738 genome. To remove bias, only the uniquely mapping reads and conserved cytosines were used for downstream analyses following a previous method[53] (also see Github: https://github.com/Anticyclone-op/Ara-genome-methly). To identify conserved regions of 1 kb or longer in *A. suecica* and Allo738, we aligned the *A. suecica* genome against the Allo738 genome by LAST (v.869) (parameters: last -q3 -m50 -e35 -P10 && last-split -m1 -s200)[119] and then swapped the sequences and extracted the best alignments. Finally, alignments with scores <1,000 were removed. The conserved cytosines between *A. suecica* and Allo738 were extracted using Python scripts. The same method was used to identify the conserved region and conserved cytosines between the A and T subgenomes. Shared methylation sites in two replicates were merged for further analysis.

DMRs between the T subgenome and *A. thaliana* or between the A subgenome and *A. arenosa* were analysed using 100-bp sliding windows, including four or more cytosines for CG and CHG contexts and 16 or more cytosines for CHH context. The hyper- and hypo-DMRs mean allotetraploid relative to parent. The weighted methylation level was calculated for each window. Significant differences were assessed using Fisher's exact test (FDR < 0.05), using the following cutoff values of the minimum difference of the methylation levels: 0.5 for CG DMRs, 0.3 for CHG DMRs and 0.1 for CHH DMRs. For DMRs between A and T genomes and in F_1, Allo733, Allo738 and Asu, using the same criteria, either as hyper-DMRs (A > T) or hypo-DMRs (T < A). DMR-overlapping genes were defined as those that overlapped with DMRs within a 2-kb region. Conserved DMRs were defined as the hypo-DMRs in Asu and consistently present in F_1, Allo733 or Allo738. Convergent DMRs were identified as the hyper-DMRs between Aar and Ath and in F_1 and resynthesized allotetraploids and decreased to a similar level to the T subgenome in Asu, while the overlap between two groups represented those DMRs convergent in newly formed allotetraploid and remained in Asu.

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

Data availability

Sequencing data are accessible under NCBI BioProject numbers (PRJNA669593). All datasets generated and/or analysed in the study are available in the main text, Table 1, Figs. 1–5, Extended Data Figs. 1–10, Supplementary Information and the Reporting Summary. Source data are provided with this paper.

Received: 5 November 2020; Accepted: 24 June 2021;
Published online: 19 August 2021

References

1. Soltis, D. E., Visger, C. J. & Soltis, P. S. The polyploidy revolution then… and now: Stebbins revisited. *Am. J. Bot.* **101**, 1057–1078 (2014).
2. Leitch, A. R. & Leitch, I. J. Genomic plasticity and the diversity of polyploid plants. *Science* **320**, 481–483 (2008).
3. Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131**, 452–462 (2007).
4. Chen, Z. J. Molecular mechanisms of polyploidy and hybrid vigor. *Trends Plant Sci.* **15**, 57–71 (2010).
5. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
6. Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A. Evolution of plant genome architecture. *Genome Biol.* **17**, 37 (2016).
7. Chen, Z. J. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.* **58**, 377–406 (2007).
8. McClintock, B. The significance of responses of the genome to challenge. *Science* **226**, 792–801 (1984).
9. Xiong, Z., Gaeta, R. T. & Pires, J. C. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proc. Natl Acad. Sci. USA* **108**, 7908–7913 (2011).
10. Chester, M. et al. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc. Natl Acad. Sci. USA* **109**, 1176–1181 (2012).
11. Wang, J. et al. Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. *Genetics* **172**, 507–517 (2006).
12. Comai, L. et al. Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. *Plant Cell* **12**, 1551–1568 (2000).
13. Novikova, P. Y. et al. Genome sequencing reveals the origin of the allotetraploid *Arabidopsis suecica*. *Mol. Biol. Evol.* **34**, 957–968 (2017).
14. Chen, Z. J. et al. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet.* **52**, 525–533 (2020).
15. Lind-Hallden, C., Hallden, C. & Sall, T. Genetic variation in *Arabidopsis suecica* and its parental species *A. arenosa* and *A. thaliana*. *Hereditas* **136**, 45–50 (2002).
16. Shimizu-Inatsugi, R. et al. The allopolyploid *Arabidopsis kamchatica* originated from multiple individuals of *Arabidopsis lyrata* and *Arabidopsis halleri*. *Mol. Ecol.* **18**, 4024–4048 (2009).
17. Novikova, P. Y. et al. Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).
18. Chen, Z. J. Genomic and epigenetic insights into the molecular bases of heterosis. *Nat. Rev. Genet.* **14**, 471–482 (2013).
19. Chen, Z. J., Comai, L. & Pikaard, C. S. Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nucleolar dominance) in *Arabidopsis* allopolyploids. *Proc. Natl Acad. Sci. USA* **95**, 14891–14896 (1998).
20. Wang, J., Tian, L., Lee, H. S. & Chen, Z. J. Nonadditive regulation of *FRI* and *FLC* loci mediates flowering-time variation in *Arabidopsis* allopolyploids. *Genetics* **173**, 965–974 (2006).
21. Ni, Z. et al. Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature* **457**, 327–331 (2009).
22. Lee, H. S. & Chen, Z. J. Protein-coding genes are epigenetically regulated in *Arabidopsis* polyploids. *Proc. Natl Acad. Sci. USA* **98**, 6753–6758 (2001).
23. Consortium, G. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
24. Hu, T. T. et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
25. Briskine, R. V. et al. Genome assembly and annotation of *Arabidopsis halleri*, a model for heavy metal hyperaccumulation and evolutionary ecology. *Mol. Ecol. Resour.* **17**, 1025–1036 (2017).
26. Paape, T. et al. Patterns of polymorphism and selection in the subgenomes of the allopolyploid *Arabidopsis kamchatica*. *Nat. Commun.* **9**, 3909 (2018).
27. Shi, X., Zhang, C., Ko, D. K. & Chen, Z. J. Genome-wide dosage-dependent and -independent regulation contributes to gene expression and evolutionary novelty in plant polyploids. *Mol. Biol. Evol.* **32**, 2351–2366 (2015).
28. Jakobsson, M. et al. A unique recent origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. *Mol. Biol. Evol.* **23**, 1217–1231 (2006).
29. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
30. Johnston, J. S. et al. Evolution of genome size in Brassicaceae. *Ann. Bot.* **95**, 229–235 (2005).
31. Pellicer, J. & Leitch, I. J. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* **226**, 301–305 (2020).
32. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
33. Zapata, L. et al. Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proc. Natl Acad. Sci. USA* **113**, E4052–E4060 (2016).
34. Burns, R. et al. Gradual evolution of allopolyploidy in *Arabidopsis suecica*. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.24.264432 (2021).
35. Navarro, A. & Barton, N. H. Chromosomal speciation and molecular divergence–accelerated evolution in rearranged chromosomes. *Science* **300**, 321–324 (2003).
36. Douglas, G. M. et al. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proc. Natl Acad. Sci. USA* **112**, 2806–2811 (2015).
37. de la Chaux, N., Tsuchimatsu, T., Shimizu, K. K. & Wagner, A. The predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative *Arabidopsis lyrata*. *Mobile DNA* **3**, 2 (2012).
38. Gan, X. et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011).
39. Jiao, W. B. & Schneeberger, K. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).
40. DeBolt, S. Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol. Evol.* **2**, 441–453 (2010).
41. Michaels, S. D. & Amasino, R. M. *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* **11**, 949–956 (1999).
42. Nah, G. & Jeffrey Chen, Z. Tandem duplication of the *FLC* locus and the origin of a new gene in *Arabidopsis* related species and their functional implications in allopolyploids. *New Phytol.* **186**, 228–238 (2010).
43. Swiezewski, S., Liu, F., Magusin, A. & Dean, C. Cold-induced silencing by long antisense transcripts of an *Arabidopsis* Polycomb target. *Nature* **462**, 799–802 (2009).

44. Heo, J. B. & Sung, S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* **331**, 76–79 (2011).
45. Ha, M. et al. Small RNAs serve as a genetic buffer against genomic shock in *Arabidopsis* interspecific hybrids and allopolyploids. *Proc. Natl Acad. Sci. USA* **106**, 17835–17840 (2009).
46. Zemach, A. et al. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **153**, 193–205 (2013).
47. Nasrallah, M. E., Yogeeswaran, K., Snyder, S. & Nasrallah, J. B. *Arabidopsis* species hybrids in the study of species differences and evolution of amphiploidy in plants. *Plant Physiol.* **124**, 1605–1614 (2000).
48. Mable, B. K. Polyploidy and self-compatibility: is there an association? *New Phytol.***162**, 803–811 (2004).
49. Takayama, S. & Isogai, A. Self-incompatibility in plants. *Annu Rev. Plant Biol.* **56**, 467–489 (2005).
50. Llaurens, V. et al. Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in *Arabidopsis halleri*? *Evolution* **62**, 2545–2557 (2008).
51. Durand, E. et al. Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. *Science* **346**, 1200–1205 (2014).
52. Wang, J. et al. Stochastic and epigenetic changes of gene expression in *Arabidopsis* polyploids. *Genetics* **167**, 1961–1973 (2004).
53. Song, Q., Zhang, T., Stelly, D. M. & Chen, Z. J. Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biol.* **18**, 99 (2017).
54. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
55. Kankel, M. W. et al. *Arabidopsis* MET1 cytosine methyltransferase mutants. *Genetics* **163**, 1109–1122 (2003).
56. Cao, X. & Jacobsen, S. E. Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. *Proc. Natl Acad. Sci. USA* **99**, 16491–16498 (2002).
57. Gong, Z. et al. ROS1, a repressor of transcriptional gene silencing in *Arabidopsis*, encodes a DNA glycosylase/lyase. *Cell* **111**, 803–814 (2002).
58. Lei, M. et al. Regulatory link between DNA methylation and active demethylation in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **112**, 3553–3557 (2015).
59. Ng, D. W., Chen, H. H. & Chen, Z. J. Heterologous protein–DNA interactions lead to biased allelic expression of circadian clock genes in interspecific hybrids. *Sci. Rep.* **7**, 45087 (2017).
60. Chen, Z. J. & Pikaard, C. S. Epigenetic silencing of RNA polymerase I transcription: a role for DNA methylation and histone modification in nucleolar dominance. *Genes Dev.* **11**, 2124–2136 (1997).
61. Zhu, W. et al. Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific *Arabidopsis* hybrid. *Genome Biol.* **18**, 157 (2017).
62. Liu, Cm,C. M. et al. Condensin and cohesin knockouts in *Arabidopsis* exhibit a *titan* seed phenotype. *Plant J.* **29**, 405–415 (2002).
63. Schubert, V. et al. Cohesin gene defects may impair sister chromatid alignment and genome stability in *Arabidopsis thaliana*. *Chromosoma* **118**, 591–605 (2009).
64. Yant, L. et al. Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. *Curr. Biol.* **23**, 2151–2156 (2013).
65. Cecchetti, V., Altamura, M. M., Falasca, G., Costantino, P. & Cardarelli, M. Auxin regulates *Arabidopsis* anther dehiscence, pollen maturation, and filament elongation. *Plant Cell* **20**, 1760–1774 (2008).
66. Bian, Y. et al. Meiotic chromosome stability of a newly formed allohexaploid wheat is facilitated by selection under abiotic stress as a spandrel. *New Phytol.* **220**, 262–277 (2018).
67. Henry, I. M. et al. The *BOY NAMED SUE* quantitative trait locus confers increased meiotic stability to an adapted natural allopolyploid of *Arabidopsis*. *Plant Cell* **26**, 181–194 (2014).
68. Chaudhury, A. M. et al. Fertilization-independent seed development in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **94**, 4223–4228 (1997).
69. Lloyd, A. H. et al. Meiotic gene evolution: can you teach a new dog new tricks? *Mol. Biol. Evol.* **31**, 1724–1727 (2014).
70. De Smet, R. et al. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl Acad. Sci. USA* **110**, 2898–2903 (2013).
71. Caryl, A. P., Armstrong, S. J., Jones, G. H. & Franklin, F. C. A homologue of the yeast HOP1 gene is inactivated in the *Arabidopsis* meiotic mutant asy1. *Chromosoma* **109**, 62–71 (2000).
72. Yuan, J. et al. Dynamic and reversible DNA methylation changes induced by genome separation and merger of polyploid wheat. *BMC Biol.* **18**, 171 (2020).
73. Gaeta, R. T., Pires, J. C., Iniguez-Luy, F., Leon, E. & Osborn, T. C. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell* **19**, 3403–3417 (2007).
74. Feldman, M. et al. Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics* **147**, 1381–1387 (1997).
75. Lu, K. et al. Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. *Nat. Commun.* **10**, 1154 (2019).
76. Xiao, C. L. et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).
77. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinform.* **13**, 238 (2012).
78. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
79. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
80. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
81. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
82. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
83. Nussbaumer, T. et al. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.* **41**, D1144–D1151 (2013).
84. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
85. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
86. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
87. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* (ed. Munro, H. N.) 21–132 (Academic Press, 1969).
88. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
89. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
90. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
91. Wu, C. H. et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187–D191 (2006).
92. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
93. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
94. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **6**, 31 (2005).
95. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
96. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
97. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
98. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
99. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
100. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucl. Acids Res.* **40**, D1178–D1186 (2012).
101. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
102. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
103. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
104. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
105. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
106. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
107. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).

108. Li, L., Stoeckert, C. J. Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
109. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
110. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486 (2002).
111. Zhang, Z. et al. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteom. Bioinforma.* **4**, 259–263 (2006).
112. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302 (2003).
113. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
114. Han, M. V., Thomas, G. W., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* **30**, 1987–1997 (2013).
115. Axtell, M. J. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* **19**, 740–751 (2013).
116. Shi, X. et al. *Cis*- and *trans*-regulatory divergence between progenitor species determines gene-expression novelty in *Arabidopsis* allopolyploids. *Nat. Commun.* **3**, 950 (2012).
117. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
118. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
119. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinforma.* **11**, 80 (2010).

Acknowledgements

We thank R. Burns and M. Nordborg at Gregor Mendel Institute of Molecular Plant Biology for sharing the *A. arenosa* sequence. We thank Bioinformatics Center at Nanjing Agricultural University for computational support and assistance in data analysis. Research at Nanjing Agricultural University is supported by grants from the National Natural Science Foundation of China (91631302) and Jiangsu Collaborative Innovation Center for Modern Crop Production. Z.J.C. is the D. J. Sibley Centennial Professor of Plant Molecular Genetics.

Author contributions

Z.J.C. and Q.S. conceived and designed the project. X.J. and W.Y. generated the data. Z.J.C., Q.S. and X.J. analysed the data and wrote the paper. All authors have read and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-021-01523-y.

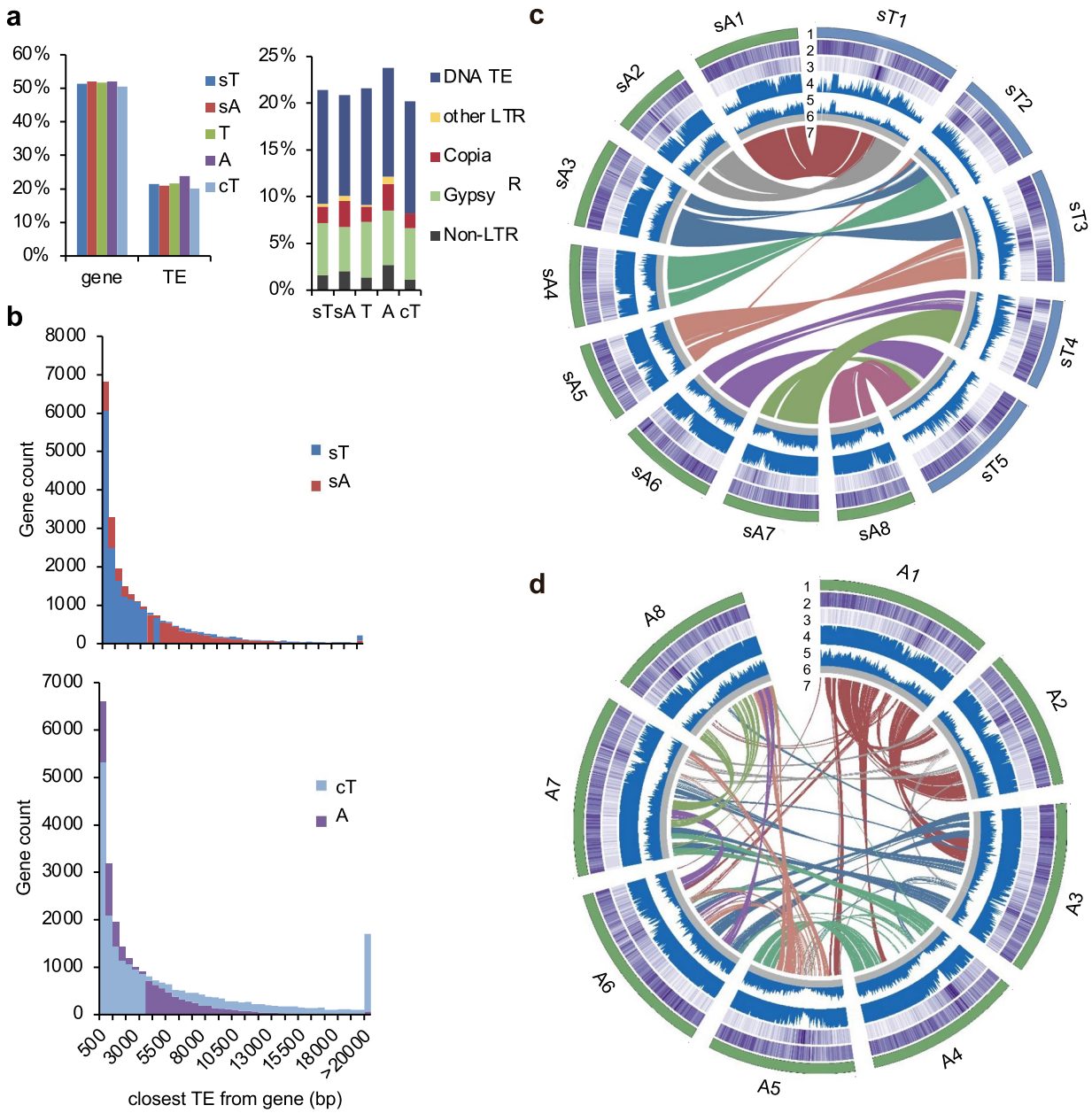
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-021-01523-y.

Correspondence and requests for materials should be addressed to Z.J.C.

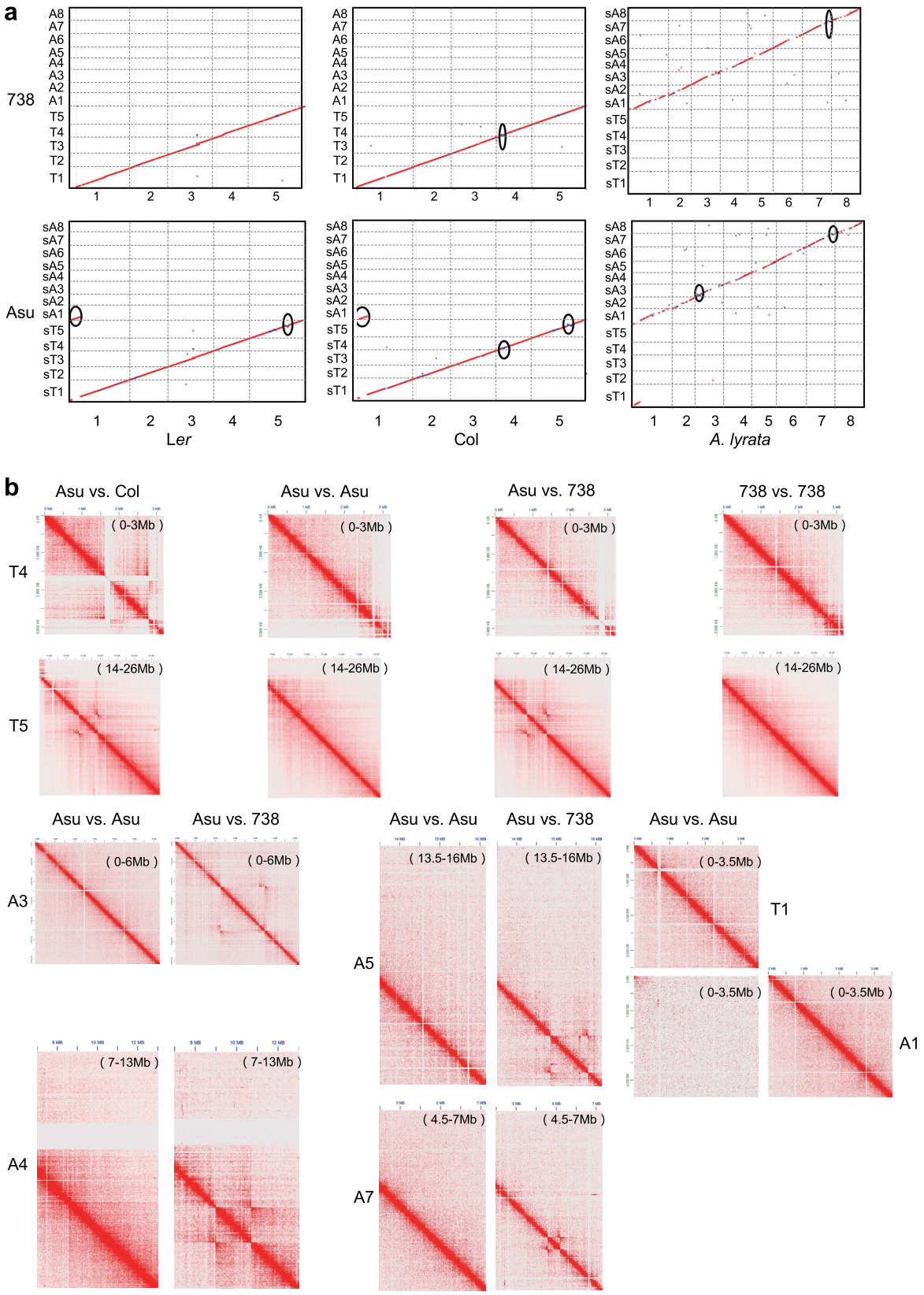
Peer review information *Nature Ecology & Evolution* thanks Adrian Gonzalo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

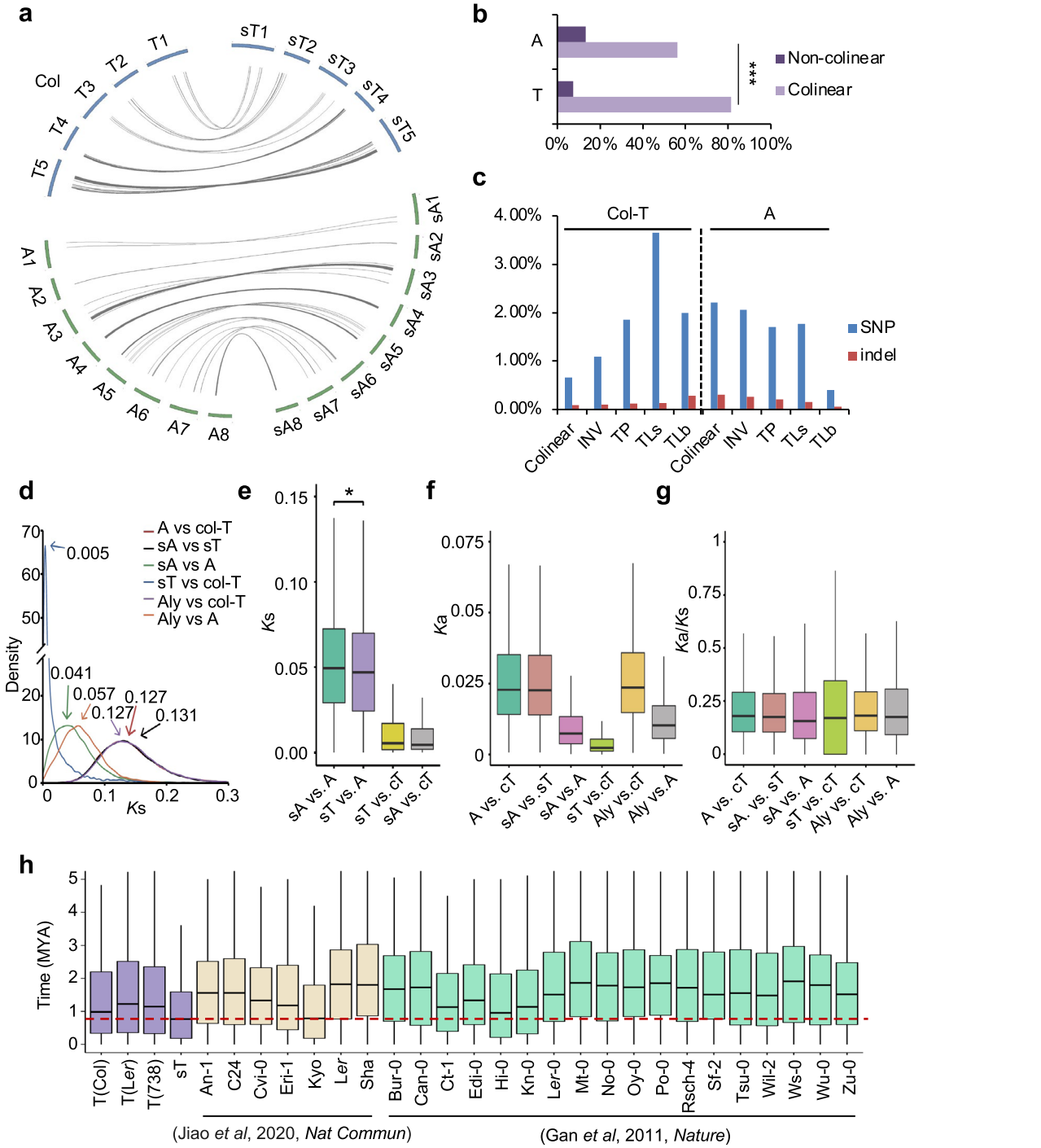
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



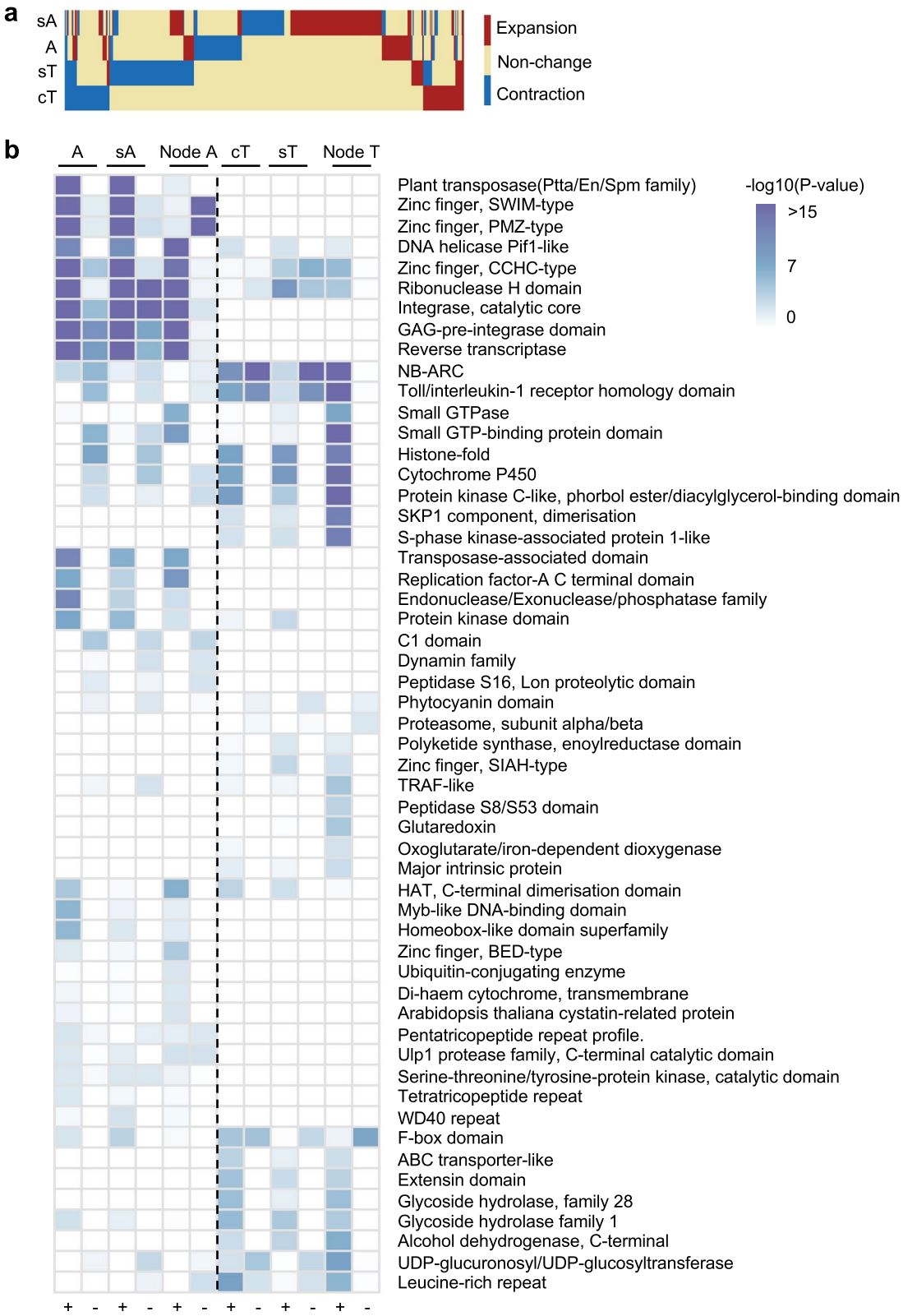
Extended Data Fig. 1 | Synteny and gene content of two subgenome in *A. suecica*. a, Proportion of genes and TEs (left) and TE classes (right) in *A. thaliana* (Col) and in A and T subgenomes of Allo738 (738) and sA and sT subgenome of *A. suecica* (Asu). **b**, Distances of the nearest TE from each gene in sT and sA subgenomes of Asu (upper panel) and in Col (cT) and A subgenome (Aar-related) of Allo738 (lower panel). **c**, **d**, Genomic features of *A. suecica* (**c**) and *A. arenosa* (**d**) genome. (1) Chromosome ideogram; (2) Gene density; (3) TE density; (4)-(6) SNP (4), indel (5) identity, and aligned regions (6) between *A. suecica* and Col or A (**c**) or between A and *A. lyrata* genomes (**d**); (7) synteny of sA and sT homologous gene pairs (**c**) or paralogous gene pairs (**d**). Colours in (1) indicate T (blue) and A (green) or related subgenomes; colour scales in (2) and (3) indicate high (dark purple) to low (white) density; densities (2)-(5) were shown per 100-kb windows; only gene pairs in syntenic blocks spanning 30 genes were shown in (7).



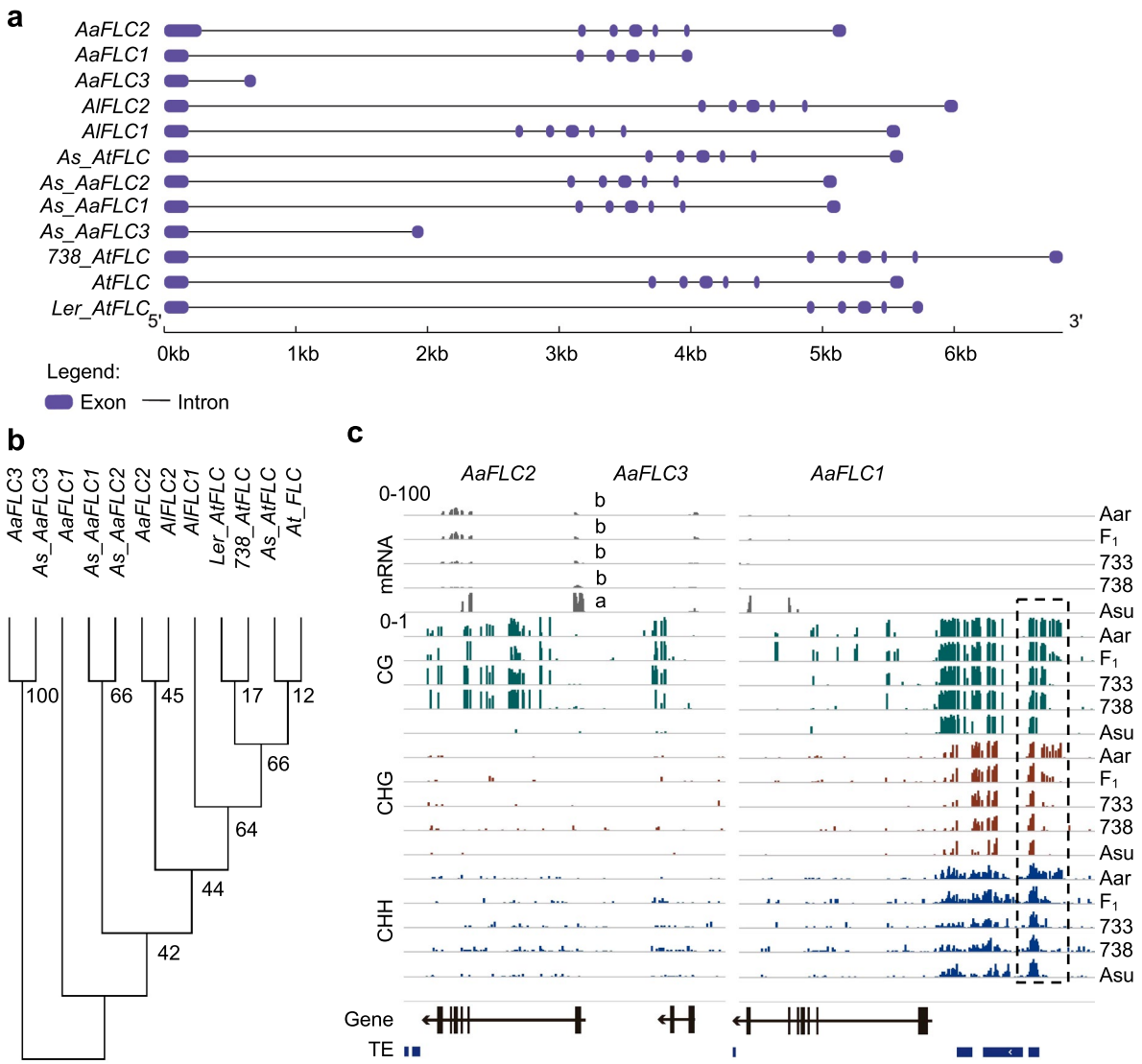
Extended Data Fig. 2 | Analysis of Allo733 and A. suecica genome assemblies using reference genomes. a, Dotplots of allotetraploid Allo738 (738) and *A. suecica* (Asu) assemblies with reference genomes of *A. thaliana* (Ler and Col) and *A. lyrata*, respectively. Two genomes were co-linear (red line) with disruptions (blue lines or dots) and inversions (blue line) or translocations (black circles). **b**, Heatmaps of Hi-C-seq chromosome contacts to show the location of structural variation in different chromosomes.



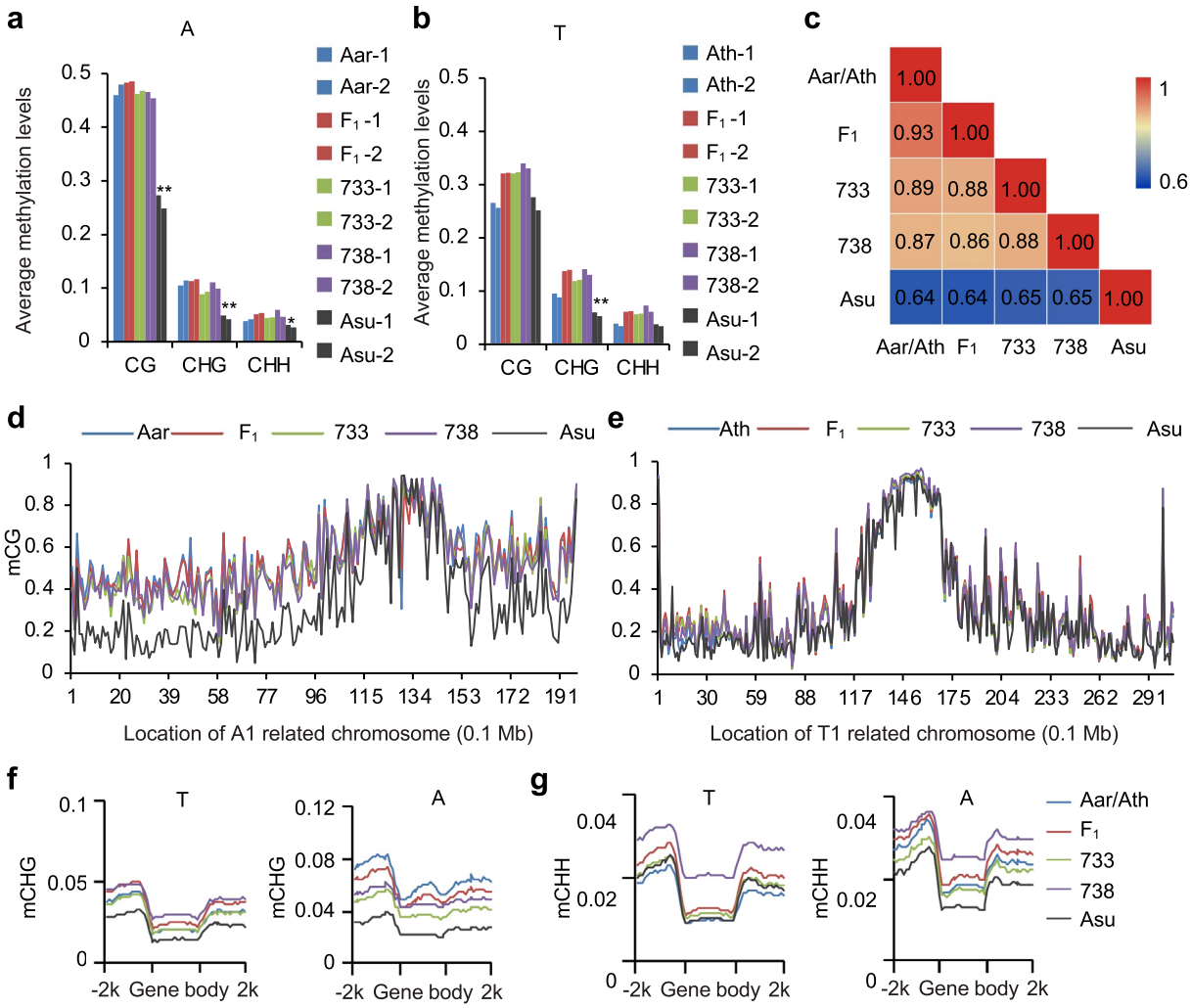
Extended Data Fig. 3 | Genomic variation between *A. suecica*, *A. thaliana*, and *A. arenosa* genomes. a, Inversions between sT of Asu and Col-T (Ath) or between sA of Asu and A (Ara-related) of Allo738. **b**, Proportion of co-linear and non-co-linear regions between A and T. Three asterisks indicate statistical significance level of $P < 0.001$ (Fisher's exact test). **c**, Proportion of SNP and indel distributions in sA and sT subgenomes of Asu relative to A and cT genomes. Non-co-linear: not collinear; INV: inversion; TP: transpositions; TLs: translocations within a subgenome; TLb: translocations between T and A subgenomes. **d**, Distribution of *Ks* values for a set of 14,668 single-copy genes among A subgenome (Aar-related) of 738, cT (Ath, Col), sT and sA subgenomes of Asu, and *A. lyrata* (Aly). **e**, Distribution of *Ks* values between genes in co-linear and TLb regions (co-linear regions: sA vs A, sT vs cT; TLb regions: sT vs A, sA vs cT). One asterisk indicates statistical significance level of $P < 0.05$ (Mann–Whitney U-test). **f**, Distribution of *Ka* values between A and col-T (cT), *A. lyrata* (Aly), and sT and sA subgenomes of Asu. **g**, Distribution of *Ka*/*Ks* values as in (**f**). **h**, Boxplots of the estimated time (million years ago, MYA) for intact LTR insertions in 25 *A. thaliana* ecotypes. The red dashed line indicates the median of time in sT (ANOVA test).



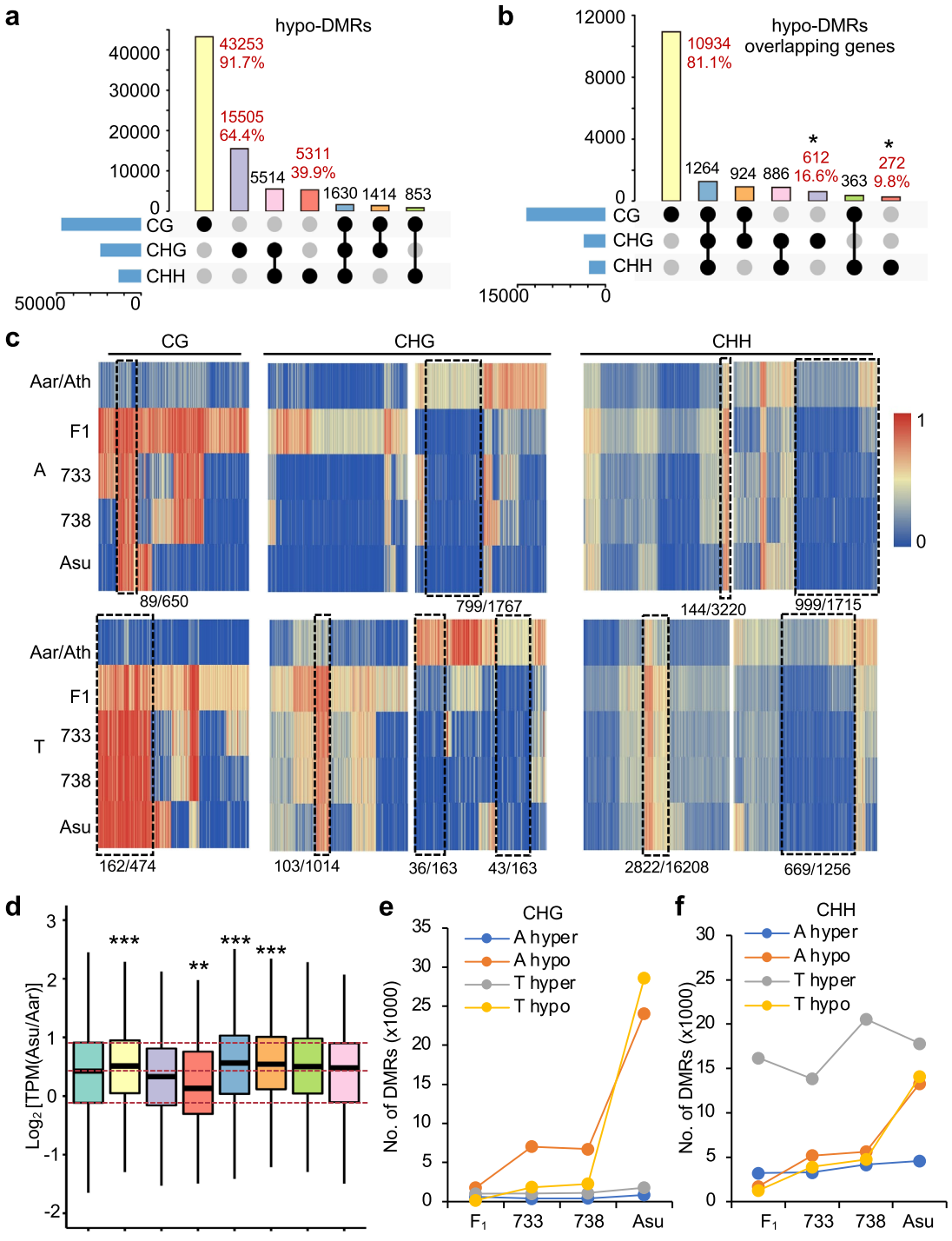
Extended Data Fig. 4 | Expansion and contraction of gene families in A and T-related genomes. a, Heatmap showing expansion and contraction of gene families in A subgenome (Aar-related) of Allo738, Col-T (cT), and sA and sT subgenomes of *A. suecica*. **b**, Domain enrichment of gene family expansion/contraction (+/-) in A, sA, and their nearest ancestor (Node A in Fig. 2c), and cT, sT, and their nearest ancestor (Node T in Fig. 2c) (Fisher's exact test).



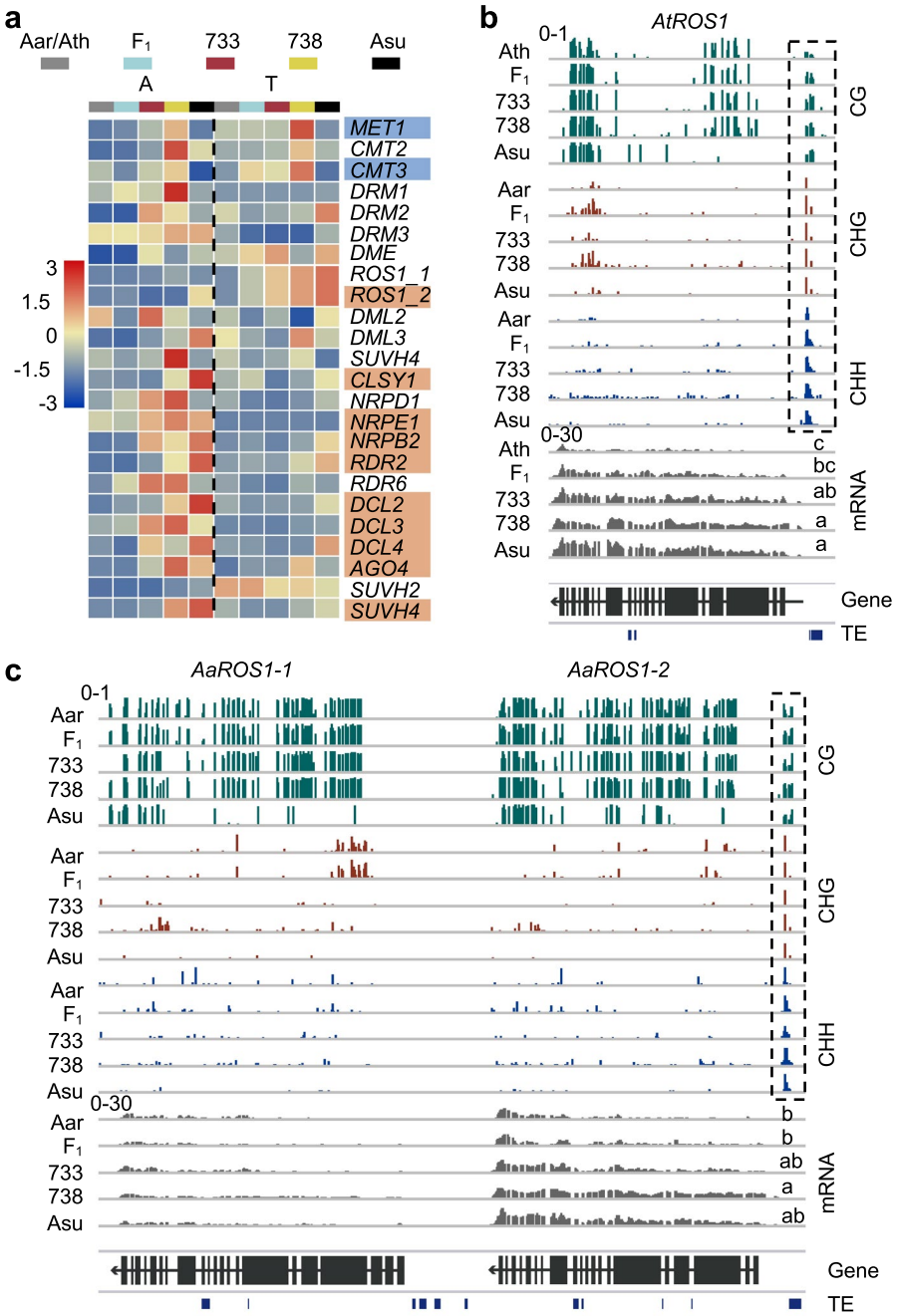
Extended Data Fig. 5 | Comparative genomics of *FLC* loci in *A. suecica* and related species. a, Gene structure of *FLC* in different species: *A. thaliana* (Col, *AtFLC;* Ler_*AtFLC)*, T subgenome of Allo738 (738_*AtFLC*), A subgenome (Aar) (*AaFLC1*, *AaFLC2*, *AaFLC3*), *A. suecica* (*As_AaFLC1*, *As_AaFLC2*, *As_AaFLC3*), *A. lyrata* (*AlFLC1* and *AlFLC2*). **b**, Phylogenetic tree of *FLC* genes. **c**, CG, CHG and CHH methylation and mRNA expression patterns of *FLC* genes and their vicinity in A (Aar), F_1, 733, 738 and *A. suecica* (Asu). The y axis scales shown above gBrowse tracks indicate mRNA (1–100) and methylation (0–1) levels. Differentially methylated regions are shown in a dashed box. Shown below are diagrams of three *FLC* loci (arrows indicating transcription direction) and TEs. Different letters in mRNA (TPM) indicate statistical significance of $P < 0.05$ (ANOVA test, $n = 3$).



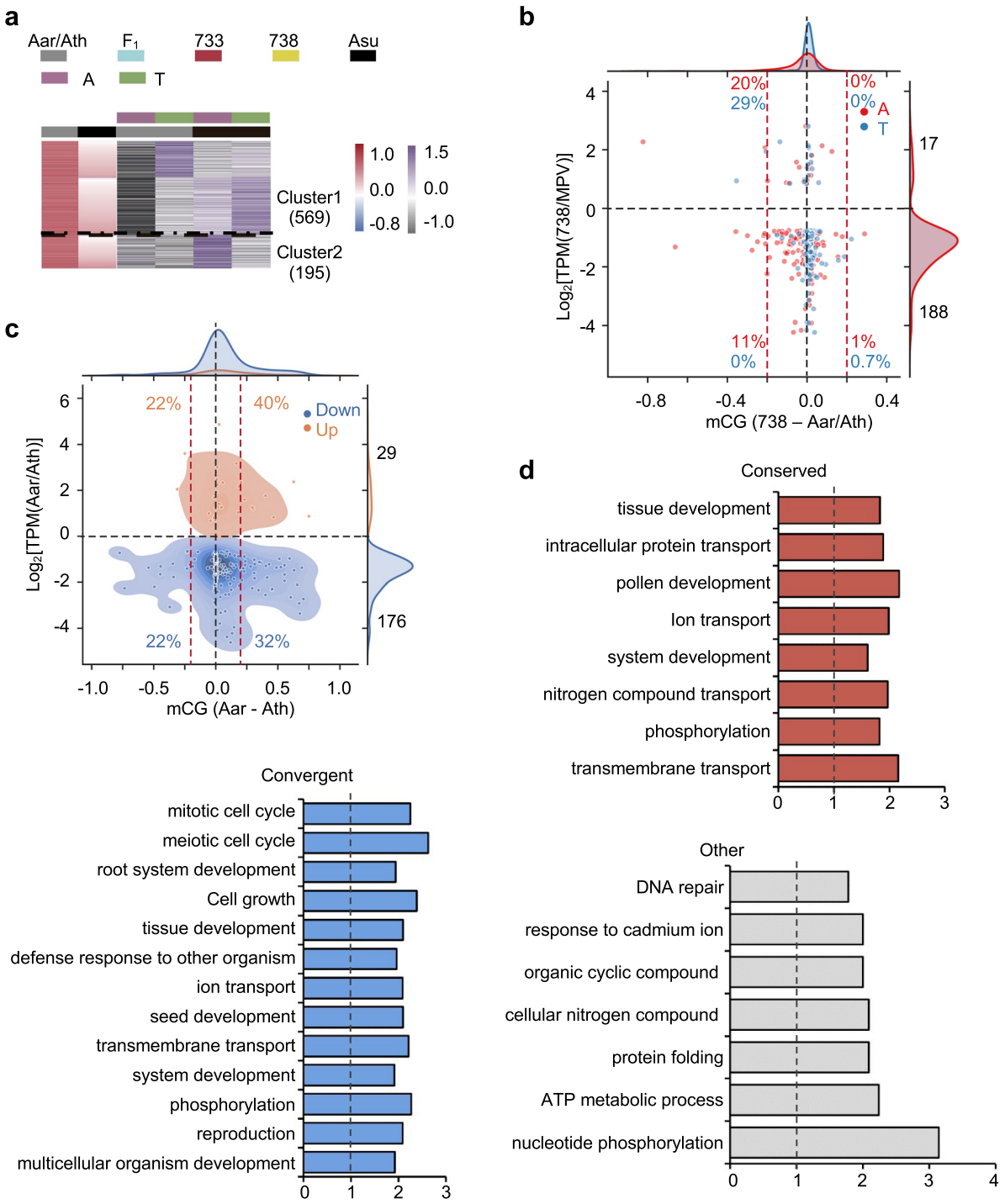
Extended Data Fig. 6 | DNA methylation levels in F_1, Allo733 (733), Allo738 (738) and *A. suecica* (Asu) allotetraploids and their related progenitors (Aar and Ath Ler4). a, **b**, Average methylation levels with two biological replicates in A (**a**) and T (**b**) related genomes or subgenomes in allotetraploids. Two asterisks indicate statistical significance level of $P < 0.01$ (Mann–Whitney U-test). **c**, Heatmap of pairwise comparisons between correlation coefficients of methylated cytosines in CG context of Aar, Ath Ler4 (T), and respective subgenomes of F_1, 733, 738 and Asu. **d**, **e**, CG methylation levels in A1 (**d**) or T1 (**e**) related chromosomes of *A. arenosa* (Aar), *A. thaliana* (Ath Ler4), F_1, Allo733 (733), Allo738 (738), and *A. suecica* (Asu). **f**, **g**, CHG and CHH methylation levels in genic regions of Ath Ler4 (T) and Aar (A) relative to respective subgenomes in allotetraploids (F_1, Allo733, Allo738, and Asu).



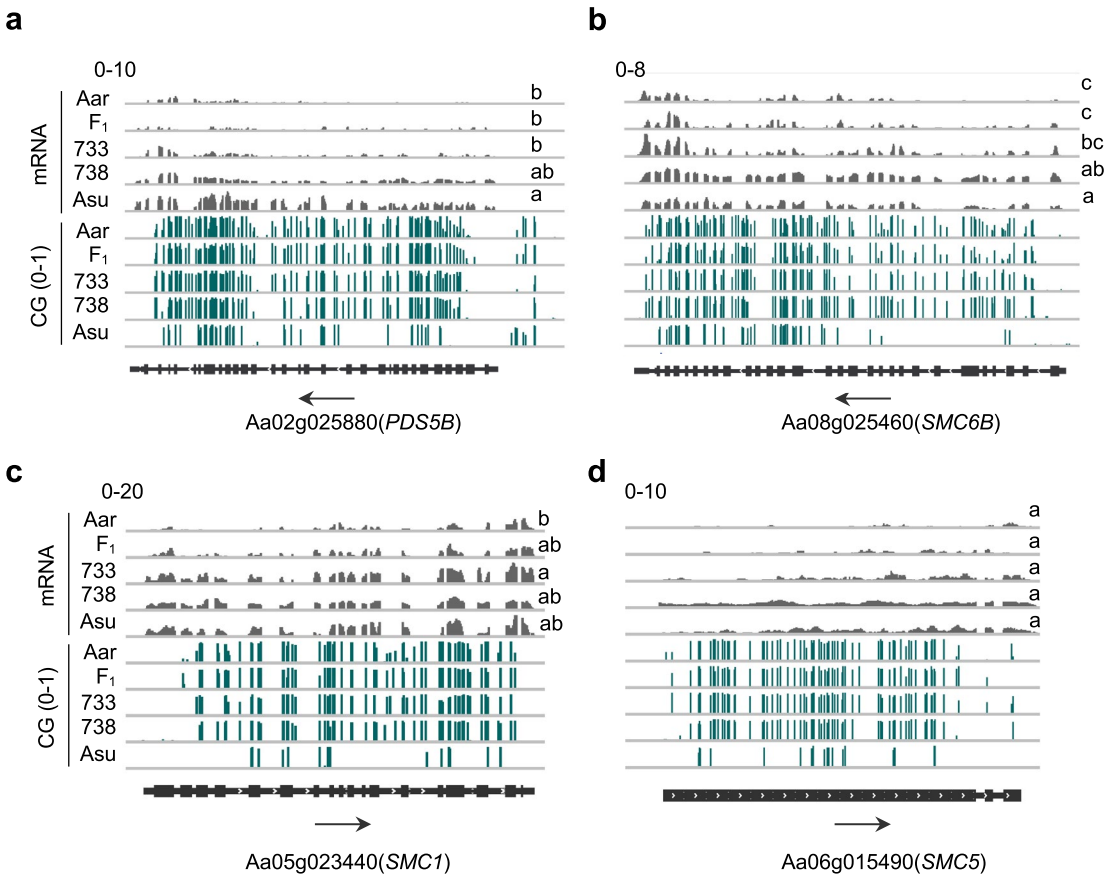
Extended Data Fig. 7 | Association of gene expression with hypo-DMRs between sA subgenome of *A. suecica* and *A. arenosa* (Aar, A) and between sT subgenome and *A. thaliana* Ler4 (Ath, T). a, Upset diagram of hypo-DMRs in CG, CHG and CHH context between sA and A. Numbers and percentages specific to each context were shown in red. **b**, Upset diagram of hypo-DMR-overlapping genes in CG, CHG and CHH context between sA and A. An asterisk indicates the fraction of the unique CHG or CHH DMRs (unique number / total number of DMRs) compared to that of their unique overlapping genes (unique number / total number of associated genes) was significantly reduced ($P < 0.05$, Fisher's exact test). **c**, Heatmap of CG, CHG, and CHH DMRs (sT-T or sA-A) in Aar/Ath, F_1, 733, 738, and Asu. Boxed regions show maintenance of DMRs from F1 to Allo733, Allo738, and *A. suecica*. **d**, Expression ratio (Log2[TPM(Asu/Aar)] of the genes as shown in (**b**). Colours (from left to right) indicate all genes (pale blue) and the genes flanked (2 kb) with hypo-DMRs in CG only (yellow), CHG only (purple), CHH only (red), all three (dark blue), both CG and CHG (orange), both CHG and CHH (green), and both CHH and CG (pink). One and three asterisks indicate statistical significance levels of $P < 0.05$ and $P < 0.001$, respectively (Mann–Whitney U-test). **e**, **f**, Number of CHG (**e**) and CHH (**f**) differentially methylated regions (DMRs) (sT-T or sA-A) in F_1, 733, 738, and Asu, respectively.



Extended Data Fig. 8 | Differential expression of methylation pathway genes including *AtROS1*, *AaROS1-1* and *AaROS1-2* in allotetraploids. a, Heatmap of transcript levels (TPM: transcripts per kilobase per million) of methylation pathway genes in A and T subgenomes of Arabidopsis allotetraploids, respectively. From parents (Aar/Ath) to *A. suecica* (Asu), the upregulated genes was marked orange, while down-regulated genes was marked blue. **b**, **c**, gBrowse tracks showing CG (green), CHG (red), and CHH (blue) methylation and mRNA expression (grey) levels in *AtROS1* (**b**) and *AaROS1-1* and *AaROS1-2* (**c**) genes and their vicinity in *A. thaliana* (Ath) and *A. arenosa* (Aar), F_1, 733, 738 and *A. suecica* (Asu). The regions associated with a TE in the 5′ sequence are shown in a dashed box. The y axis scales shown above the gBrowse tracks indicate mRNA (0–30) and methylation (0–1) levels. Different letters in mRNA (TPM) indicate statistical significance of $P < 0.05$ (ANOVA test, $n = 3$).



Extended Data Fig. 9 | Gene ontology (GO) enrichment terms for hypo-DMR-associated genes and association of CG methylation with non-additive gene expression in allotetraploids. a, The first two columns show CG DMR (A-T) levels (difference threshold > 0.5) of homologous genes in *A. thaliana* (*Ler*4, T) and *A. arenosa* (Aar/Ath) and in *A. suecica* (Asu). The remaining four columns indicate expression levels (TPM) of these genes in Aar/Ath (*Ler*4 and Aar) and Asu, respectively. **b**, Correlation between expression fold changes of shared differentially expressed genes (DEGs) between Aar and Ath (Wang *et al*., 2016) (y axis) and methylation differences (x axis). **c**, Correlation between expression fold changes of shared DEGs between Allo738 and MPV (Wang *et al*., 2016) (y axis) and methylation differences between A (red) and T (blue) subgenomes in Allo738 against Aar/Ath (x axis). The red dashed line indicates 0.2 value of methylation difference; the percentage statistics indicates the fraction of genes with more than 0.2 methylation differences in each quadrant. The genes used in (**b**) and (**c**) were the shared DEGs of Aar vs. *Ler*4 and Allo738 vs. MPV. **d**, GO term overrepresentation for the genes showing conserved (red), convergent (blue), and other (grey) CG hypo-DMRs between sA of *A. suecica* and *A. arenosa* (Aar, A). Dashed line indicates onefold enrichment.



Extended Data Fig. 10 | Association of CG methylation with expression of reproduction-related genes in *Arabidopsis* allotetraploids. a-d, CG methylation near genic regions of *PSD5B* (**a**), *SMC6B* (**b**), *SMC1* (**c**), and *SMC5* (**d**) and their mRNA expression patterns in *A. arenosa* (Aar), F_1, Allo733 (733), Allo738 (738), and natural *A. suecica* (Asu). Black arrow indicate the orientation of gene. PDS5B: One of 5 PO76/PDS5 cohesion cofactor orthologs of *Arabidopsis*; SMC: STRUCTURAL MAINTENANCE OF CHROMOSOMES. Scales indicate mRNA (0–10 and 0–8) and CG methylation density (0–1) levels. Different letters in mRNA (TPM) indicate statistical significance of $P < 0.05$ (ANOVA test, $n = 3$).

nature research

Corresponding author(s): Z. Jeffrey Chen

Last updated by author(s): Feb 20, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection

(1) DNA sequencing was performed using Illumina X10, PacBio - SEQUEL and corresponding software from the manufacturers.
(2) RNA-seq data were generated using Illumina X10 (2X150 bp paired-end reads) and its software.
(3) Methylome (MethylC seq) data were generated using paired-end sequencing using Illumina X10.
(4) Hi-C sequencing was performed using Illumina X10 (2X150 bp paired-end reads), and reads were mapped using Juicer (v1.6.2).

Data analysis

(1) Assembly and annotation: We used MECAT (v1.0), ARROW (v5.0.1.9585), Pillon (v1.22), Juicer (v1.5.6), 3D-DNA (v180114) and Juicebox (v1.9.0) for genome assembly. Following tools were used for genome annotation: Augustus (v3.2.2); TransDecoder (v5.3.0); PASA (v2.3.3); Exonerate (v2.2.0); EVidenceModeler (v1.1.1); InterProScan (v 5.32-71.0); RepeatModeler (v1.0.11); Repeatmasker (v4.0.7); LTR_retriever (v2.0); LTR-FINDER (v1.07); infernal (v1.1.2); tRNAscan-SE (v2.0).

(2) Assessment of genome completeness: We evaluated the genome assembly completeness by BUSCO (v3.0.2) and the accuracy of the assembly through whole-genome alignment against the reference genome of A. thaliana (TAIR10) or A. lyrata (Alyrata_384_v2.1) by MUMmer (v4.0.0beta2). Genome comparisons using HiC data: HiC libraries A. suecica (Asu) and A. thaliana x A. arenosa (Allo738) were aligned to published Ath and Aly reference genomes using BWA-MEM. Heatmaps were generated using the JUICER-pre command, and visualized using JUICEBOX. Inversions and rearrangements were further identified using JUICEBOX.

(3) Analysis of chromosomal collinearity, structural rearrangements and gene family composition between A. suecica and the combination of its assumed progenitors, A. thaliana and A. arenosa: Ath (TAIR10) and Aar (A subgenome of Allo738) assemblies were aligned to the Asu assemblies generated in this study using MUMmer with parameters (nucmer --mum -l 50 -c 100 -b 500 -g 100 && delta-filter -l 100 -i 90). The resulting alignments were used to identify structural rearrangements and local variations using SyRI. Synthetic blocks were identified by MCscan of jcvi (v0.8.12). The gene copy numbers and gene families between assemblies were identified using OrthoFinder based on all annotated protein coding sequences.

(4) LTR analyses: LTR-FINDER (v1.07) and LTR-harvest (v1.5.10) was used to identify full-length retrotransposons. LTR-retriever was used to

integrate those TEs generated by LTR-finder and LTR-harvest, as well as to predict the TE insertion time using the Arabidopsis mutation rate (r=7x10-9). Box plots of insertion time were generated using ggplot2 in R.

(5) Analysis of orthologs and homoeologs: We used diamond (v0.9.24) and OrthoFinder (v2.2.7) to identify homoeologous and orthologous sequences. GO functional enrichment analysis was performed using the clusterProfiler R package.

(6) The non-synonymous/synonymous (Ka/Ks) values estimate: The 14,668 orthologs pairs of each Arabidopsis species were used for estimating Ka/Ks values by KaKs_Calculator (v1.2).

(7) Evolutionary analysis: We used OrthoFinder (v2.2.7), RAxML (v8.2.11), r8s (v1.81) and CAFE (v4.2.1) for phylogenetic analysis and contraction and expansion of gene families estimates. Domain enrichment analysis of contraction/expansion gene families using a Fisher's-exact-test and FDR correction for multiple test.

(8) RNA-seq analysis of homoeolog expression: To exclude expression bias between Ath and Aar due to depth difference, reads of Ath and Aar were down-sampled to the same level and combined. Reads of Ath, Aar, F1, Allo733, and A. suecica were mapped to the Allo738 genome using HISAT2 (v2.1.0) (--score-min L, 0.0,-0.4). Reads of Allo738 were mapped to the Allo738 genome using HISAT2 with default parameters. Only uniquely mapped reads were kept for further analysis. The expression level of each gene was calculated using StringTie (v1.3.3b).

(9) MethylC seq analyses: MethylC-seq reads of Asu and Allo738 were mapped to the Asu and Allo738 genome using Bismark (v0.15.1) with parameters (--score_min L,0,-0.2), respectively. MethylC-seq reads of Ath, Aar, F1, Allo733 were mapped to the Allo738 genome using Bismark with parameters (--score_min L,0,-0.4). To remove bias, only the conserved cytosines were used for downstream analyses using custom Python scripts. To identify conserved regions of 1 kb or longer in A. suecica and Allo738, we aligned the Asu genome against the Allo738 genome by LAST (v869), swapped the sequences and extracted the best alignments. Finally, alignments with scores less than 1000 were removed. The same method was used to identify the conserved region and conserved cytosines between the A and T subgenomes. Differentially methylated regions (DMRs) between the T subgenome and Ath or between the A subgenome and Aar were analyzed using 100-bp sliding windows, including four or more cytosines for CG and CHG contexts and sixteen or more cytosines for CHH context. The weighted methylation level was calculated for each window. Significant differences were assessed using Fisher's-exact-test and FDR correction for multiple test (FDR<0.05), using the following cut-off values of the methylation levels: 0.5 for CG DMRs, 0.3 for CHG DMRs, and 0.1 for CHH DMRs.

(10) Variation calling and Phylogenetic analysis: The paired-end resequencing reads of 39 A. arenosa and 15 A. suecica were downloaded from PRJNA309923 and PRJNA284572 in NCBI Short Reads Archive. The downloaded reads and the reads of Asu, Allo733 and Allo738 were filtered using Trimmomatic (version 0.39). The clean reads of A. arenosa were mapping to the Aar assembly and reads of A. suecica, Allo733 and Allo738 were mapping to the combination of Aar and TAIR10 assembly by BWA program (version 0.7.17-r1188). Only uniquely mapped paired reads were used for the detection of genetic variation and remove PCR duplicates using Picard (version 2.18.15). Variation was called through the Genome Analysis Toolkit (GATK, version 4.1.3.0). Finally, we generate variants of A genome and T genome separately. The variants of 1035 individuals and the variants of T subgenome of A. suecica, Allo733 and Allo738 were merged to the final variants file of T genome. The independent SNPs from A genome with MAF>0.05, missing rate >0.05 were filtered by PLINK (version 1.9). While for SNPs of T genome were filtered same with A genome except the missing rate > 0.02. The filtered SNPs were used to construct phylogenetic trees using the Neighbor-Join method in TASSEL (version 5.0).

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

SUBID	BioProject	BioSample	Accession	Organism
SUB8369755	PRJNA669593	SAMN16534086	JAEFBK000000000	Arabidopsis arenosa x Arabidopsis thaliana (Allo738)
SUB8369755	PRJNA669593	SAMN16534085	JAEFBJ000000000	Arabidopsis suecica (As)
SUB8902864	PRJNA669593	SAMN17369459	JAESVC000000000	Arabidopsis arenosa x Arabidopsis thaliana (Allo733)
SUB8323092	PRJNA669593	SAMN16456086	SRR12880892	Allo738_seedling_RNA-seq
SUB8323092	PRJNA669593	SAMN16456085	SRR12880893	As_pod_RNA-seq
SUB8323092	PRJNA669593	SAMN16456084	SRR12880894	As_flower_RNA-seq
SUB8323092	PRJNA669593	SAMN16456083	SRR12880895	As_seedling_RNA-seq
SUB8323092	PRJNA669593	SAMN16456082	SRR12880896	Allo738_HiC-seq
SUB8323092	PRJNA669593	SAMN16456081	SRR12880897	As_HiC-seq
SUB8323092	PRJNA669593	SAMN16456080	SRR12880898	Allo738_DNA-seq
SUB8323092	PRJNA669593	SAMN16456103	SRR12880899	As_BS-seq_rep2
SUB8323092	PRJNA669593	SAMN16456102	SRR12880900	As_BS-seq_rep1
SUB8323092	PRJNA669593	SAMN16456101	SRR12880901	Allo738_BS-seq_rep2
SUB8323092	PRJNA669593	SAMN16456100	SRR12880902	Allo738_BS-seq_rep1
SUB8323092	PRJNA669593	SAMN16456099	SRR12880903	Allo733_BS-seq_rep2
SUB8323092	PRJNA669593	SAMN16456098	SRR12880904	Allo733_BS-seq_rep1
SUB8323092	PRJNA669593	SAMN16456097	SRR12880905	F1_BS-seq_rep2
SUB8323092	PRJNA669593	SAMN16456079	SRR12880906	As_DNA-seq
SUB8323092	PRJNA669593	SAMN16456096	SRR12880907	F1_BS-seq_rep1
SUB8323092	PRJNA669593	SAMN16456095	SRR12880908	Aa_BS-seq_rep2
SUB8323092	PRJNA669593	SAMN16456094	SRR12880909	Aa_BS-seq_rep1

SUB8323092 PRJNA669593 SAMN16456093 SRR12880910 At_BS-seq_rep2
SUB8323092 PRJNA669593 SAMN16456092 SRR12880911 At_BS-seq_rep1
SUB8323092 PRJNA669593 SAMN16456091 SRR12880912 Allo738_leaf_RNA-seq_rep3
SUB8323092 PRJNA669593 SAMN16456090 SRR12880913 Allo738_leaf_RNA-seq_rep2
SUB8323092 PRJNA669593 SAMN16456089 SRR12880914 Allo738_leaf_RNA-seq_rep1
SUB8323092 PRJNA669593 SAMN16456088 SRR12880915 Allo738_pod_RNA-seq
SUB8323092 PRJNA669593 SAMN16456087 SRR12880916 Allo738_flower_RNA-seq
SUB8323092 PRJNA669593 SAMN16456078 SRR12880917 Allo738_PacBio
SUB8323092 PRJNA669593 SAMN16456077 SRR12880918 As_PacBio
SUB8902848 PRJNA669593 SAMN17371748 SRR13452155 Allo733_PacBio
SUB8902848 PRJNA669593 SAMN17371749 SRR13452154 Allo733_DNA-seq
Note: Assemblies are still in manual review and will be released under those accession numbers.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Sample size per group or condition was determined based on the minimum number of biological replicates required to perform differential expression and methylation analysis as per software tools used and previously published literature.
Data exclusions	Samples were excluded if they failed at the library preparation stage or those that displayed poor correlation between biological replicates.
Replication	Findings were consistent between biological replicates and different sequencing plates/batches. The replications of methylation data were merged to increase coverage.
Randomization	Order of sample processing for library preparation and sequencing were processed in multiple batches as and when they were received from collaborating laboratories, kind of randomization in itself, but following stringent standardized protocols.
Blinding	No blinding took place. To alleviate any complications from non-blinded analyses all samples were analyzed simultaneously in the same manner regardless of their condition/origin. All specimens' identities were encoded before submission for genotyping.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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