Title

Structural variation reshapes population gene expression and trait variation in 2,105 Brassica napus accessions.

Journal

Nature Genetics, 56(11)

Authors

Zhang, Yuanyuan

Yang, Zhiquan

He, Yizhou

et al.

Publication Date

2024-11-01

DOI

10.1038/s41588-024-01957-7

Peer reviewed

nature genetics

Article

https://doi.org/10.1038/s41588-024-01957-7

Structural variation reshapes population gene expression and trait variation in 2,105 *Brassica napus* accessions

Received: 19 April 2023

Accepted: 23 September 2024

Published online: 5 November 2024

Check for updates

Yuanyuan Zhang[1,7], Zhiquan Yang[2,3,7], Yizhou He[1], Dongxu Liu[2], Yueying Liu[1], Congyuan Liang[2], Meili Xie[1], Yupeng Jia[2], Qinglin Ke[1], Yongming Zhou[2], Xiaohui Cheng[1], Junyan Huang[1], Lijiang Liu[1], Yang Xiang[4], Harsh Raman[5], Daniel J. Kliebenstein[6], Shengyi Liu[1] ✉ & Qing-Yong Yang[2] ✉

Although individual genomic structural variants (SVs) are known to influence gene expression and trait variation, the extent and scale of SV impact across a species remain unknown. In the present study, we constructed a reference library of 334,461 SVs from genome assemblies of 16 representative morphotypes of neopolyploid *Brassica napus* accessions and detected 258,865 SVs in 2,105 resequenced genomes. Coupling with 5 tissue population transcriptomes, we uncovered 285,976 SV-expression quantitative trait loci (eQTLs) that associate with altered expression of 73,580 genes. We developed a pipeline for the high-throughput joint analyses of SV-genome-wide association studies (SV-GWASs) and transcriptome-wide association studies of phenomic data, eQTLs and eQTL-GWAS colocalization, and identified 726 SV–gene expression–trait variation associations, some of which were verified by transgenics. The pervasive SV impact on how SV reshapes trait variation was demonstrated with the glucosinolate biosynthesis and transport pathway. The study highlighting the impact of genome-wide and species-scale SVs provides a powerful methodological strategy and valuable resources for studying evolution, gene discovery and breeding.

SVs such as deletions, insertions, duplications and inversions represent a potentially important source of genetic diversity that can control phenotypic variation and enable organismal adaptation to diverse environments[1]. Recent studies have shown that SVs can cause large-scale perturbations of regulatory regions and account for a greater proportion of genomic variation than SNPs[2,3]. Studies on individual genes have shown *cis*- and *trans*-regulatory mechanisms of SVs on gene expression[2,4,5]. Further studies on three-dimensional (3D) interactions of chromosomes have indicated that SVs can alter higher-order chromatin organization to affect the expression of neighboring genes in humans[6,7]. In polyploid plants, large SVs frequently occur after genome duplication and can contribute to genome differentiation by removing genes[8]. Therefore, SVs may have a greater impact on polyploid genomes and accompanying trait innovation[9–11]. However, the extent and scale of genome-wide SV impact on gene expression, especially in polyploids, remain unknown. Previous studies have largely focused on

[1]The Key Laboratory of Biology and Genetic Improvement of Oil Crops, the Ministry of Agriculture of PRC, Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, China. [2]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan, China. [3]Innovative Center of Molecular Genetics and Evolution, School of Life Sciences, Guangzhou University, Guangzhou, China. [4]Guizhou Rapeseed Institute, Guizhou Academy of Agricultural Sciences, Guiyang, China. [5]NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, New South Wales, Australia. [6]Department of Plant Sciences, University of California-Davis, Davis, CA, USA. [7]These authors contributed equally: Yuanyuan Zhang, Zhiquan Yang. ✉e-mail: liusy@oilcrops.cn; yqy@mail.hzau.edu.cn

https://doi.org/10.1038/s41588-024-01957-7

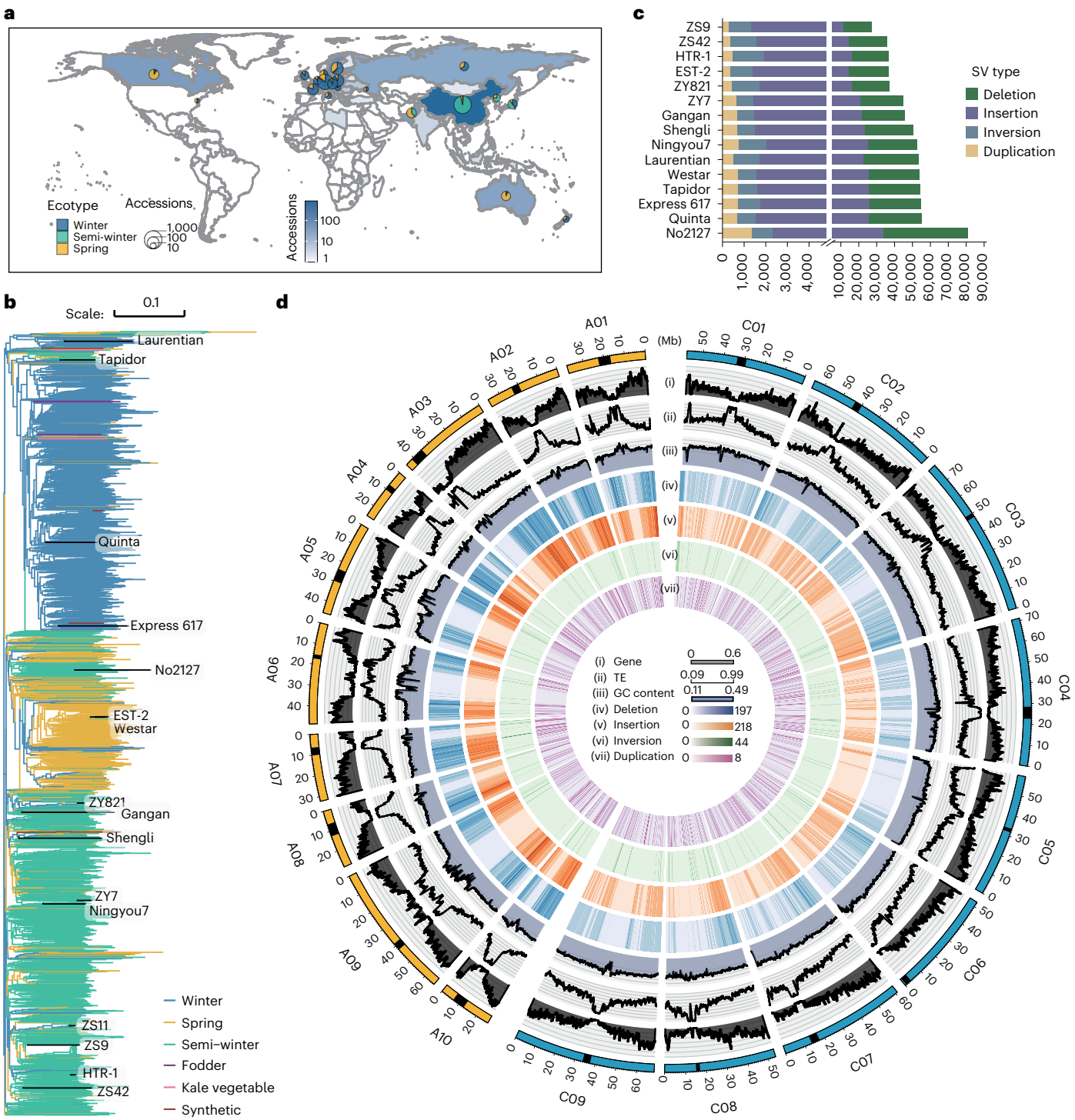


Fig. 1 | Identification and characterization of the *B. napus* pan-SV. **a**, Geographic distribution of 2,105 *B. napus* accessions that were sequenced. The map was created using the map data function in the ggplot2 package. **b**, Phylogenetic analysis of 2,105 accessions based on SNPs. Line colors represent three rapeseed ecotypes, another two botanical varieties (var. *pabularia* and var. *napobrassica*) and resynthetics developed from hybridization between *B. rapa* and *B. oleracea*. The 16 representative *B. napus* accessions, including one rutabaga (swede) root fodder (Laurentian) and one resynthetic (No2127) for de novo assembling are indicated as black lines. **c**, The SV types and numbers of the 15 *B. napus* genome assemblies based on the reference cv. ZS11 genome. In **b** and **c**, ZS11 is for Zhongshuang11, ZS9 for Zhongshuang9, ZY821 for Zhongyou821 and ZY7 for Zheyou7. **d**, Distribution feature of SVs from 2,105 *B. napus* accessions. Different tracks (i–vii) indicate the densities of genes, TE and GC content (i–iii) or abundance of SVs (iv–vii).

studying associations between SNPs and gene expression quantity, namely eQTLs (SNP-eQTLs)[12,13].

A problem confronting the population analysis of SV impact on gene expression and trait variation is the unreliability of SV identification using short-read sequencing[1]. In polyploids, short-read sequences from a large number of similar genomic segments can cross-map on to multiple genomic sites[14]. Consequently, the vast majority of SVs have remained hidden, along with their impact on genomic and phenotypic diversity[1,14]. Therefore, a genomic resource of SVs and associated phenotypic variation is required at a species scale to empower our understanding of SV impact.

Neopolyploid *B. napus* rapeseed/canola (AACC, $2n = 4x = 38$) is an important crop for healthy edible oil, stockfeed and biofuel markets worldwide. Previous studies have identified small insertion/deletion (indel) variations and their associations with traits in the rapeseed accession population[15], as well as SVs from eight genotypes and the corresponding variation in presence/absence associated with three traits in a nested association mapping population[16]. However, there is no report on the SV regulatory effect of large accession populations on gene expression and trait variation.

The present study aimed to explore the potential of integrating high-throughput population genomics with gene-editing technologies and reveal the extent and scale of genome-wide SV impact on gene expression, and thereby trait variation of the *B. napus* population comprising 2,105 diverse core accessions. To achieve this, we constructed a reference pan-SV library for reliable detection of SVs from the large resequenced population, thus enabling the establishment of extensive catalogs of SVs together with their regulated genes. Furthermore, coupling the population genomic, transcriptomic and phenomic data, we established a joint analysis method involving SV-GWAS, SV-eQTL, TWAS (transcriptome-wide association study)[17] and GWAS-eQTL colocalization[18] and revealed a substantial number of SV loci with a surprisingly wide regulatory effect on variation of gene expression and agronomically important traits.

Results

Identification and characterization of species-scale SVs

To investigate genome-wide SVs and their impact on gene expression at the *B. napus* species scale, we constructed a high-confidence reference pan-SV library using 16 genomes, 6 assembled in the present study (Supplementary Tables 1–4 and Supplementary Note 1) and 10 published ones[16,19,20]. These 16 accessions were representatives of global *B. napus* diversity, covering both subspecies (*B. napus* subsp. *napus* with three rapeseed ecotypes (spring, winter and semi-winter) and *B. napus* subsp. *rapifera*) and a resynthesized oilseed rape line. They were chosen from the 2,105 accessions collected from various countries (Fig. 1a,b and Supplementary Table 1). The six genomes of ZY821, Laurentian, ZS9, ZS42, HTR-1 and EST-2 were de novo assembled using combined Oxford Nanopore Technologies long reads (79-fold genome coverage) and Illumina short reads (67-fold genome coverage) with multiple assemblers[21], yielding 6 new assemblies with an average contig N50 of 5.18 Mb and an average length of 937.19 Mb (Supplementary Tables 2–4). The cv. Laurentian assembly was the first fodder crop genome available to date to the *Brassica* community.

We identified SVs by comparing the assemblies and long-read alignment to the ZS11 reference genome (Extended Data Fig. 1a). After highly stringent quality filtering, a total of 334,461 high-confidence, nonredundant SVs (>50 bp) were identified (Fig. 1c). Of them, a large inversion of 26.67 Mb was illustrated and verified using Hi-C link patterns and PCR amplification (Extended Data Fig. 1b–d and Supplementary Note 2). Using this reference pan-SV library, we constructed a population SV map of 2,105 *B. napus* accessions resequenced by Illumina HiSeq with an average of 8.6-fold coverage by mapping the short reads on to the reference SVs using the Paragraph package[22]. The precision and recall rate of SVs excluding translocations were 0.84 and 0.91 (Supplementary Table 5), respectively, comparable to human genomes[22]. In total, we identified 258,865 SVs including 125,611 insertions, 124,744 deletions, 6,146 inversions and 2,364 duplications, and their occurrence frequency, genomic distribution pattern and sizes are different, particularly between A_n and C_n (Fig. 1d, Extended Data Fig. 1e–h, Supplementary Tables 6 and 7 and Supplementary Note 3). These characteristics may reflect the distinct evolutionary features of the two subgenomes in neopolyploid *B. napus*.

To assess the contribution of SVs to genomic diversity, we constructed and compared SV- and SNP-based phylogenetic trees of the 2,105 accessions and found apparent differences between the trees (Fig. 1b and Extended Data Fig. 1i), suggesting the independent distribution of SVs and SNPs, both of which contributed to genomic diversity.

Identification and characterization of eQTLs in population

To reveal the extent and scale of SV impact on population gene expression, we conducted SV-eQTL analyses in the two *B. napus* subpopulations. After filtering the 258,865 SVs for those with minor allele frequency (MAF) <0.01 and call rate <0.7, a total of 93,505 high-quality/-confidence SVs were generated. For the SV-eQTL analysis, RNA sequencing (RNA-seq) data were generated from five tissues, shoot apical meristems (SAMs), leaves, siliques at 18 d after pollination (d.a.p.) and developing seeds at 20 d.a.p. and 40 d.a.p., each sample with reads of ~6 Gb (Fig. 2a and Supplementary Table 8). The eQTL mapping was performed using transcripts from a total of 81,424 genes, each expressing in >5% of the accessions of the two subpopulations, against the 93,505 SV genotypes. A total of 285,976 SV-eQTLs was identified (Supplementary Table 9). Among the SVs of polyploid plants, homoeologous nonreciprocal translocation, an interesting kind of homoeologous exchange (HE) detectable by the method previously described[23,24], occurs in *B. napus*[23,24]; however, we could not analyze them in SV-eQTL mapping owing to their low frequencies and some false positives that were difficult to circumvent in the populations (Supplementary Note 4).

In each SV-eQTL region, the highest significantly associated eSV (peak eSV) was defined as the lead eSV and, in total, 47,897 lead eSVs were identified (repeated ones in different tissues were just counted once), which regulated the expression of 73,580 target genes (referred to as eGenes) (Fig. 2b and Supplementary Table 9), accounting for 90% of the total expressed genes (81,424), which represents an unprecedented gene quantity and ratio. Based on the physical distance between lead eSVs and the associated eGenes, we divided these eQTLs into *cis*-eQTLs (≤1 Mb away from eGene) and *trans*-eQTLs (>1 Mb or on different chromosomes), identifying 66,003 *cis*-eQTLs (23%, corresponding to 30,827 lead eSVs) and 219,973 *trans*-eQTLs (77%, corresponding to 39,609 lead eSVs). Of the 47,897 lead eSVs, 17% (8,288) are purely *cis*-eSVs, 36% (17,070) are purely *trans*-eSVs and the remaining 47% (22,539) have both *cis* and *trans* effects. Within a 1-Mb range, the number of *cis*-eSVs decreased with distance from the eGene (Extended Data Fig. 2a). More than half the lead eSVs regulate the expression of more than one gene either in *cis* or in *trans*, and 74% and 43% of eGenes are *trans* and *cis* regulated, respectively, by more than one lead eSV (Fig. 2c,d). Of these, there are 8,631 *trans*-lead eSVs, each of which simultaneously regulates >5 genes, and 10,217 eGenes, each of which is simultaneously regulated by >5 *trans*-lead eSVs, suggesting complex regulatory effects of SVs on gene expression.

In the regulation by *trans*-lead eSVs (Fig. 2e,f), 15% of the total lead eSVs regulate their eGenes on the same chromosome (intrachromosome), whereas 54% regulate interchromosome eGenes. Of the *trans*-lead eSVs on interchromosomes, 44% regulate intra-subgenome eGenes; the remaining 56% (66% on A_n and 34% on C_n) regulate inter-subgenome eGenes and, despite C_n being 1.5× larger than A_n (Supplementary Table 7), the number of lead eSVs on A_n is 1.7× greater (Fig. 2e), indicating remarkably asymmetrical subgenome regulation between A_n and C_n.

We next investigated the genomic distribution of SV-eQTL hotspots that might harbor master regulators affecting a suite of downstream genes[25]. We identified 495 *trans*-eQTL hotspots regulating the expression of 59,914 genes (Fig. 2g and Supplementary Table 10). Enrichment and pathway analyses of eGenes strongly show that each *trans*-eQTL hotspot regulates a set of functionally connected genes or a gene network (Supplementary Table 11), revealing a way to link *trans*-SVs to their specific biological functions or the nature of SV effect.

https://doi.org/10.1038/s41588-024-01957-7

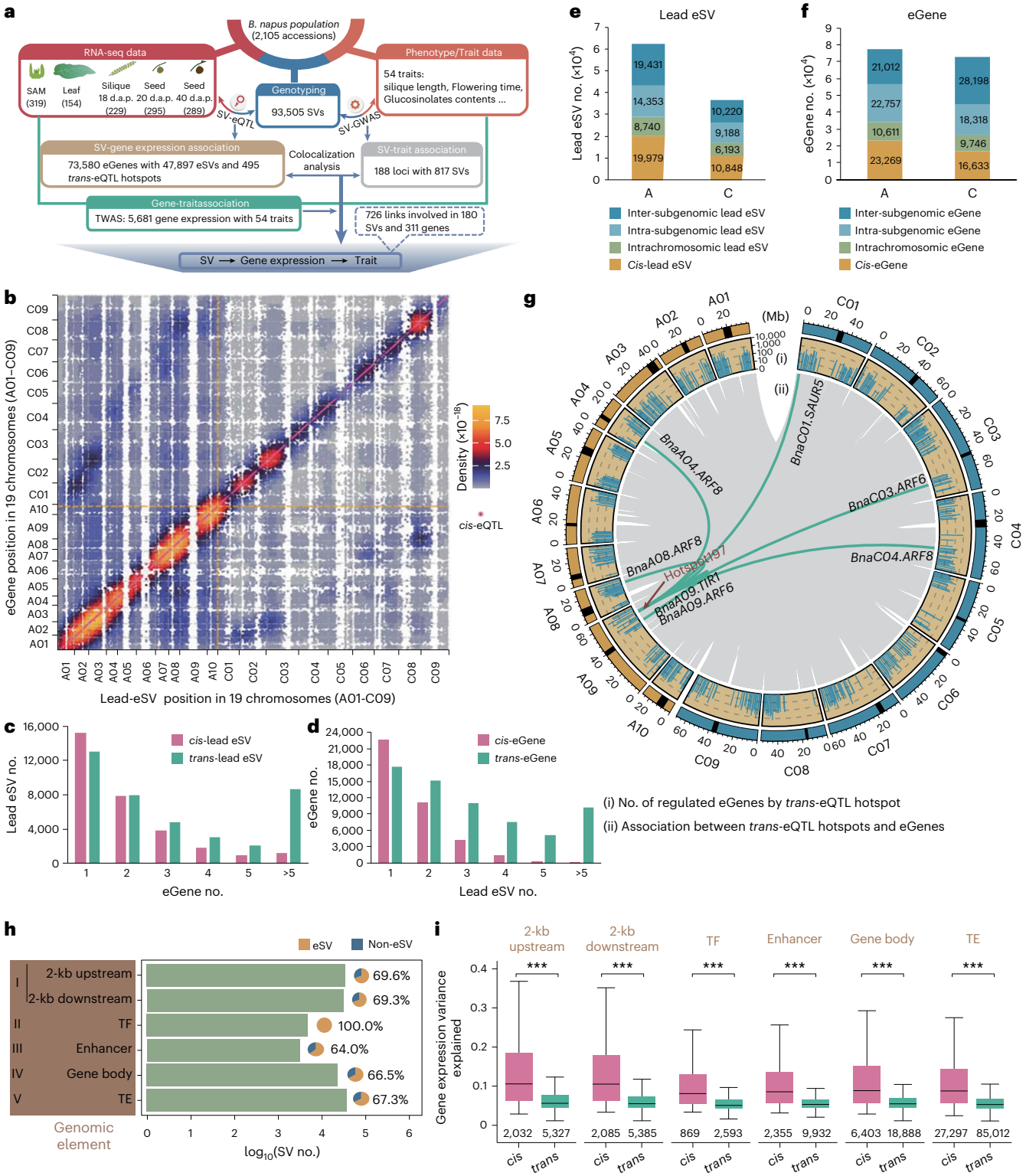


Fig. 2 | SV impact on gene expression and its regulatory mechanisms. **a**, Overview of the study of SV impact on gene expression and trait variation. The number of accessions for each RNA-seq tissue are shown in parentheses. **b**, Heatscatter plot showing the genome-wide distribution of SV-eQTLs. Each dot represents a lead eSV and eGene association. The pink dots along the diagonal represent those for *cis*-lead eSVs. **c**,**d**, Relationships between the numbers of lead eSVs (**c**) and eGenes (**d**). **e**,**f**, The number and proportion of lead eSVs (**e**) and eGenes (**f**) in *cis*- and *trans*-eQTLs in A_n and C_n subgenomes. **g**, Distribution of 495 *trans*-eQTL hotspots and their eGenes. The emanative bars stand for hotspots and their height for the number of eGenes in the mid-circle (i), of which each eSV and its eGene(s) in *trans*-eQTLs are connected by gray lines (ii) and specifically the green lines for the *trans*-eQTL Hotspot-197 that was studied in Extended Data Fig. 6. **h**, The number and proportion of eSVs sorted in terms of their regulatory mechanisms. Each pie chart at the right of each column indicates the ratio of eSVs to all annotated SVs in each category. **i**, Gene expression variance explained by *cis*- and *trans*-eQTLs in individual categories of regulatory mechanisms. All *P* values are based on the two-tailed Wilcoxon's rank-sum tests. The horizontal lines within boxes represent median value, the bottom and top edges of the boxes the 25th (***P value almost 0) and 75th percentile values and the lower and upper whiskers showing the 5th and 95th percentile values (the same for all other boxplots). The number (n) under the box represents the corresponding number of samples in each group.

https://doi.org/10.1038/s41588-024-01957-7

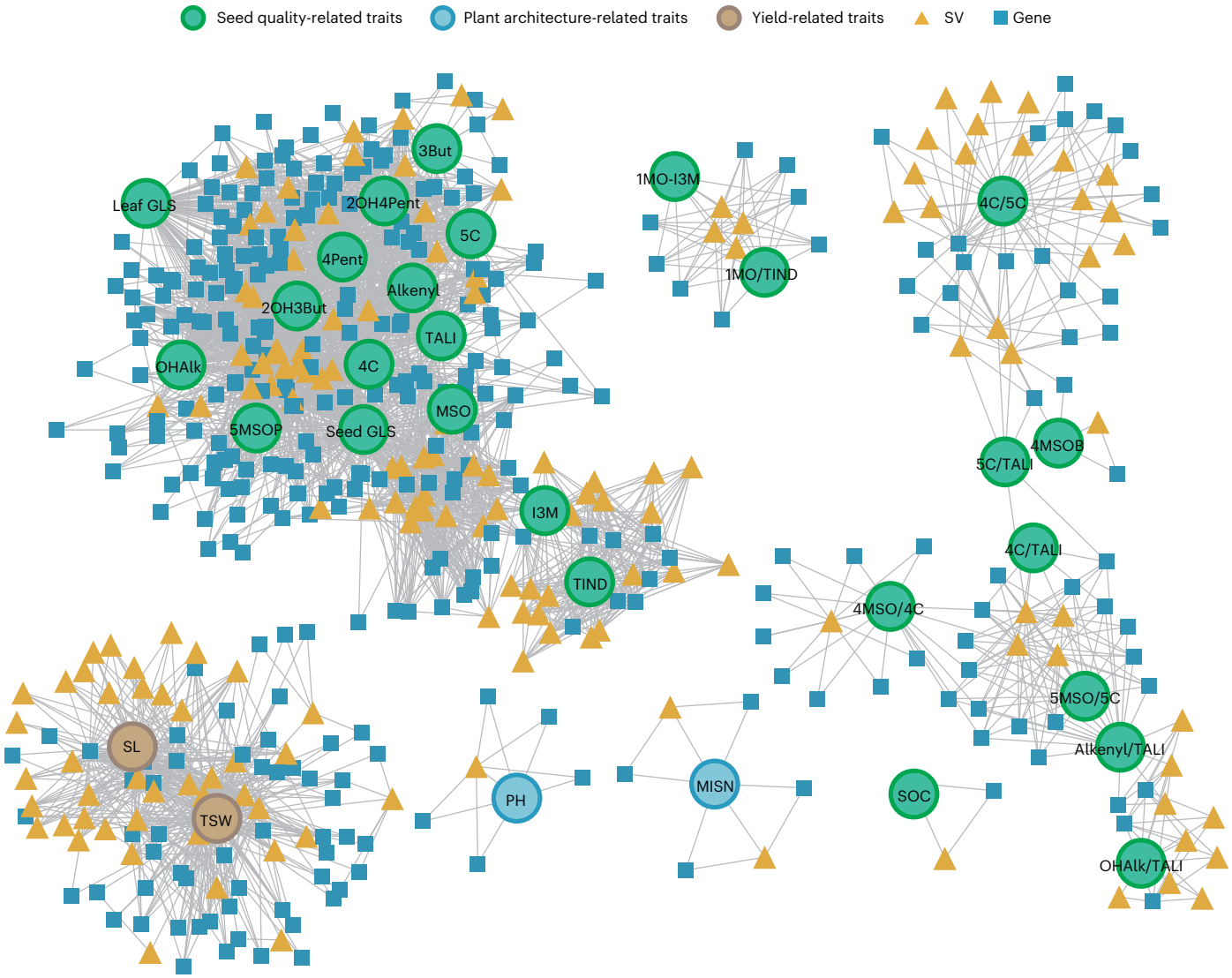


Fig. 3 | Networks illustrating causal SV–gene expression–trait variation association integrated from the data of SV-eQTLs, SV-GWASs and TWASs. See Supplementary Tables 14 and 18 for more details.

The eSV-mediated regulatory mechanisms of gene expression

Based on eSV annotation, sequence characteristics and/or its position in relationship to the eGenes, we classified eSV-mediated regulatory mechanisms of gene expression alteration into eight categories (Fig. 2h, Supplementary Table 12 and Supplementary Note 5). These were (I) SV-mediated changes of regulatory sequences: of 2,485 SVs, 1,476 at 2 kb upstream and, of 2,616 SVs, 1,598 at 2 kb downstream of genes are eSVs, and they putatively regulate 48,264 and 49,765 eGenes, respectively; (II) the SV effect through its transcription factors (TFs): there are 4,865 eSVs significantly associated with 260 TF genes, and another 5,042 eGenes could be matched to the putative upstream 260 TFs by retrieving the previously established *B. napus* 'TF-target gene' database[26] (Supplementary Table 13), suggesting that the eSVs in either *cis* or *trans* regulate expression of some eGenes by affecting the gene expression of TFs that target these eGenes; (III) the SV-changed activity of distal regulatory elements (enhancer); (IV) SV-mediated disruption of gene body; (V) transposable element (TE)-mediated SV effect; and moreover (VI–VIII), the SV effect that could also occur through its epigenetic regulation on gene expression. These eight categories are summarized in Supplementary Table 12 and Supplementary Note 5. In addition, 6,384 eSVs do not fall into the above categories and have no identifiable regulatory components. Overall, the effects of eSVs explain significantly greater gene expression variance in *cis* than in *trans* and those in categories I and V are higher than the others (Fig. 2i, Extended Data Fig. 2b,c and Supplementary Note 5).

These results revealed an intriguing landscape of the genome-wide impact of SVs on gene expression via *cis* and *trans* regulation in large populations and provided new insight into the complex mechanism of gene and trait regulation by SVs.

Identification of SV–gene expression–trait associations

With the above metadata and megadata, we proceeded to test the ability of the joint analyses of SV-eQTL, SV-GWAS, TWAS and eQTL-GWAS colocalization for high-throughput identification of associations linking the SV effect on gene expression to trait variation in the *B. napus* populations (Fig. 2a). This would give an image of the extent and scale of the SV impact on population trait variation.

We first carried out SV-GWASs using 54 sets of phenotypic data, including seed quality, morphology and yield components of the above populations. We identified 817 SVs in 188 loci that were significantly associated with the traits (Fig. 2a and Supplementary Tables 14 and 15). Of them, 686 SVs are lead eSVs for eQTLs associated with 5,084 eGenes; 84 loci are overlapped between eQTL hotspot and GWAS QTLs (Supplementary Table 16), suggesting that eSVs of these eQTL hotspots

https://doi.org/10.1038/s41588-024-01957-7

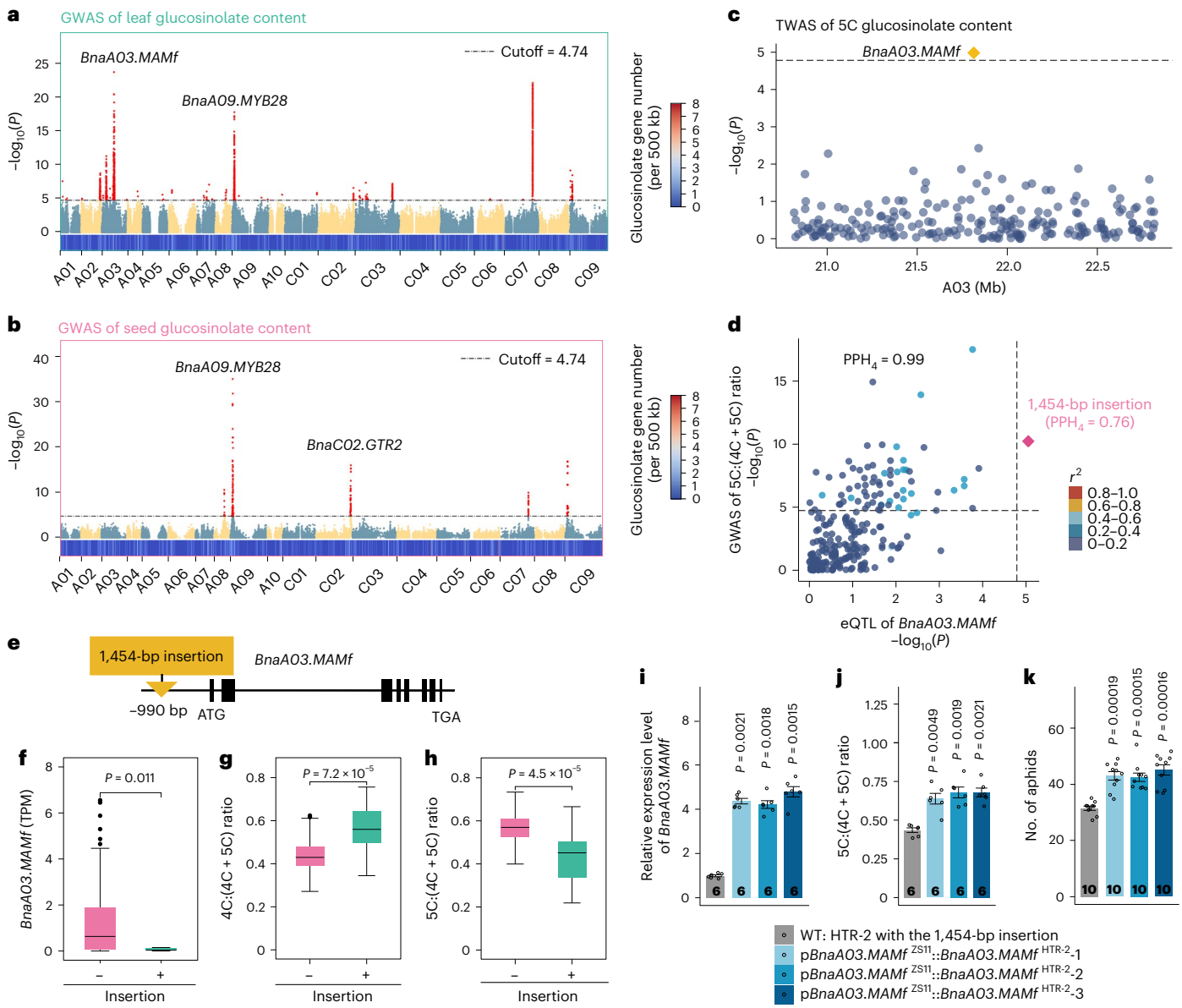


Fig. 4 | An insertion repressing downstream gene expression. **a**,**b**, Manhattan plots presenting composite SV-GWAS loci of the individual and total glucosinolate contents in leaves (**a**) and seeds (**b**). The red dots indicate significant associations and the blue-and-red heatmap above the *x* axis shows density of the genes orthologous to *Arabidopsis* glucosinolate pathway genes. **c**, Local Manhattan plot of TWASs showing a significant association between 5C-glucosinolate (glucosinolate with a five-carbon side-chain) content and the gene expression level of *BnaA03.MAMf*. Each dot represents a gene and the yellow square is a significantly associated eGene. **d**, Colocalization analysis of the loci of an eQTL regulating *BnaA03.MAMf* expression (*x* axis) and GWAS of 5C:(4C + 5C) ratio (*y* axis) in leaves. The black PPH_4 (0.99) indicates the posterior probability of colocalization, whereas the pink PPH_4 (0.76) is the posterior probability of the causal variant shared by GWAS and eQTL. Each dot is an SV, and the color bar indicates LD (r^2). In **a**–**d**, the gray dashed lines represent Bonferroni's corrected significance threshold (two sided) for GWASs ($P = 1.82 \times 10^{-5}$) (**a** and **b**), TWASs ($P = 1.65 \times 10^{-5}$) (**c**), eQTLs (vertical, $P = 1.83 \times 10^{-5}$) and GWASs (horizontal, $P = 1.82 \times 10^{-5}$) of colocalization analysis (**d**). **e**, Diagram showing a 1,454-bp insertion at 990 bp upstream of *BnaA03.MAMf* in a *cis*-eQTL. **f**–**h**, Population allelic variation in *BnaA03.MAMf* expression level (**f**), 4C:(4C + 5C) ratio (**g**) and 5C:(4C + 5C) ratio (**h**) between the accessions with ($n = 132$) and without ($n = 21$) the 1,454-bp insertion. For the legends of boxplots and *P* values, see Fig. 2i. **i**–**k**, Characterization of three independent transgenic lines showing the relative expression level of *BnaA03.MAMf* (**i**), 5C:(4C + 5C) ratios (**j**) and aphid proliferation assay (**k**). HTR-2 (with the 1,454-bp insertion, WT) was transformed with a construct containing the native promoter sequence of *BnaA03.MAMf* without the insertion and HTR-2 CDS. Data are shown as mean ± s.e. *P* values indicate the significance of differences across each of three transgenic lines and the control, determined by two-tailed Student's *t*-tests. The number (*n*) in each column represents biological replicates.

can influence trait variation. Then we used the above two sets of phenotypic and transcriptomic data to perform TWASs and identified 3,487 nonredundant genes that significantly associated with at least one trait (Supplementary Table 17). Of these genes, 311 genes are eGenes with their SV-eQTLs overlapping the GWAS loci, in which the significant associations between phenotypic traits and eGenes were also detected in TWASs. Finally, we identified 726 SV–gene expression–trait variation associations involving 180 eSVs that regulate 311 eGenes which further regulate trait variation. Of these associations involving 97 eSVs and 119 eGenes, 278 are supported by high colocalization posterior probability of eQTL and GWAS loci and are thus causal SVs of trait variation (Fig. 3 and Supplementary Table 18). Figure 3 summarized part of the networks of causal SV–gene expression–trait variation associations and some detailed examples have been presented on an exemplar trait glucosinolate content in the next two sections to show how to identify these associations and underlying mechanisms,

https://doi.org/10.1038/s41588-024-01957-7

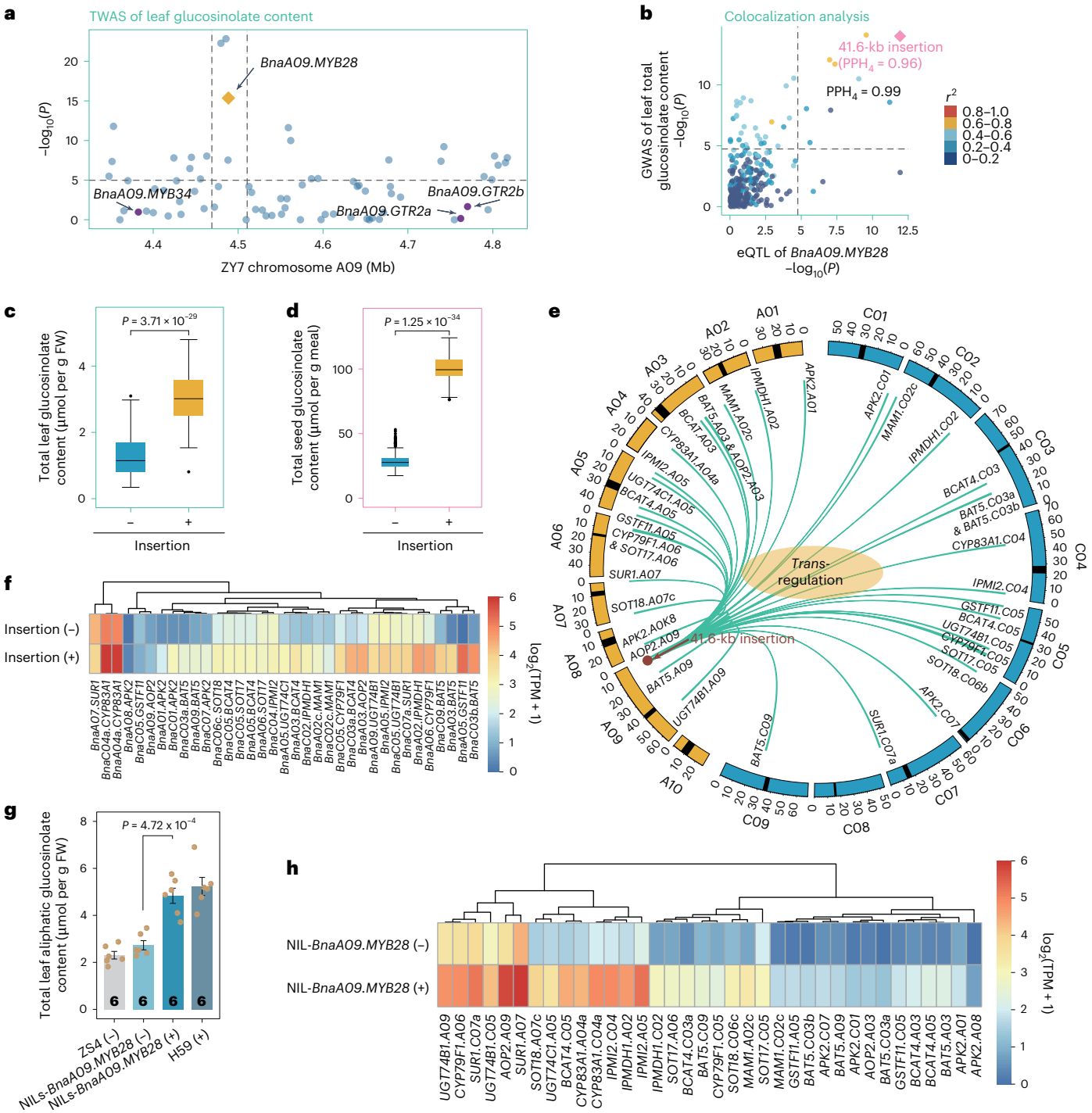


Fig. 5 | Insertion effect originated from a harbored TF gene. a, Local Manhattan plot of TWASs showing a significantly associated locus between gene expression and total leaf aliphatic glucosinolate content. The vertical dashed lines indicate the physical position of a 41.6-kb insertion harboring *BnaA09.MYB28*. **b**, Colocalization analysis of the eQTL regulating *BnaA09.MYB28* expression (*x* axis) and the GWAS QTL of total leaf aliphatic glucosinolate content (*y* axis). The dashed lines represent Bonferroni's corrected significance thresholds that were set at $P = 1.68 \times 10^{-5}$ for TWAS (**a**) and $P = 1.82 \times 10^{-5}$ for GWAS (horizontal) and $P = 1.80 \times 10^{-5}$ for eQTL (vertical) of colocalization analysis (**b**); all *P* values are from two-sided tests. **c**,**d**, Population allelic variation in leaf (**c**) and seed (**d**) glucosinolate contents between the accessions with ($n = 34$) and without ($n = 117$) the 41.6-kb insertion. For the legends of boxplots and *P* values, see Fig. 2i. **e**,**f**, Genomic distribution (**e**) and expression patterns (**f**) of 36 glucosinolate biosynthesis genes regulated by the 41.6-kb insertion carrying *BnaA09.MYB28*. In **e**, local Manhattan plots of eQTLs of the 36 genes are presented in Supplementary Note 7 and, in **f**, $n = 117$ accessions are without the insertion and $n = 34$ with the insertion. **g**, Comparison of total seed aliphatic glucosinolate contents of NILs and their recurrent parent line ZS4 without the insertion and donor parent line H59 but with the insertion carrying *BnaA09.MYB28*. Each error bar is mean ± s.d. Statistical significance (*P* value) was determined using the two-tailed Wilcoxon's rank-sum test. The number (*n*) at the bottom of each column represents the number of samples. **h**, Expression patterns of genes putatively regulated by the 41.6-kb insertion harboring *BnaA09.MYB28* in aliphatic glucosinolate biosynthesis in NILs in the presence (+) or absence (−) of *BnaA09.MYB28*. FW, fresh weight.

which also revealed an integrative landscape of how SVs reshaped variation of a trait.

The case studies on SV impact

To illustrate how complex trait variation is influenced by the SV–gene expression associations established above, we focused on molecular mechanisms affecting glucosinolate content (the biosynthesis and transport pathway). Glucosinolates specific to Brassicales are important compounds in plant defense against disease and insects, and in human nutrition/health such as anti-cancer effect[27], and their biosynthesis in the green organs and subsequent transportation to seeds by transporters were extensively studied in *Arabidopsis* spp.[28–30]. Using total and individual glucosinolates measured in leaves and seeds in the two *B. napus* subpopulations, we identified 119 significantly associated SV-GWAS loci, many of them being new (Fig. 4a,b and Supplementary Tables 15 and 19). In the present study, we highlight key loci as cases to present detailed analyses of glucosinolates and underlying molecular mechanisms that were not dissected previously.

An insertion repressing downstream gene expression. We identified a significant and new locus that controls the glucosinolate ratio of 4C:(4C + 5C) and 5C:(4C + 5C) (representing the enzyme activities of side-chain elongation in glucosinolate biosynthesis) on chromosome A03 (Extended Data Fig. 2d,e). Within this SV-GWAS locus, only one gene, *BnaA03.MAMf*, the expression level of which is significantly associated with 5C-glucosinolate content in TWASs (Fig. 4c) and the link between SVs and *BnaA03.MAMf* expression variation, was established by SV-eQTLs (Extended Data Fig. 2f). The *Arabidopsis MAM1/2/3* orthologs of *BnaA03.MAMf* encode methylthioalkylmalate synthase which determines the natural variation of glucosinolate side-chain elongation[31], and its products significantly correlate with aphid feeding behavior[32]. The regional eQTL and GWAS QTL are colocalized at a strong probability of $PPH_4 = 0.99$, and a 1,454-bp insertion at 990 bp upstream of *BnaA03.MAMf* was identified as the causal variant for both *BnaA03.MAMf* expression and 5C:(4C + 5C) ratio ($PPH_4 = 0.76$) (Fig. 4d,e and Extended Data Fig. 2g). Allelic analyses showed that the accessions without the 1,454-bp insertion had higher expression levels of *BnaA03.MAMf* and lower 4C:(4C + 5C) or higher 5C:(4C + 5C) ratios (Fig. 4f–h).

To verify this SV effect, we constructed an expression vector containing the 2.5-kb native promoter sequence of *BnaA03.MAMf* from a low seed glucosinolate cultivar, ZS11, without the 1,454-bp insertion and the *BnaA03.MAMf* coding sequence from an accession HTR-2 with the insertion in its promoter (Supplementary Note 6), and the vector was transformed into HTR-2. The increases of the transgenic lines in both *BnaA03.MAMf* expression and 5C:(4C + 5C) ratio (Fig. 4i,j) confirmed the 1,454-bp insertion effect (the category I). The aphid feeding bioassay showed that the transgenic lines with elevated *BnaA03.MAMf* expression were more attractive to aphids (*Brevicoryne brassicae*) than wild-type (WT) (HTR-2), indicating that the aphid prefers longer-chain glucosinolates which predominately exist in modern cultivars (Fig. 4k).

Insertion effect originated from harbored TFs. An A09 locus was identified by SV-GWASs (Extended Data Fig. 3a–c), which explains the highest phenotypic variance of total glucosinolate content in leaves (38%) and seeds (60%) among all GWAS loci. Among 25 TWAS genes significantly associated with total glucosinolate content, just one gene, *BnaA09.MYB28*, is involved in glucosinolate accumulation in the locus, which is orthologous to *Arabidopsis MYB28* (*AT5G61420*) (Fig. 5a and Extended Data Fig. 3b,c) and *BnaA09.MYB28* had a significant SV-eQTL (Extended Data Fig. 3d). Pairwise colocalization analysis indicated that a 41.6-kb insertion harboring *BnaA09.MYB28* is the most significantly shared causal variant for both eQTL and GWAS signals (Fig. 5b and Extended Data Fig. 3c–e). The population allelic variation indicated that the presence of the insertion carrying *BnaA09.MYB28* contributes significantly higher glucosinolate content (Fig. 5c,d). Consistently, RNA-seq data from 22 tissues/stages showed no expression of *BnaA09.MYB28* when the insertion was absent in the low glucosinolate cultivar ZS11, contrasting to the WT ZY821 (Extended Data Fig. 3f). Therefore, the regulatory effect of the 41.6-kb insertion on aliphatic glucosinolate content is from *BnaA09.MYB28*.

The *BnaA09.MYB28*/the eSV (41.6-kb insertion) within the above eQTL regulated the expression of a set of eGenes (Fig. 5e and Supplementary Note 7). The previous study indicated that *Arabidopsis* MYB28 is a key TF and regulates a suite of its downstream genes involved in aliphatic glucosinolate biosynthesis[30,33]. To check the *trans*-regulatory effect of *BnaA09.MYB28*/eSV in *B. napus*, we first measured the eGene expression pattern in the accessions with or without *BnaA09.MYB28*/eSV. The results showed that transcripts involved in aliphatic glucosinolate biosynthesis were significantly higher in the accessions with *BnaA09.MYB28*/eSV (Fig. 5f). Furthermore, we developed near-isogenic lines (NILs) of the presence/absence of *BnaA09.MYB28*/eSV and confirmed not only significantly higher total leaf glucosinolate content in the lines with *BnaA09.MYB28* (Fig. 5g), but also the *trans*-regulation pattern of expression (Fig. 5h). The results revealed *trans* regulation of *BnaA09.MYB28*/eSV on downstream eGenes, causing variation in aliphatic glucosinolate content (category II).

Insertion effect originated from its enhancer elements. To elucidate the effect of an SV carrying enhancer elements on remote gene expression and its mechanism (category III), we examined the above datasets (Fig. 3 and Supplementary Tables 9–19) and revealed an overlapped GWAS QTL and eQTL of *BnaC02.GTR2*, the *Arabidopsis* ortholog of which plays a key role in glucosinolate transport[29] (Figs. 4a,b and 6a and Extended Data Fig. 3g–i). *BnaC02.GTR2*, but not *BnaC02.MYB28*, significantly associated with total seed glucosinolate content in TWASs (Fig. 6a) and both genes belonged to different linkage disequilibrium (LD) blocks (Extended Data Fig. 3i). Colocalization of eQTL and GWAS loci pinpoints a 7,365-bp insertion to be the causal eSV at 20.3 kb upstream of *BnaC02.GTR2* (Fig. 6b,c). The dynamic expression and population allelic analysis indicated significant contribution of the insertion to *BnaC02.GTR2* expression and total seed glucosinolate content (Fig. 6d–f). Total seed aliphatic glucosinolate content was significantly positively correlated with the expression level of *BnaC02.GTR2* and related to the insertion (Fig. 6g).

The insertion contains a cluster of enhancer elements, especially CAAT-box motifs (Supplementary Table 20), which was thus predicted as a *cis*-acting DNA sequence to regulate *BnaC02.GTR2* expression. To examine this, we first investigated the chromatin accessibility by assay for transposase-accessible chromatin with high-throughput sequencing (ATAC–seq)[34] and chromatin folding/interaction (Hi-C)[35] in the representative high and low glucosinolate accessions, and ZY821 with the insertion and ZS11 without the insertion. By comparing read coverage depth in the regions surrounding this insertion after read alignment, we identified highly enriched strong ATAC–seq signals in both the enhancer sequence and flanking regions of the insertion in ZY821, but not in ZS11, indicating more accessibility of the chromatin regions with the insertion and its flanking regions (Fig. 6h,i and Supplementary Note 8). By developing an iterative approach to pinpoint local clusters of chromatin interaction frequencies in Hi-C, we captured abundant interactions between the enhancer region and 12 nearby genes including *BnaC02.GTR2* in the ZY821 locus (Fig. 6h). We next detected the active enhancers mark, acetylation of histone 3 at lysine 27 (H3K27ac), by chromatin immunoprecipitation sequencing (ChIP–seq)[36]. We found that the enhancer sequence with the insertion in ZY821 was highly enriched for H3K27ac whereas the signals were absent in the corresponding region of ZS11 (Fig. 6j). As expected, the expression levels of 12 nearby genes, including *BnaC02.GTR2*, were noticeably higher in ZY821 than those in ZS11 (Fig. 6k). These results indicated that the 7,365-bp insertion confers seed glucosinolate content variation by rewiring chromatin spatial structure (Fig. 6l) to regulate at least *BnaC02.GTR2* expression.

https://doi.org/10.1038/s41588-024-01957-7

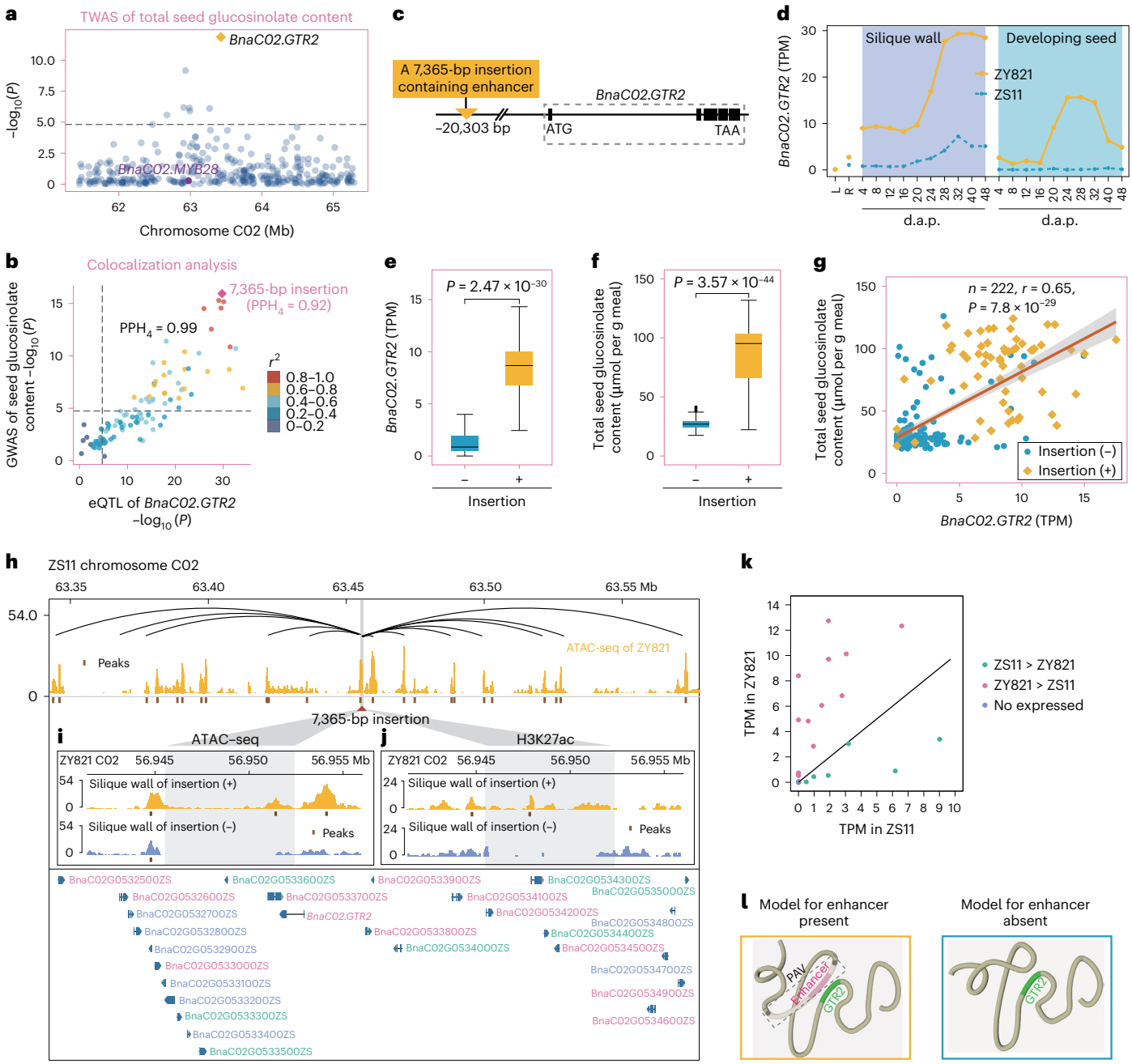


Fig. 6 | Insertion effect originated from its enhancer elements. a, Local Manhattan plot of TWASs showing a significantly associated locus between total seed glucosinolate content and gene expression in developing seeds at 40 d.a.p. **b**, Colocalization analysis of the eQTL regulating *BnaC02.GTR2* expression in seeds at 40 d.a.p. (*x* axis) and the GWAS QTL of total seed glucosinolate contents (*y* axis). The dashed lines represent Bonferroni's corrected significance thresholds that were set at $P = 1.51 \times 10^{-5}$ for TWAS (**a**) and $P = 1.82 \times 10^{-5}$ for GWAS (horizontal) and $P = 1.83 \times 10^{-5}$ for eQTL (vertical) of the colocalization analysis (**b**); all *P* values are from two-sided tests. **c**, Diagram showing a 7,365-bp insertion upstream of *BnaC02.GTR2* in a *cis*-eQTL. **d**, *BnaC02.GTR2* expression pattern in ZY821 with the insertion and ZS11 without the insertion. L, leaves; R, roots. **e**,**f**, Allelic variation in *BnaC02.GTR2* expression levels (**e**) and seed glucosinolate contents (**f**) between the accessions with ($n = 58$) and without ($n = 196$) the insertion. For the legends of boxplots and *P* values, see Fig. 2i. **g**, Correlation between seed glucosinolate contents and expression levels of *BnaC02.GTR2* in developing seeds at 40 d.a.p. The correlation scatter plot shows best fit linear regression line (orange) with 95% confidence intervals (gray). The *r* is Pearson's correlation coefficient with the two-tailed test. **h**–**k**, An enhancer-contained insertion that spatially interacts with target genes for enhanced expression, resulting in glucosinolate content variation. **h**, The enhance–promoter interactions (curves) inferred from Hi-C-captured chromatin folding/interactions by comparing ZY821 and ZS11. **i**,**j**, Enrichments of open chromatin regions (ATAC–seq) (**i**) and histone modifications (H3K27ac) (**j**) in the enhancer and its flanking regions. Short brown vertical lines indicate the peaks from ATAC–seq and ChIP–seq data. The bottom panel shows the annotated genes in this region and upregulated genes in the insertion-present accession ZY821 are marked in pink (**h** and **k**). **k**, Comparison of expression level of genes in the enhancer function region. Each dot represents a gene. Color of dot is same as gene's color in the bottom of (**h**). **l**, Models showing the interaction of the enhancer-contained insertion with target genes. The brown line indicates the chromosome. The pink and green segments represent the enhancer and target genes (*GTR2*), respectively.

https://doi.org/10.1038/s41588-024-01957-7

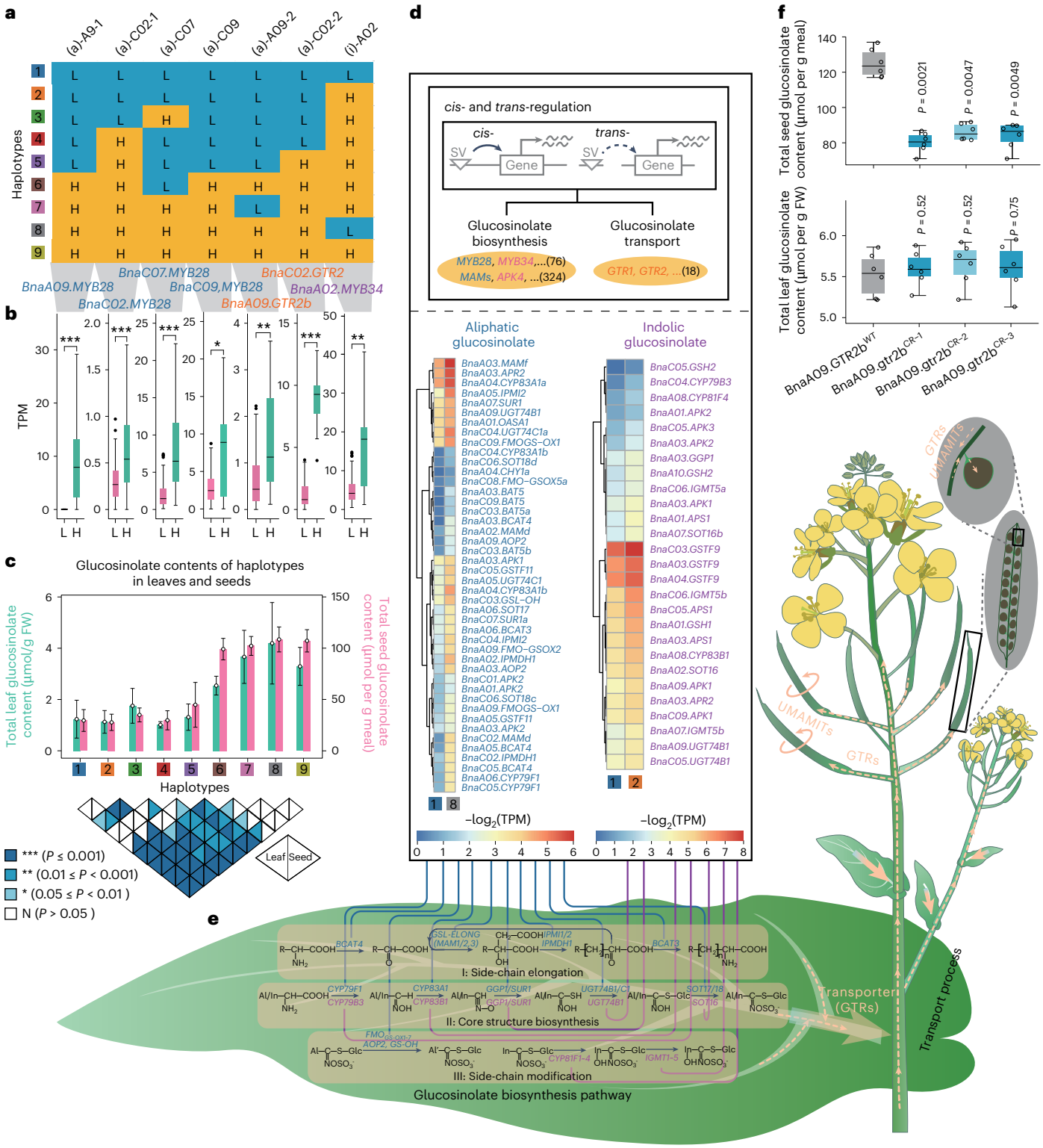


Fig. 7 | A landscape of SVs affecting glucosinolate biosynthesis and transport and its application for breeding. a,**b**, The eSV haplotype alleles and corresponding key gene expression levels determining glucosinolate biosynthesis and transport. H (yellow) and L (blue) indicate the alleles for high or low aliphatic (**a**) or indolic (**b**) glucosinolate contents. For the legends of boxplots and *P* values, see Fig. 2i. **c**, Nine eSV haplotypes representing different contents of leaf and seed total glucosinolates. Data are represented as mean ± s.e. Statistical significance (*P* value) was determined using two-tailed Wilcoxon's rank-sum test. **d**,**e**, A model of eSVs regulating *cis* and *trans* effects on gene expression of the glucosinolate biosynthesis and transport pathway. The regulatory modes of key gene expression (**d**) is mapped to the glucosinolate biosynthesis pathway (orange shadings) shown in an enlarged leaf (**e**). For simplicity, the pathway is shown only in a leaf and not in a silique which is also a major source for biosynthesis. The orange arrows stand for glucosinolate transport paths to seeds for accumulation and the circular arrows for glucosinolate transport in siliques. The pathway construction was based on previous publications[28,30,50]. **f**, Leaf and seed glucosinolate contents in WT and *BnaA09.GTR2*-edited lines. For the legends of boxplots, see Fig. 2i. *P* values show the significance of differences between each transgenic line (three kinds of blue boxes) and WT (gray box) in the two-tailed Student's *t*-tests ($n = 6$ biologically independent samples for each group). In **b** and **c**, for the detailed data of the number of samples (*n*) and *P* values, see Supplementary Table 21.

The other cases. Following the above methods (Figs. 3–6), we can expand many examples based on Supplementary Tables 12 and 18. For glucosinolate content, we briefly presented additional five cases in Extended Data Figs. 4 and 5: identification of three indels that play contrast roles in regulating the other three *BnaMYB28s*' expression and glucosinolate content, identification of an insertion decreasing *BnaA02.MYB34* expression and altering indolic glucosinolate content and identification of a deletion that upregulates the second *BnaGTR2* gene expression to boost glucosinolate contents. We also presented two examples of other traits: one is a 3.7-kb CACTA-like TE insertion (category V) upstream of *BnaA09.CYP78A9* identified within an eQTL hotspot (Hotspot-197; Fig. 2g) which enhances the expression of functionally validated auxin biosynthesis gene *BnaA09.CYP78A9* (ref. 37) in *cis* that *trans*-regulated seven well-known, auxin-responsive genes[38–41] and thereby increases silique length (Fig. 3 and Extended Data Fig. 6). The second (categories VI and VII) is an insertion that putatively mediates *cis* and *trans* regulation of a series of cascade reactions through DNA methylation and histone modification to affect flowering (Supplementary Note 9).

SVs reshape genomic diversity for accelerating breeding

To uncover and demonstrate the role of the eSV–eGene associations in trait enhancement, we took an exemplar trait, the glucosinolate biosynthesis and transport (GBT) pathway and, presenting a landscape of SVs and the regulated genes, especially polyploid duplicated genes (Figs. 3–7 and Extended Data Figs. 2–5 and 7). SV-GWASs detected 549 SVs in 119 loci significantly associated with 31 leaf and seed glucosinolates and derived statistic indices (Supplementary Tables 15 and 19), all containing *Arabidopsis* orthologous genes involved in GBT. The 141 eQTL-GWAS locus pairs were identified as colocalized loci ($PPH_4 > 0.50$) that influence both gene expression and glucosinolate contents. Within the paired loci, 61 colocalized eSVs were identified as causal variants for alteration in the expression of 80 eGenes and contents of glucosinolates (Supplementary Table 18). The genes revealed by complementary and mutually evidence-supported analyses of GWASs, eQTLs and TWASs include 76 TFs, 324 enzyme genes and 36 transporter genes, providing an almost complete list of genes in the pathway (Figs. 3 and 7d and Extended Data Fig. 7a).

Further analyses revealed seven key loci, corresponding to nine haplotype (Hap) combinations, which dominate genetic variation and determine profiles and contents of different glucosinolates in *B. napus* (Fig. 7a–c). The responsible genes in all loci exhibited significant expression differences between the accessions (Fig. 7b–e). Hap7–Hap9 are most common in high glucosinolate accessions, whereas Hap1 and Hap2 are most common in low ones (Extended Data Fig. 7b). Hap7–Hap9 are characterized by the *BnaA09.MYB28* alleles that confer 1.4- to 2.6-fold total leaf and seed glucosinolate contents when compared with Hap1 and Hap 2, respectively (Fig. 7c).

Germplasm with high leaf and low seed glucosinolates has been sought for resistance to fungal pathogens and insects[42], because of the canola breeding for low erucic acid in seed oil required for human health and low glucosinolates in seed meal required for feeding animal initiated in the mid-twentieth century[43]. To our best knowledge, there is no such accession with desirable levels. The reason revealed from the above data is that glucosinolate biosynthesis in green tissues was disrupted by modern low-seed glucosinolate breeding, resulting in a highly positive correlation between leaf and seed glucosinolate contents (Extended Data Fig. 7c). Furthermore, we found that all the significantly associated loci with a large effect on seed glucosinolate content contain *BnaMYB28*, *BnaMYB34* and *BnaGTR2* on A09, C02 and C09, all locating within a syntenic block on each *B. napus* chromosome with strong selective sweeps (Extended Data Fig. 7d–f). This synteny along with the allelic variation analysis showed that all haplotypes with low glucosinolate have nonfunctional *BnaMYB28* and/or *BnaMYB34* TFs that are in linkage with functional *BnaGTR2s*. To breed for a true zero canola would require creating nonfunctional *BnaGTR2s* in each narrow locus. To address the possibility with genome editing, we mutated *BnaA09.GTR2* using clustered regularly interspaced short palindromic repeats (CRISPR)–cas9 (Extended Data Fig. 7g). The result showed the crucial role of *BnaA09.GTR2* in decreasing seed glucosinolates while keeping or even increasing leaf glucosinolates (Fig. 7f). These indicate that SV analysis could enable the discovery of a full range of variants to fine-tune the levels, by conventional and biotechnological breeding, of different glucosinolate contents in leaves and seeds for canola quality of seeds and control of fungal diseases and aphids.

Discussion

Previous studies on individual SVs or in a few individual accessions have highlighted the effect of SVs on gene expression and trait variation[2,4,5]. Furthermore, the present study reveals the extent and scale of genome-wide SV impact across a species by developing the above described strategy to identify population SVs. Our data suggest a more widespread effect of SVs on gene expression (73,580 representing 76% of the *B. napus* genes) than that of SNPs revealed in maize and cotton (44% and 33% of their respective whole-genome genes, detected by SNP-eQTLs[44,45]). Furthermore, a powerful high-throughput joint analysis revealed 726 SV–gene expression–trait variation associations (Figs. 2–7 and Extended Data Figs. 2–7), in which the trait case studies illustrate how SVs reshaped gene expression and trait variation. If using more tissue transcriptomic and phenomic data, the number of associations is expected to noticeably increase. These SVs and related information provide opportunities for conventional breeding to combine SVs with expected effect to breed expected varieties, but also open new avenues to create new/groundbreaking germplasm by genome editing[46,47].

It may be worthwhile highlighting the advantages of the polyploid SV identification and high-throughput joint analysis method described above: the SV identification strategy is cost saving and thus more feasible for detection of SVs from a large population and largely avoids crossmapping of polyploid duplicated genes/sequences[1,8]; the joint analysis method enables high-throughput identification of SV–gene expression–trait variation associations, which thus partially overcome the GWAS/eQTL disadvantage of hardly determining causal/target genes within mapped loci, each usually containing many genes, and provides the capability to dissect how SVs regulate expression of a group of genes and thus enables construction of gene networks/pathways linking traits (Fig. 3 and Extended Data Fig. 7), in all of which genes to be identified are either functionally known orthologs in other plants or functionally uncharacterized, particularly polyploid duplicated genes that have undergone subfunctionalization or neofunctionalization. More details are discussed in Supplementary Note 1.

HE is a noteworthy SV event in polyploids because it has been shown to affect trait variation[23,24] and may play a role in genome evolution[48,49]. However, our analysis attempting to dissect the HE effect encountered difficulties, as a result of either HE's peculiar features such as unusual genome distribution and very low frequency that do not meet the requirement of the current population genomics methods or inaccurate HE identification by the method previously described[23,24], for example, uncertain HE boundaries and false HEs (Supplementary Note 4). New methods are needed to solve or avoid these issues (see Supplementary Note 4 for more details).

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01957-7.

References

1. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).

2. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161 (2020).
3. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
4. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176 (2020).
5. Li, N. et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat. Genet.* **55**, 852–860 (2023).
6. Hadi, K. et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell* **183**, 197–210 (2020).
7. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
8. Schiessl, S.-V., Katche, E., Ihien, E., Chawla, H. S. & Mason, A. S. The role of genomic structural variation in the genetic improvement of polyploid crops. *Crop J.* **7**, 127–140 (2019).
9. Cai, X. et al. Impacts of allopolyploidization and structural variation on intraspecific diversification in *Brassica rapa*. *Genome Biol.* **22**, 166 (2021).
10. Wang, M. et al. Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*. *Nat. Genet.* **54**, 1959–1971 (2022).
11. Wellenreuther, M., Mérot, C., Berdan, E. & Bernatchez, L. Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.* **28**, 1203–1209 (2019).
12. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).
13. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
14. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
15. Hu, J. et al. Genomic selection and genetic architecture of agronomic traits during modern rapeseed breeding. *Nat. Genet.* **54**, 694–704 (2022).
16. Song, J.-M. et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* **6**, 34–45 (2020).
17. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
18. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
19. Lee, H. et al. Chromosome-scale assembly of winter oilseed rape *Brassica napus*. *Front. Plant Sci.* **11**, 496 (2020).
20. Zou, J. et al. Genome-wide selection footprints and deleterious variations in young Asian allotetraploid rapeseed. *Plant Biotechnol. J.* **17**, 1998–2010 (2019).
21. Zimin, A. V. et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).
22. Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
23. Sharpe, A. G., Parkin, I. A. P., Keith, D. J. & Lydiate, D. J. Frequent nonreciprocal translocations in the amphidiploid genome of oilseed rape (*Brassica napus*). *Genome* **38**, 1112–1121 (1995).
24. Chalhoub, B. et al. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**, 950–953 (2014).
25. Kliebenstein, D. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annu. Rev. Plant Biol.* **60**, 93–114 (2009).
26. Yang, Z. et al. BnIR: A multi-omics database with various tools for *Brassica napus* research and breeding. *Mol. Plant* **16**, 775–789 (2023).
27. Kliebenstein, D. J. in *Plant-derived Natural Products: Synthesis, Function, and Application* (eds Osbourn, A. E. & Lanzotti, V.) 83–95 (Springer, 2009).
28. Harun, S., Abdullah-Zawawi, M.-R., Goh, H.-H. & Mohamed-Hussein, Z.-A. A comprehensive gene inventory for glucosinolate biosynthetic pathway in *Arabidopsis thaliana*. *J. Agric. Food Chem.* **68**, 7281–7297 (2020).
29. Nour-Eldin, H. H. et al. NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. *Nature* **488**, 531–534 (2012).
30. Sønderby, I. E., Geu-Flores, F. & Halkier, B. A. Biosynthesis of glucosinolates—gene discovery and beyond. *Trends Plant Sci.* **15**, 283–290 (2010).
31. Abrahams, R. S., Pires, J. C. & Schranz, M. E. Genomic origin and diversification of the glucosinolate MAM locus. *Front. Plant Sci.* **11**, 711 (2020).
32. Zust, T. et al. Natural enemies drive geographic variation in plant defenses. *Science* **338**, 116–119 (2012).
33. Gigolashvili, T., Yatusevich, R., Berger, B., Muller, C. & Flugge, U. I. The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in *Arabidopsis thaliana*. *Plant J.* **51**, 247–261 (2007).
34. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213 (2013).
35. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **162**, 687–688 (2014).
36. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
37. Shi, L. et al. A CACTA-like transposable element in the upstream region of *BnaA9.CYP78A9* acts as an enhancer to increase silique length and seed weight in rapeseed. *Plant J.* **98**, 524–539 (2019).
38. Ulmasov, T., Hagen, G. & Guilfoyle, T. J. Activation and repression of transcription by auxin-response factors. *Proc. Natl Acad. Sci. USA* **96**, 5844–5849 (1999).
39. Franco, A. R., Gee, M. A. & Guilfoyle, T. J. Induction and super-induction of auxin-responsive mRNAs with auxin and protein synthesis inhibitors. *J. Biol. Chem.* **265**, 15845–15849 (1990).
40. Li, M. et al. Grape small auxin upregulated RNA (*SAUR*) 041 is a candidate regulator of berry size in grape. *Int. J. Mol. Sci.* **22**, 11818 (2021).
41. Ruegger, M. et al. The TIR1 protein of *Arabidopsis* functions in auxin response and is related to human SKP2 and yeast Grr1p. *Genes Dev.* **12**, 198–207 (1998).
42. Chhajed, S. et al. Glucosinolate biosynthesis and the glucosinolate–myrosinase system in plant defense. *Agronomy* **10**, 1786 (2020).
43. Kondra, Z. P. & Stefansson, B. R. Inheritance of the major glucosinolates of rapeseed (*Brassica napus*) meal. *Can. J. Plant. Sci.* **50**, 643–647 (1970).
44. Wang, X. et al. Genome-wide analysis of transcriptional variability in a large maize-teosinte population. *Mol. Plant* **11**, 443–459 (2018).
45. You, J. et al. Regulatory controls of duplicated gene expression during fiber development in allotetraploid cotton. *Nat. Genet.* **55**, 1987–1997 (2023).
46. Li, S. et al. Genome-edited powdery mildew resistance in wheat without growth penalties. *Nature* **602**, 455–460 (2022).

47. He, Y. et al. Enhancing canola breeding by editing a glucosinolate transporter gene lacking natural variation. *Plant Physiol.* **188**, 1848–1851 (2022).
48. Mason, A. S. & Wendel, J. F. Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Front. Genet.* **11**, 1014 (2020).
49. Deb, S. K., Edger, P. P., Pires, J. C. & McKain, M. R. Patterns, mechanisms, and consequences of homoeologous exchange in allopolyploid angiosperms: a genomic and epigenomic perspective. *New Phytol.* **238**, 2284–2304 (2023).
50. Xu, D. et al. Export of defensive glucosinolates is key for their accumulation in seeds. *Nature* **617**, 132–138 (2023).

Methods

Plant materials and growth conditions

The *B. napus* accessions in the present study were planted in Wuhan (30° 34′ N and 114° 20′ E), Yangluo (30° 43′ N and 114° 31′ E), Qinghai (36° 44′ N and 101° 45′ E) and Yangzhou (32° 37′ N and 119° 24′ E), in the six growth seasons (2012–2018). All accessions listed in Supplementary Table 1 were maintained by self-pollination. Crop management of field experiments for agronomic trait tests followed the standard protocol of the China National Rapeseed Variety Field Test. The biparental segregating population (ZS4 × H59) and NILs were planted in the experimental fields in Wuhan (30° 34′ N and 114° 20′ E), Yangluo (30° 43′ N and 114° 31′ E) and Qinghai (36° 44′ N and 101° 45′ E) in the growing seasons of 2014–2018.

Phenotyping

All field experiments for phenotypic evaluations were conducted in a randomized block design with at least 3 replicates, in which each plot had at least 30 plants with a space of 30 cm between rows and 20 cm between plants in a row. Sampling methods was dependent on specific traits. For morphological and yield-related traits, measurements were done directly in field at different plant development stages or in labs by randomly sampling ten plants from each plot. For quality traits and chemical compounds, leaves at the seedling or flowering stages and siliques or seeds from the main inflorescence were sampled and analyzed in labs. All measurements followed the standard protocol of the China National Rapeseed Variety Field Test or other specific protocols cited in the paper. For experiments in multiple environments, the best linear unbiased prediction values were estimated using R package 'lme4' (v.1.1-27) as the final phenotypic data for further analysis[51]. For the phenotype description and measurement, see Supplementary Methods.

Genome assembling

For the DNA extraction and library preparation of long-read sequencing for assembly of the six *B. napus* genomes, see Supplementary Methods. The assembly of de novo genomes was performed based on Nanopore reads using different genome assemblers, including Canu (v.1.8)[52], wtdbg2 (v.2.5)[53], Miniasm (v.0.3)[54], Flye (v.2.4.1)[55] and SMARTdenovo (v.1.0)[56]. For each accession, a best assembly based on contiguity metrics (N50, N90 and total genome size) was used for the downstream analysis. The assembled contigs were polished three times using Racon (v.1.3.1)[57] with Nanopore reads as input to correct systematic errors of Nanopore reads. For the short reads were then mapped to contigs using BWA-MEM (v.0.7.15-r1140)[58] and polished 3× using Pilon (v.1.22)[59].

Anchoring and validation of *B. napus* cv. ZY821 assembly

For the Hi-C library preparation, sequencing and data processing, see Supplementary Methods. For contig anchoring, Hi-C reads were aligned to the polished contigs by Burrows–Wheeler Alignment (BWA)-MEM and Hi-C files were obtained using the Juicer pipeline[60]. The polished contigs were sorted and oriented using Hi-C data by 3D DNA pipeline (v.180922)[61]. Of 3,159 contigs 2,598 were anchored into 19 scaffolds. These scaffolds were aligned to the reference genome ZS11 using Mummer4 (v.4.0.0beta2) and named chromosomes A01–A10 and C01–C09 (ref. 62). To evaluate the quality of the ZY821 genome assembly, two genome sequences, ZS11 (ref. 16) and Darmor-*bzh* (v.5)[24] were aligned to the ZY821 genome using Mummer4 (ref. 62) with the parameters '-c 100 -L 1000' for collinearity analysis. Centromeric repeat sequences including CentBr, CRB, TR238 and PCRBr[63] were aligned to the ZY821 assembly using NUCmer-MUMmer4 to identify the locations of centromeres. For the gene annotation of ZY821 genome, see Supplementary Methods.

SNP/indel calling for *B. napus* accessions

For the DNA extraction and library preparation of short-read sequencing for *B. napus* accessions, see Supplementary Methods. The short reads of each accession were aligned to the ZS11 reference genome[16] using BWA-MEM with default parameters. Then the reads with a mapping quality value <10 were filtered out by SAMtools (v.1.6)[64]. SNPs and small indels (≤50 bp) were identified using the Sentieon DNAseq pipeline (v.201911)[65] for each accession. We filtered these data to include only biallelic SNPs/indels using GATK SelectVariants (GATK, v.3.6-0-g89b7209)[66]. To obtain high-quality SNPs, we discarded the SNPs with low mapping quality ('QUAL < 30.0 || MQ < 50.0 || QD < 2') by GATK VariantFiltration. At the population level, all SNPs/indels with MAF < 0.05 and missing >0.1 were discarded using VCFtools software (v.0.1.15)[67]. In addition, SNPs and indels with a heterozygosity rate of >50% were removed.

SV identification and pan-SV library construction

Both contig alignment and long-read alignment strategies were used to identify SVs from 16 *B. napus* genome assemblies (6 from the present study, 10 downloaded from National Center for Biotechnology Information (NCBI) BioProjects: accession nos. PRJNA526961, PRJNA546246 and PRJNA587046) (Supplementary Table 2).

First, we compared the contigs between the ZS11 reference genome and the other 15 assembled genomes using the NUCmer program with the parameters 'nucmer - mum – noextend -L 1000' in MUMmer4. After filtering one-to-one alignments with a minimum alignment length of 50 bp using the delta-filter program from MUMmer4 with parameters '-1 -l 50 -i 95' (ref. 62), NucDiff (v.2.0.3) was used to extract the features and coordinates of SVs with the MD flag[68]. All SVs were filtered out if they were low quality (flag: UNRESOLVED) and had ambiguous breakpoints (flag: IMPRECISE), fewer than four supporting reads, <50 bp and duplicate calling.

Second, for long-read aligning of SV calling, we used NGMLR (v.0.2.8)[14] to map the long read (>500 bp) of each accession on to the ZS11 reference genome, carried out the SV calling using Sniffles (v.1.0.7)[14] and filtered the SVs with the same steps as above for the whole-genome alignment. Then, we merged SV sets from these two identification approaches using SURVIVOR (v.1.0.3) with the parameters '10 1 1 1 0 50' (ref. 69) and combined them into a single variant call format (VCF) using the population-calling method of the Sniffles pipeline[14].

To genotype SVs in 2,105 *B. napus* accessions, we aligned the Illumina short reads from each accession on to the pan-SV library using Paragraph (v.2.0)[22] with default parameters and identified SVs for each accession according to the SV breakpoints in the pan-SV library. For all kinds of SVs (insertion, deletion, inversion and duplication), only split reads were used as evidence and each breakpoint was supported by at least four split reads. For deletions and duplications, the read coverage in the corresponding regions were checked further.

RNA-seq data analysis

For the RNA extraction and library preparation of RNA-seq, see Supplementary Methods. After clipping the adapter sequences and removing the low-quality reads by Trimmomatic (v.0.36)[70], the RNA-seq clean reads from each sample were mapped to the ZS11 reference genome using Hisat2 (v.2.1.2) with default parameters[71]. RNA-seq reads with mapping quality <10, and nonunique and unmapped reads were filtered using SAMtools (v.1.6)[64]. The abundance of genes or transcripts was calculated as transcripts per million (TPM) using StringTie software (v.1.3.6) with default settings[72,73].

GWASs

After quality control and population-level SV filtering with MAF < 0.01 and call rate <0.7, 93,505 high-confidence and high-quality SVs and 7,452,135 SNPs + small indels (≤50 bp) of 2,105 accessions were obtained for further association analysis. The variants (SVs and SNPs) of GWAS accessions were selected by MAF > 0.05 of each of two subpopulations and then were imputed and phased using Beagle (v.5.1)[74]. GWASs were

performed for all traits using GEMMA (v.0.98.1)[75]. The population structure was controlled by including the first three principal components as covariates and an IBS kinship matrix derived from all variants (SNPs or SVs) calculated by GEMMA. We used Wald's test in GEMMA to test the null hypothesis that no association exists between the SNPs/SVs and any of the trails, and Bonferroni's corrected significance threshold ($P = 1/n$, where n represents the total number of SVs or SNPs) was set to determine the significance of associations.

Identification of SV-eQTLs

The expressed genes in five tissues (SAMs, young leaves, siliques at 18 d.a.p. and developing seeds at 20 and 40 d.a.p.) of the accessions of the two subpopulations were analyzed for SV-eQTL mapping, but those genes with expression levels <0.1 (TPM < 0.1) in >95% of accessions for each tissue were discarded. We then filtered genes with the expression level change within twofold between the 5th and the 95th percentile expression levels. In addition, we assessed mismapping of RNA-seq reads between A_n and C_n subgenomes and the results indicated that mismapping of RNA-seq reads is very low (<5%) (Supplementary Note 10), so its influence on eQTL detection is negligible.

To address the preconditions for identification of eQTLs, the gene expression values from a single tissue of which must follow a Gaussian distribution, we performed the quantile normalization of gene expression levels using the 'qqnorm' function in R (v.4.1.2) (http://www.r-project.org). This normalization eliminates potential noise and could increase the power of detecting eQTLs[76]. The normalized gene expression values were then used as the phenotype for subsequent eQTL mapping.

SVs with MAF > 0.05 and call rate >0.7 of each eQTL population were used to perform eQTL mapping with GEMMA (v.0.98.1)[75] to detect associations of SV–gene pairs. For each of the above SV–gene pairs, the P values were obtained from GEMMA by default (Wald's test, two tailed) and Bonferroni's corrected significance threshold ($P = 1/n$, where n represents the total number of SVs in eQTL mapping) was set for determining the significance of associations. We grouped all SVs significantly associated with target genes into a cluster if the distance between two consecutive SVs <50 kb, and the cluster with at least two significant SVs, was considered as a candidate eQTL, which was represented by its most significant SV (named as the lead eSV). Among eQTLs with strong LD ($r^2 > 0.2$), only the most significant eQTL was retained as its representative. Based on the distance between the eQTL and the target genes, we subdivided an eQTL into *cis*-eQTL if its lead eSV was found within 1 Mb of the TSS or the TES of the target gene, and otherwise as *trans*-eQTL.

An eQTL hotspot was defined as influencing expression of many downstream target genes. We identified eQTL hotspots using the Hotscan program (v.05Oct2013)[77]. Different initial window sizes (5, 10, 50, 100, 200 and 500 kb) were tested, the significance level of the adjusted P value was set to 0.05 and finally 200 kb was used to achieve single-gene level resolution.

Gene set enrichment analysis on *trans*-eQTL hotspots

B. napus genes were annotated based on orthologous gene pairs across *B. napus* and *Arabidopsis thaliana* and *A. thaliana* gene ontology (GO) terms[78]. Then, GO enrichment analysis of the genes with expression regulated by 495 *trans*-eQTL hotspots was performed using R package clusterProfiler (v.3.10.1)[79]. The false discovery rate (FDR) threshold of 0.01 was assessed as significant terms overlapping.

Analyses of ChIP–seq and ATAC–seq

The silique wall at 28 d after pollination (at the stage, *BnaA06.GRT2* has the highest expression level) of the accessions ZY821 and ZS11 was harvested and immediately flash frozen into liquid nitrogen for ChIP–seq (H3K27ac) and ATAC–seq. For the library preparation and sequencing for ChIP–seq and ATAC–seq, see Supplementary Methods. Low-quality reads from raw data of ATAC–seq and ChIP–seq were filtered out using Trimmomatic (v.0.36) with parameters 'ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36' (ref. 70). Then, clean reads were mapped to ZS11 and ZY821 genomes using Bowtie (v.2.3.2)[80] with the default parameters. PCR duplicates were removed using Picard tools (v.2.19, http://picard.sourceforge.net). Peaks of ATAC–seq and ChIP–seq were called using the callpeak module of MACS2 software (v.2.1.3.3)[81] with the parameters of '--shift -100 --extsize 200 --board -B -g 9.6e8' and '-B -g 9.6e8', respectively. All other kinds of ChIP–seq data analyses in the present study (H3K4me1, H3K4me3, H3K9me2 and H3K27me3) were also followed as above. For the assessment of quality control for whole ChIP–seq and ATAC–seq data in the present, study see Supplementary Note 10.

SV annotation (regulatory effect annotation of eSVs)

To identify various potential regulation mechanisms of SVs on gene expression, we carried out SV annotation, including SV genome annotation (gene model, TF and TE), regulatory elements, epigenetic modifications, small RNA and so on. Generally, we annotated each eSV for potential regulation of eGenes by vcfanno[82]. In brief, the first annotated SV genomic elements identified relationships of SV physical positions with flanking genomic elements. SV genomic elements, such as gene bodies, TFs and 2-kb upstream and 2-kb downstream regions of genes, were annotated based on alignment on *B. napus* genome assemblies. For a possible regulatory relationship between TF genes and the other eGenes of SV-eQTL, we retrieved data from the previously established database BnIR with the 'TF–target gene' relationships[26]. For SV epigenetic regulation annotation, sequence reads from ChIP–seq of histone modifications (H3K4me1, H3K4me3, H3K9me2, H3K27me3 and H3K27ac), chromatin accessibility (ATAC–seq) and RNA polymerase II (RNAPII) occupancy of accessions from the present study and published data (GSA Bioproject, accession no. PRJCA013095 and NCBI BioProject, accession no. GSE143287) were mapped to the ZS11 reference genome to call peaks. If a peak exists, an epigenetic signal exists in the genomic region under the peak. If >50 bp of an SV or >20% of its length overlaps with the genomic region, the SV was considered to have an epigenetic signal. For enhancers, genomic regions for overlapping peaks of ATAC–seq and H3K27ac were computed using intersectBed in BEDtools (v.2.26.0-114-g4c407ce)[83], but the sequences from 2 kb upstream of a TSS to 2 kb downstream of a TES were excluded; if a genomic region with overlapping peaks overlaps with an SV, the SV was considered to carry or disturb an enhancer. For post-transcriptional regulation annotation, we downloaded *B. napus* primary small RNA sequence data from PmiREN (https://pmiren.com) and mapped them to the pan-SV genome to annotate SV-encoding small RNA information. We further classified the primary small RNA derived from TEs within SVs. For DNA methylation, we collated DNA methylation data (whole-genome bisulfite sequencing (WGBS–seq)) from our methylation sequencing and published data (NCBI BioProject, accession no. GSE143287) and mapped these data to the ZS11 reference genome. If >50 bp of an SV or >20% of its length overlaps with a methylation signal, the SV was considered to have a DNA methylation signal.

TWASs

We conducted TWASs using the sets of transcriptomic and phenomic data used above for SV-eQTL and GWAS analyses[17,84]. The EMMAX (v.beta-07Mar2010) module was used to perform the TWAS tests[17,85]. Bonferroni's corrected significance threshold ($P = 1/n$, where n represents total number of genes in TWASs) was set for determining the significance of associations.

Colocalization analysis of GWASs and eQTLs

To assess the probability of a causal SV that causes both gene expression variation detected by SV-eQTLs and an agronomic trait variation detected by SV-GWASs, we applied COLOC (v.5.1.0)[18] and LocusCompare

(v.0.2.1)[86] to compute posterior causal probabilities for each SV in SV-eQTLs and SV-GWASs. Against the five hypotheses of the Bayesian colocalization (COLOC) method: (1) a genetic locus has no associations with either the SV-eQTL or the SV-GWAS investigated (H_0); (2) the locus is associated only with gene expression (H_1); (3) the locus is associated only with the agronomic trait (H_2); (4) the locus is associated with both via independent SVs (H_3), even if their associated loci are overlapped; and (5) the locus is associated with both traits through a shared SV (H_4). All SVs within overlapped loci of eQTLs and GWAS QTLs were tested for colocalization using default parameters of the software. The posterior probability of H_4 (PPH_4) > 0.50 was considered strong evidence for a colocalized locus of an eQTL-GWAS pair that influence both target gene expression and GWAS trait, and PPH_4 of eSV (>0.50) was considered as a causal SV shared by eQTLs and GWASs[18].

Construction of the NILs of *BnaA09.MYB28*

F_1 seeds were obtained by crossing the *B. napus* homozygous line H59 (*BnaA09.MYB28*^{present} in a 41.6-kb insertion) with ZS4 (no insertion and *BnaA09.MYB28*^{absent}). The heterozygous plants were backcrossed with ZS4 to the BC_4F_2 generation and, in each generation, the *BnaA09.MYB28*^{present} allele was genotyped by SNPs/indels near the insertion. The BC_4F_2 plants were self-pollinated to obtain BC_4F_3 plants for further analysis. The leaf and seed glucosinolate contents were compared between NILs homozygous for either present or absent *BnaA09.MYB28*.

Generation of *BnaA03.MAMf* transgenic plants

Total RNA was extracted from seedling leaves of *B. napus* cv. HTR-2 and used for complementary DNA synthesis using PrimeScript RT reagent Kit with genomic DNA Eraser (Takara, cat. no. RR047A). The full-length CDS sequence of *BnaA03.MAMf* from HTR-2 cDNA was amplified by PCR. The genomic fragment of the *BnaA03.MAMf* promoter sequence (2.44 kb) from ZS11 was amplified by PCR. The amplified coding sequences (CDSs) and promoter sequences were synchronously cloned into *Hand*III–*Bam*HI-digested PBI121 vector by homologous recombination using ClonExpress MultiS One Step Cloning Kit (Vazyme, cat. no. C113-01) to obtain p*BnaA03.MAMf*^{ZS11}::*BnaA03.MAMf*$^{HTR-2}$ recombinant vector. The physical structure of binary vectors used in the present study was shown in Supplementary Note 6. The construct was introduced into HTR-2 by an *Agrobacterium*-mediated transformation method[87]. The positive p*BnaA03.MAMf*^{ZS11}::*BnaA03.MAMf*$^{HTR-2}$ transgenic plants were identified by PCR-based genotyping on genomic DNA. The primers used in the present study are listed in Supplementary Table 22. For the expression measurement (reverse transcriptase–quantitative PCR (RT-qPCR)) of transgenic plants, see Supplementary Methods.

Aphid proliferation assay

Adult aphids were collected and transferred to the seedlings of *B. napus* cv. HTR-2. After 2 d, adults were removed and only their progenies were kept on the plants to ensure that all aphids were the same age (±1 d) at the start of the experiment. After 1 week, six aphids of the same origin were transferred by a soft brush to each test plant (WT and three p*BnaA03.MAMf*^{ZS11}::*BnaA03.MAMf*$^{HTR-2}$ transgenic lines) at 90 d after sowing. For p*BnaA03.MAMf*^{ZS11}::*BnaA03.MAMf*$^{HTR-2}$-1, p*BnaA03.MAMf*^{ZS11}::*BnaA03.MAMf*$^{HTR-2}$-2 and p*BnaA03.MAMf*^{ZS11}::*BnaA03.MAMf*$^{HTR-2}$-3 and WT, each of ten plants was randomly placed on a flat surface and evenly spaced ~10 cm apart, and all plants were grown in 16-cm pots in a greenhouse and under a regimen of 12 h light (21 °C):12 h dark (18 °C). A set of morphologically uniform plants was chosen for the bioassay. The experiments were with three replicates. Then, after another 5 d within the linear proliferation phase, the number of aphids per plant were counted manually.

Gene editing of *BnaA09.GTR2* by CRISPR–Cas9

Two small guide (sg)RNA target sequences, specifically targeting the first and second exon regions of *BnaA09.GTR2* in ZY821 genome, were designed. CRISPR–Cas9 plasmid construction was conducted according to He et al.[47]. In brief, to assemble two guide RNAs, a single PCR fragment flanked by two sgRNA targets was amplified from the pCBC-DT1T2 vector with two pairs of partially overlapping primers, of which two forward and two reverse primers, respectively, contain one of the two target sites. Then the PCR fragment was purified and inserted into the binary vector pHSE401 by a restriction–ligation reaction using *Bsa*I restriction enzyme (New England Biolabs, cat. no. M0202V) and T4 Ligase (New England Biolabs, cat. no. M0202V). The CRISPR–Cas9 binary vector pHSE401-2gR-BnaA09.GTR2b, verified by sequencing, was transformed into *B. napus* cv. ZY821 hypocotyls using the *Agrobacterium*-mediated method[87]. The physical structure of binary vectors used in the present study has been shown in Supplementary Note 6. Genomic DNA from individual transgenic plants was extracted for PCR analysis. The PCR products containing the target sites were amplified with specific primers BnaA09GTR2-F/BnaA09GTR2-R and then cloned into pEASY-T3 vector (TransGen Biotech, cat. no. CT301-01) for sequencing. The primers used in plasmid construction and mutant identification are listed in Supplementary Table 22.

Statistical analysis

All statistics applied in the present study were performed in R (v.4.1.2) and provided alongside the respective analysis in Methods, the main text and figure legends. The statistical tests of significance in GWASs, eQTLs, TWASs and colocalization analysis have been described above.

Inclusion and ethics

We have read the Nature Portfolio Authorship Policy and confirm that this manuscript complies with the policy information about authorship: inclusion and ethics in global research. The researchers who fulfill the authorship criteria are included as co-authors.

Reporting summary

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

Data availability

All sequencing and genome assembly data used in the present study have been deposited into the NCBI database: the long-read sequencing data and genome-assembled contigs data of six *B. napus* accessions (BioProject, accession nos. PRJNA1155214 and PRJNA1149936), short-read data of genome resequencing of 366 accessions (BioProject, accession no. PRJNA1156901), population RNA-seq data of SAMs, leaves and siliques (BioProject, accession nos. PRJNA1149544, PRJNA1157560 and PRJNA1153365, respectively), and the data of Hi-C sequencing, ATAC–seq and Chip–seq (H3K27ac) of ZY821 and ZS11 accessions (BioProject, accession no. PRJNA1155718). All the above data have also been deposited into the National Genomics Data Center (https://ngdc.cncb.ac.cn/?lang=en) database under the GSA Bioproject, accession no. PRJCA013095. The other data generated in the previous studies are publicly available under: the GSA Bioproject, accession no. PRJCA002836 for RNA-seq data of 20-d.a.p. and 40-d.a.p. developing seeds; the NCBI BioProject, accession nos. PRJNA526961, PRJNA546246 and PRJNA587046 for the 10 genome assemblies; the NCBI BioProject, accession nos. SRP067370, SRP125656 and SRP155312, the GSA BioProject, accession no. PRJCA002835 and the ENA Project, accession nos. PRJEB5974 and PRJEB6069 for short-read data of genome resequencing of 1,739 accessions. All genome assemblies, annotations and SV information are available at the BnaOmics Portal[88] (https://BnaOmics.ocri-genomics.net). Source data are provided with this paper.

Code availability

All software and tools used in the present study are publicly available as described in Methods and the Nature Portfolio Reporting Summary. The customized scripts and codes used in the present study

are available via Zenodo at https://doi.org/10.5281/zenodo.13365025 (ref. 89).

References

51. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
52. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
53. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).
54. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
55. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
56. Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a *de novo* assembler using long noisy reads. *GigaByte* https://doi.org/10.46471/gigabyte.15 (2021).
57. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
59. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
60. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
61. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
62. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
63. Mason, A. S. et al. Centromere locations in *Brassica* A and C genomes revealed through half-tetrad analysis. *Genetics* **202**, 513–523 (2016).
64. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
65. Kendig, K. I. et al. Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. *Front. Genet.* **10**, 736 (2019).
66. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
67. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
68. Khelik, K., Lagesen, K., Sandve, G. K., Rognes, T. & Nederbragt, A. J. NucDiff: in-depth characterization and annotation of differences between two sets of DNA sequences. *BMC Bioinf.* **18**, 338 (2017).
69. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
70. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
71. 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).
72. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
73. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
74. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next- generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
75. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821 (2012).
76. Pickrell, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
77. Silva, I. T., Rosales, R. A., Holanda, A. J., Nussenzweig, M. C. & Jankovic, M. Identification of chromosomal translocation hotspots via scan statistics. *Bioinformatics* **30**, 2551–2558 (2014).
78. Yang, Z. et al. BnVIR: bridging the genotype–phenotype gap to accelerate mining of candidate variations underlying agronomic traits in *Brassica napus*. *Mol. Plant* **15**, 779–782 (2022).
79. Tian, T. et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129 (2017).
80. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
81. Gaspar, J. M. Improved peak-calling with MACS2. Preprint at *bioRxiv* https://doi.org/10.1101/496521 (2018).
82. Pedersen, B. S., Layer, R. M. & Quinlan, A. R. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol.* **17**, 118 (2016).
83. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinform.* **47**, 11.12.11–34 (2014).
84. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
85. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
86. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).
87. Zhou, Y. et al. Control of petal and pollen development by the plant cyclin-dependent kinase inhibitor ICK1 in transgenic *Brassica* plants. *Planta* **215**, 248–257 (2002).
88. Cui, X. et al. BnaOmics: a comprehensive platform combining pan-genome and multi-omics data from *Brassica napus*. *Plant Commun.* **4**, 100609 (2023).
89. Yang, Z. Code repository for 'Structural variation reshapes population gene expression and trait variation in 2,105 *Brassica napus* accessions' (v0.0.1). *Zenodo* https://doi.org/10.5281/zenodo.13365025 (2024).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (grant nos. 2022YFD1200400, 2021YFD1600502, 2018YFE0108000 and 2016YFD0101007 to S.L. and 2021YFF1000100 to Q.Y.), the National Natural Science Foundation of China (grant nos. U20A2034 to S.L., 32322061 and 32070559 to Q.Y., 32370681 to Y. Zhang and 32301868 to Y.H.), the Agricultural Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences (grant nos. CAAS-ASTIP-2013-OCRI and CAAS-ZDRW202105 to S.L.) and the China Postdoctoral Science Foundation (grant no. 2022M710875 to Z.Y.).

Author contributions

S.L., Q.Y. and Y. Zhang conceived the project and discussed with H.R. Y. Zhang and Z.Y. conducted genome assembly, pan-SV genome construction and association analysis. Y.H., Y. Zhang, Y.L., M.X., X.C., J.H., L.L. and Y. Zhou collected data and performed experiments

for SV function validation. D.L., M.X., C.L., Q.K. and Y.J. analyzed data from RNA-seq, ChIP–seq, ATAC–seq and WGBS–seq. Y. Zhang and S.L. wrote the manuscript with input from Q.Y., D.J.K., H.R., Y.X. and Z.Y. All authors read, edited and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

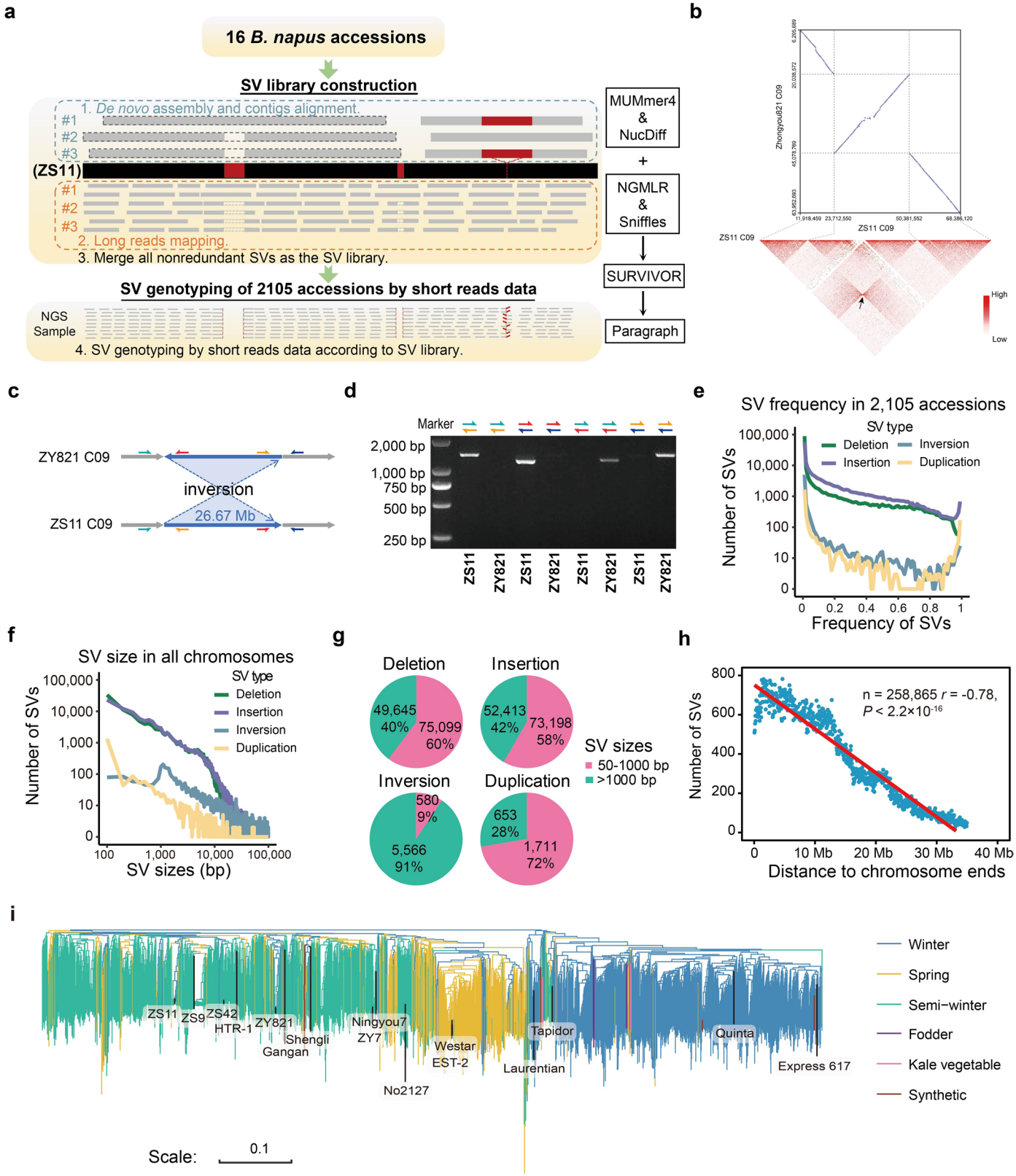
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01957-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01957-7.

Correspondence and requests for materials should be addressed to Shengyi Liu or Qing-Yong Yang.

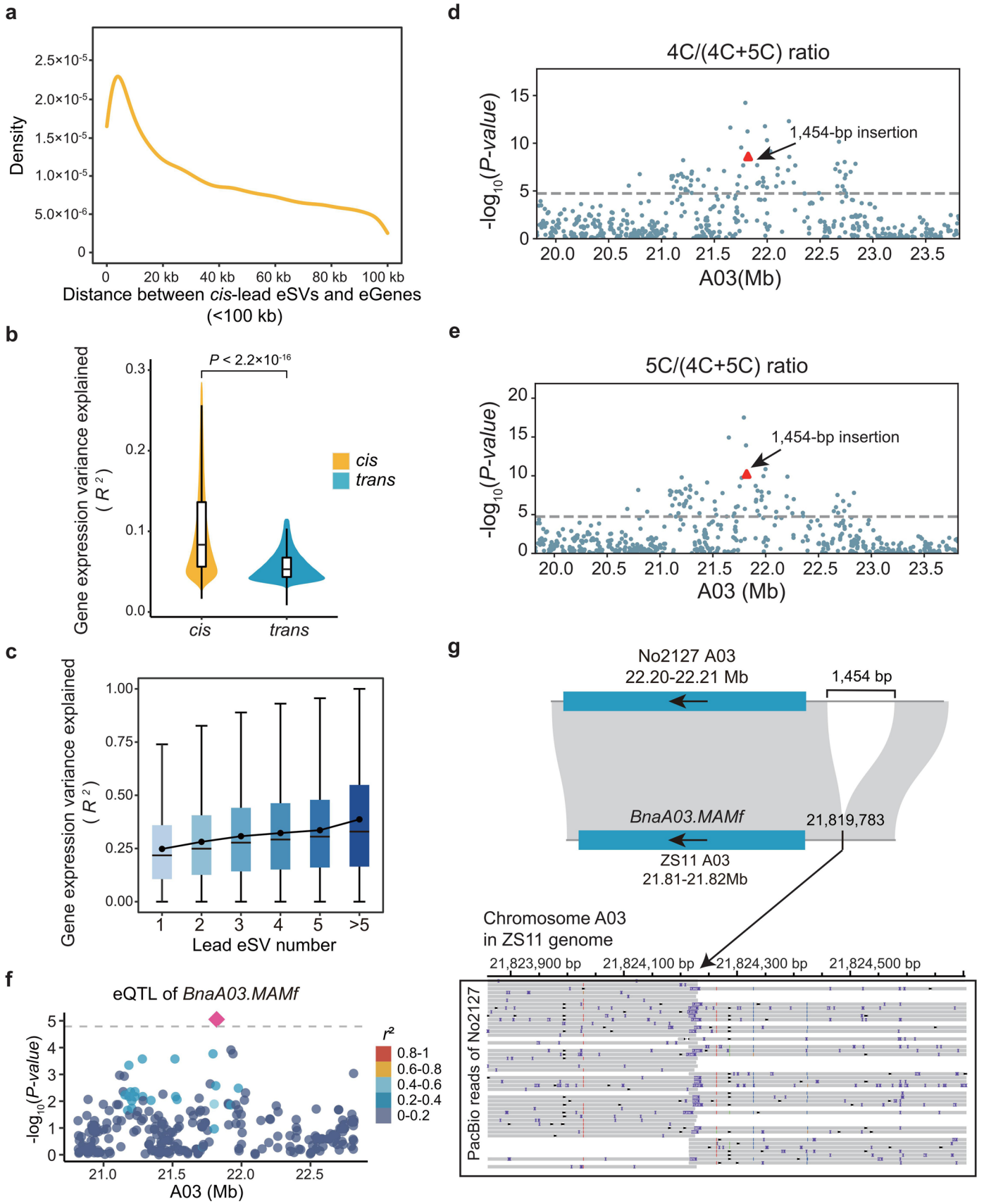
Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-024-01957-7



Extended Data Fig. 1 | Construction and characterization of the *B. napus* panSV genome. (**a**) Overview of SV analysis workflow for panSV construction. #1, #2, #3 and up to 15 accessions in step 1 and 2 indicate accession genomes. (**b**) A demonstration of a large inversion with 26.67 Mb in length detected between ZS11 and ZY821 by genome assembly alignment (top) and Hi-C contact maps (bottom). (**c**, **d**) Verification of the large inversion by PCR amplification of its break point. PCR primers for ZS11 and ZY821 are indicated as corresponding color arrows in schematic diagram (**c**) and gel electrophoresis plot (**d**). The experiments were repeated three times with similar results. (**e**) Relationship between the frequency and the number of SVs in 2,105 accessions (bin width is 100 bp). (**f**) Relationship between size and the number of SVs in the 2,105 accessions (bin width is 100 bp). (**g**) The numbers and ratios of SVs with different sizes. (**h**) Correlation between SV number and distance of SVs to chromosome arm ends. For each SV, the distance was calculated and divided into 500-kb bins. *r* is Pearson correlation coefficient. *P* value was calculated using *F* test for the linear regression model with two-tailed test. Here, the observed value of *P* value is almost zero. (**i**) Phylogenetic analysis of 2,105 *B. napus* accessions based on SVs. The 16 assembled accessions are indicated as black lines.

https://doi.org/10.1038/s41588-024-01957-7

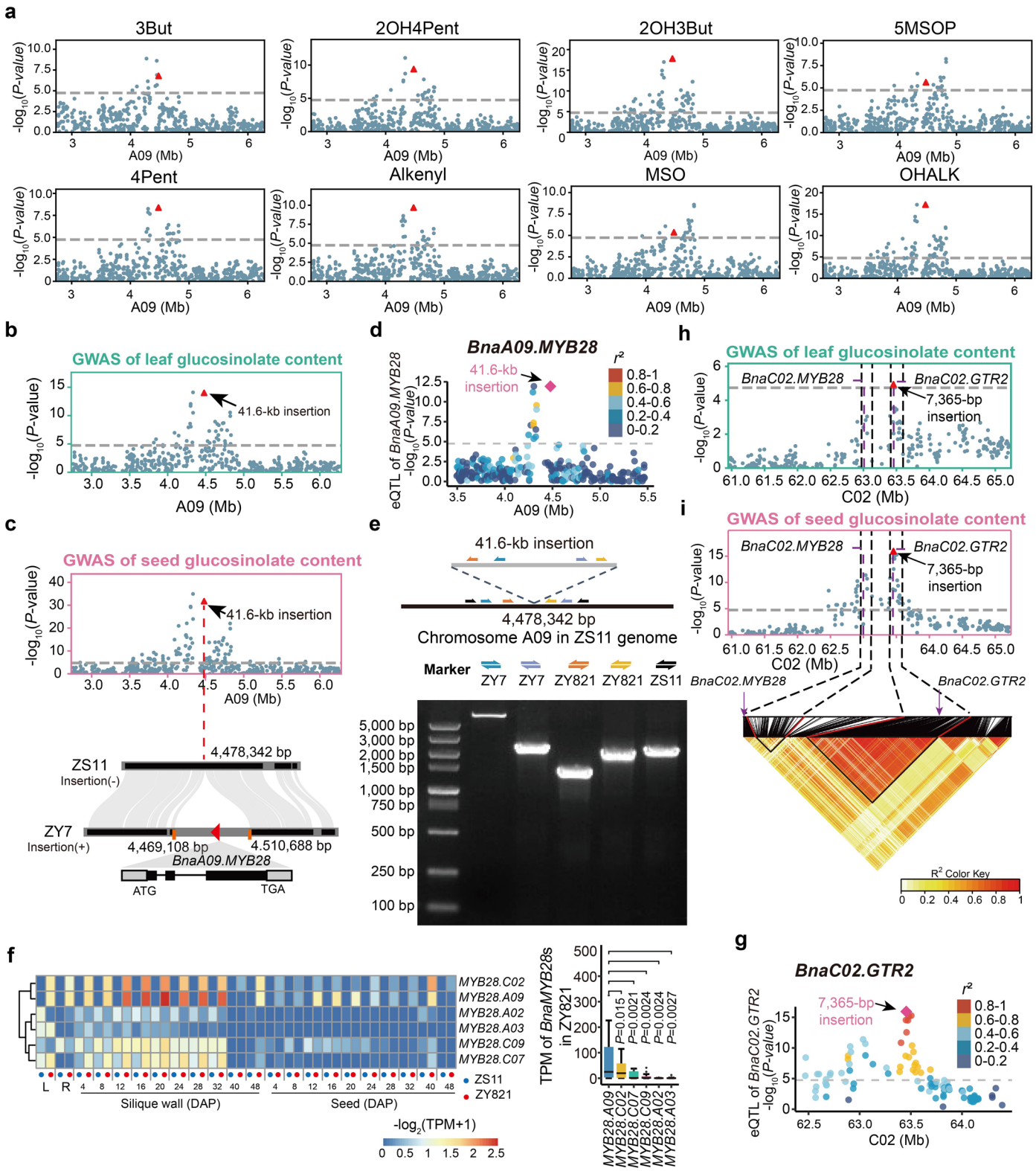


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41588-024-01957-7

Extended Data Fig. 2 | Lead eSV distance to eGenes, the difference in gene expression variance between *cis*-eQTL and *trans*-eQTL and effects of 1,454-bp insertion on short-chain glucosinolates. (**a**) Distribution of distance (<100 kb) between lead eSVs in eQTLs and corresponding eGenes. (**b**) Gene expression (as a phenotype) variance explained by *cis*-eQTLs (n = 66,003) and *trans*-eQTLs (n = 219,973). The violin plots show the distribution density, and the box plots show the distribution quantiles. Here, the observed value of P value is almost zero. (**c**) Accumulated effect of multiple lead eSVs on gene expression. The dots within boxes represent average values. (n = 33,609, 45,855, 46,349, 39,636, 29,220 and 59,368 for six groups, respectively.) (**d**, **e**) Local Manhattan plots of SV-GWAS for the ratios of 4C/(4C + 5C) (**d**) and 5C/(4C + 5C) (**e**), both represent the side-chain 4C and 5C aliphatic glucosinolates in leaves. Deep blue dots represent SVs; red triangles indicate causal SV (1,454-bp insertion) in the promoter of *BnaA03.MAMf*. The gray dashed line represents the Bonferroni-corrected significance threshold (two-sided $P = 1.82 \times 10^{-5}$). (**f**) Local Manhattan plot of eQTL on *BnaA03.MAMf* expression. Each dot stands for an SV and pink square for the causal SV significantly associated with expression of *BnaA03.MAMf*. The gray dashed line represents the Bonferroni-corrected significance threshold (two-sided $P = 1.86 \times 10^{-5}$). The color bar indicates linkage disequilibrium (r^2). (**g**) Identification of the 1,454-bp insertion located in the promoter region of *BnaA03.MAMf* through assembled genome comparison (top) and long-read alignment (bottom). In (**b**, **c**), see Fig. 2i for the legends of boxplots and P values.

https://doi.org/10.1038/s41588-024-01957-7

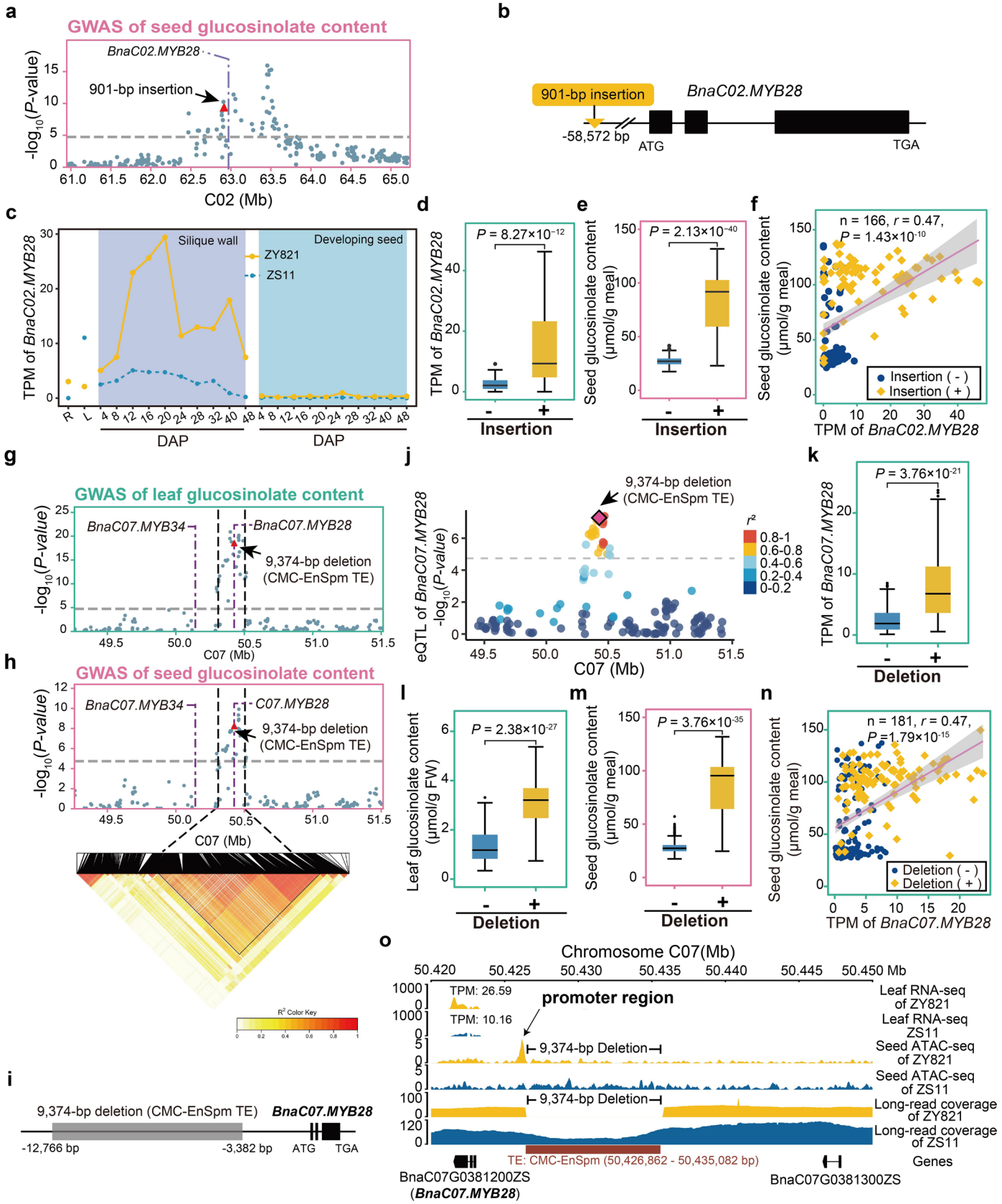


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41588-024-01957-7

Extended Data Fig. 3 | The additional evidence of eSV case studies in Figs. 5 and 6. (**a**–**c**) Local Manhattan plots of SV-GWAS for 8 kinds of leaf aliphatic glucosinolate contents (**a**), total leaf (**b**) and seed (**c**, top) glucosinolate contents on chromosome A09. See Supplementary Table 14 for glucosinolate abbreviations. The bottom of (**c**) shows the 41.6-kb insertion leading to the present/absent of *BnaA09.MYB28* in different accessions. (**d**) Local Manhattan plot of eQTL of *BnaA09.MYB28* expression. (**e**) PCR amplification to verify the break point of the insertion. PCR primers for ZY7 and ZY821, both with the insertion, and ZS11 without the insertion are indicated as corresponding color arrows. The experiments were repeated three times with similar results. (**f**) Expression patterns and statistics of all *BnaMYB28* family members in low (ZS11) and high (ZY821) glucosinolate accessions. L: leaves; R: roots; DAP: days after pollination. The box plots (right part) show the statistics of the expression levels of each *BnaMYB28* from 22 tissues in ZY821. *P* values show the significance of differences between the expression level of *BnaA09.MYB28* and that of each of other *BnaMYB28s* in the two-tailed paired *t* tests. (**g**, **h**) Local Manhattan plots of SV-GWAS of the leaf (**g**) and seed (**h**, top) glucosinolate contents. In the bottom of (**h**), the red triangle represents the causal SV locating together with *BnaGTR2.C02* in the same LD block which separates (vertical dash lines) from the other LD block with *BnaC02.MYB28*. (**i**) Local Manhattan plot of eQTL of *BnaC02.GTR2* expression. In (**a**–**c**) and (**g**, **h**), see Extended Data Fig. 2d for the legends. The gray dashed line represents the Bonferroni-corrected significance threshold for GWAS (two-sided $P = 1.82 \times 10^{-5}$). In (**d**) and (**i**), see Extended Data Fig. 2f for the legends. The gray dashed line represents the Bonferroni-corrected significance threshold for eQTL (two-sided $P = 1.80 \times 10^{-5}$ for *BnaA09.MYB28* and two-sided $P = 1.83 \times 10^{-5}$ for *BnaC02.GTR2*). In (**f**) and (**n**), see Fig. 6g for the legends of statistical test and *P* value.

https://doi.org/10.1038/s41588-024-01957-7

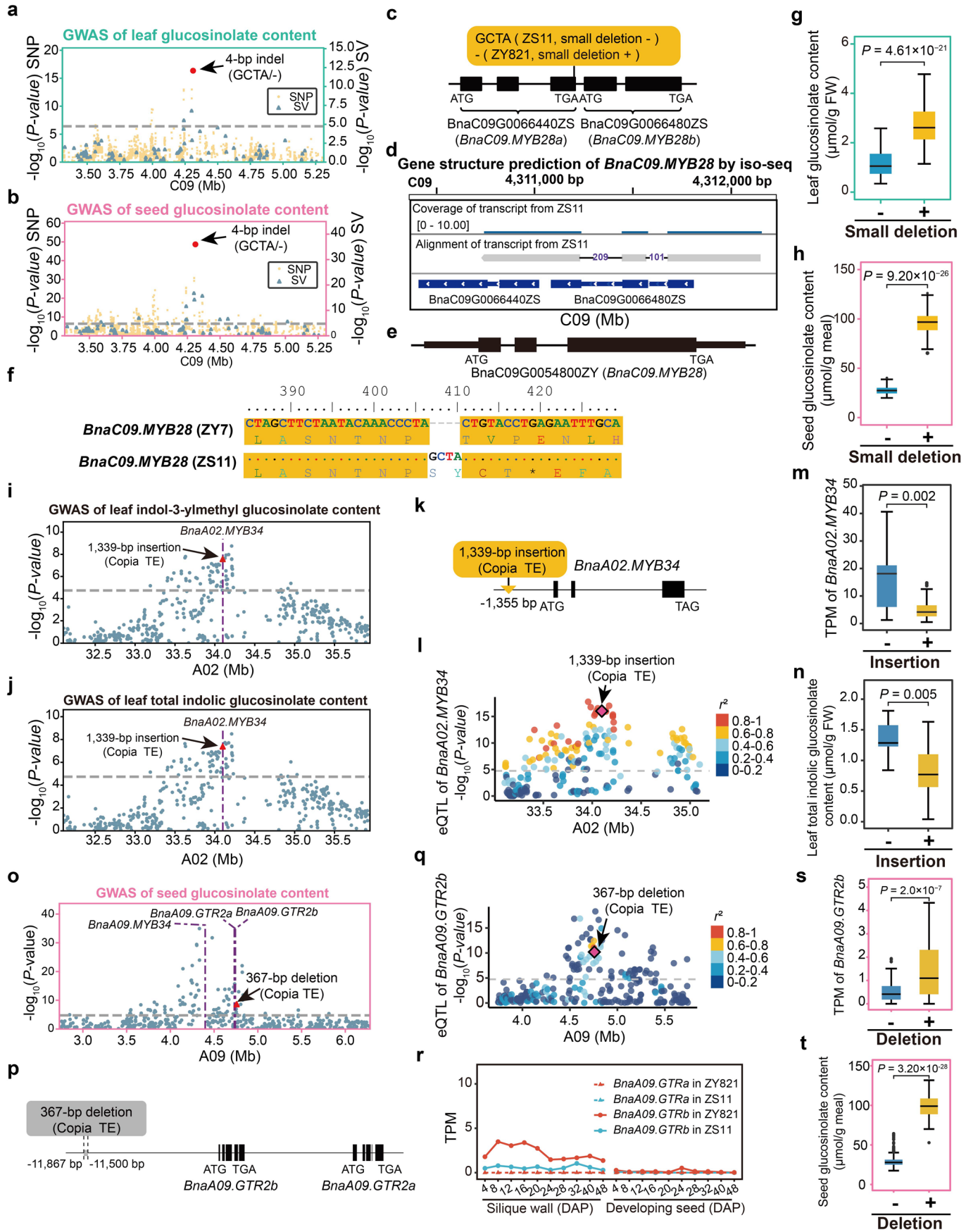


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41588-024-01957-7

Extended Data Fig. 4 | Identification of two insertions/deletions that play contrast roles in regulating the expression of the other two *BnaMYB28s* and contents of glucosinolates. (**a**) Local Manhattan plot of SV-GWAS of seed glucosinolate contents highlighting a significantly associated 901-bp SV. (**b**) Diagram showing the 901-bp insertion at the upstream of *BnaC02.MYB28* in a *cis*-eQTL. (**c**) *BnaC02.MYB28* expression pattern in ZS11 (without the insertion) and ZY821 (with the insertion). (**d**, **e**) Allelic variation in *BnaC02.MYB28* expression (**d**) and seed glucosinolate contents (**e**) between the accessions with presence ($n = 77$) or absence ($n = 100$) of the 901-bp insertion. (**f**) Correlation between seed glucosinolate content and *BnaC02.MYB28* expression in *B. napus* population. (**g**–**i**) Local Manhattan plots of SV-GWAS of leaf (**g**) and seed (**h**) glucosinolate contents. The causal SV near *BnaC07.MYB28* (**i**) was separated from the other LD block containing *BnaC07.MYB34*. (**j**) Local Manhattan plot of eQTL of *BnaC07.MYB28* expression. For the legends, see Extended Data Fig. 2f. The gray dashed line represents the Bonferroni-corrected significance threshold (two-sided $P = 1.86 \times 10^{-5}$). (**k**–**m**) Allelic variation in *BnaC07.MYB28* expression levels (**k**) and total glucosinolate contents in leaves (**l**) and seeds (**m**) between the accessions with ($n = 100$) or without ($n = 173$) the 9,374-bp deletion. (**n**) Correlation between seed glucosinolate content and *BnaC07.MYB28* expression in *B. napus* population. (**o**) Screenshot showing that the 9,374-bp deletion increases chromatin accessibility in the promoter region of *BnaC07.MYB28*, potentially enhancing gene expression for higher glucosinolate content. The middle two panels show difference of the enrichment of chromatin accessibility (ATAC-Seq) in the promoter region (indicated by an arrow) of *BnaC07.MYB28* in two representative accessions (ZY821 with the deletion and ZS11 without the deletion). The bottom three panels show long-reads coverage supporting the SV's existence with *BnaC07.MYB28*. In (**a**), (**g**) and (**h**), see Extended Data Fig. 2d for the legends of symbols and statistical test for GWAS. In (**d**), (**e**), (**k**–**m**), see Fig. 2i for the legends of boxplots and *P* values. In (**f**) and (**n**), see Fig. 6g for the legends of statistical test and *P* value.

https://doi.org/10.1038/s41588-024-01957-7

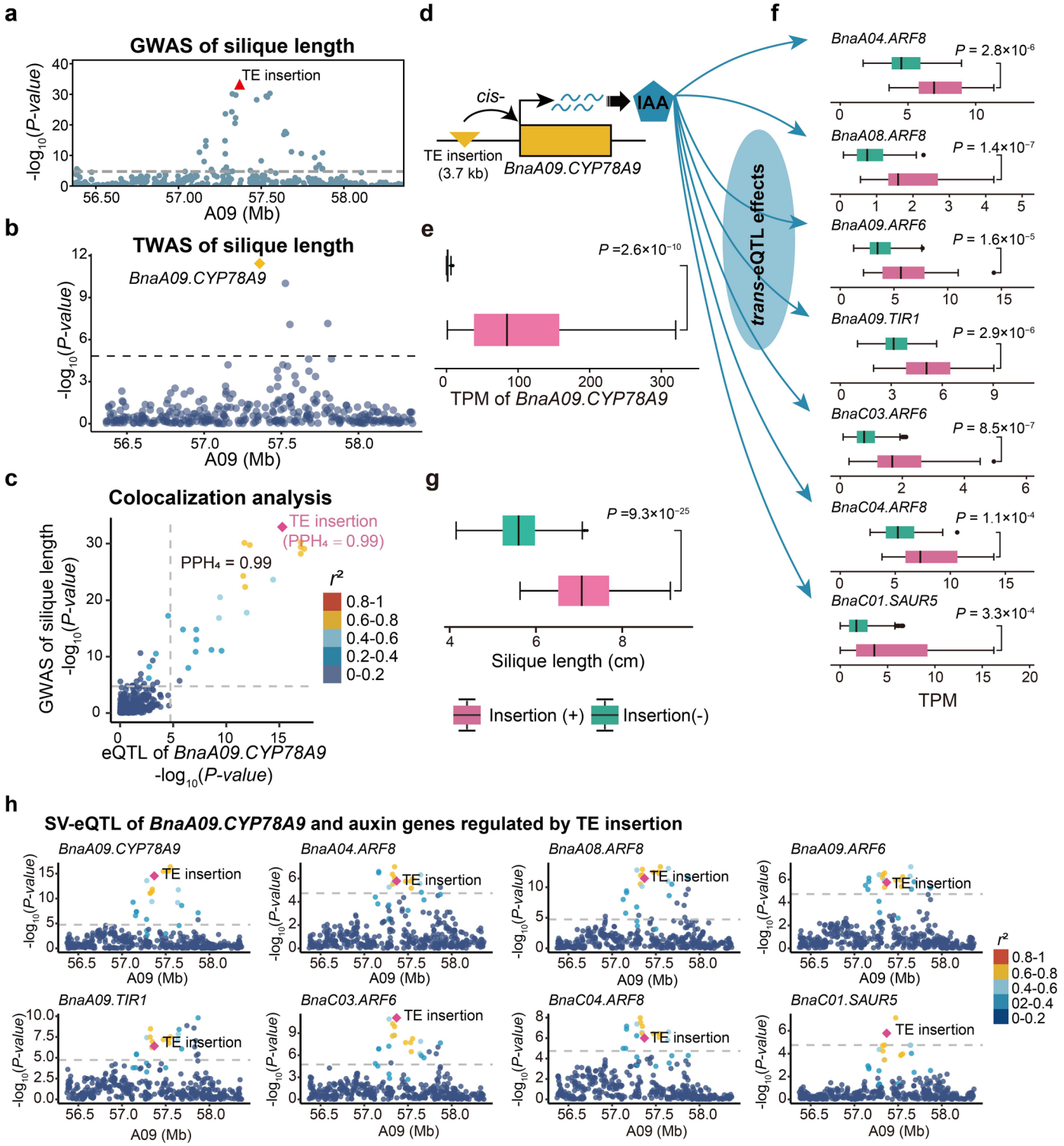


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41588-024-01957-7

Extended Data Fig. 5 | Identification of three pairs of SVs/InDel—key genes affecting glucosinolate contents. (**a**, **b**) Local Manhattan plots of SV-GWAS highlighting a target InDel significantly associated with total glucosinolate contents in leaves (**a**) and seeds (**b**). (**c**–**e**) The causal InDel identification and gene re-annotation. This 4-bp InDel locates in one of two contiguous *BnaA09.MYB28s* in the ZS11 reference genome (**c**), but the single-molecule long-read isoform sequencing (Iso-seq) revealed only one gene in this region (**d**), and leading to re-annotation as one single gene *BnaC09.MYB28ZY* (**e**). (**f**) Sequence comparison of DNA and amino acids of *BnaC09.MYB28* between ZS11 and ZY7. (**g**, **h**) Population allelic variation in glucosinolate contents in leaves (**g**) and seeds (**h**) between the accessions with or without the 4-bp deletion ($n = 237$ and $n = 95$ for leaves; $n = 218$ and $n = 92$ for seeds), suggesting the 4-bp deletion increases glucosinolate contents. (**i**, **j**) Local Manhattan plots of SV-GWAS highlighting a target 1,339-bp insertion significantly associated with Indol-3-ylmethyl glucosinolate content (**i**) and total indolic glucosinolate content (**j**) in leaves. (**k**, **l**) Diagram showing the 1,339-bp insertion in the promoter region of *BnaA02.MYB34* (**k**) in a *cis*-eQTL (**l**). (**m**, **n**) Population allelic variation in *BnaA02.MYB34* expression (**m**) and total indolic glucosinolate content (**n**) between the accessions with ($n = 11$) or without ($n = 143$) the 1,339-bp insertion, indicating the insertion decreases total indolic glucosinolate content through downregulating *BnaA02.MYB34* expression. (**o**) Local Manhattan plot of SV-GWAS highlighting a 367-bp deletion significantly associated with total seed glucosinolate contents. (**p**, **q**) Diagram showing the 367-bp deletion in upstream of *BnaA09.GTR2b* (**p**) in a *cis*-eQTL (**q**). (**r**) Expression patterns of *BnaA09.GTR2b* and *BnaA09.GTR2a* in ZS11 (with the deletion) and ZY821 (without the deletion). (**s**, **t**) Population allelic variation in *BnaA09.GTR2b* expression (**s**) and total seed glucosinolate contents (**t**) between the accessions with ($n = 241$) or without ($n = 49$) the 367-bp deletion. In (**a**, **b**), (**i**, **j**) and (**o**), see Extended Data Fig. 2d for the legends of symbols and statistical test for GWAS. In (**g**, **h**), (**m**, **n**) and (**s**, **t**), see Fig. 2i for the legends of boxplots and *P* values. In (**l**) and (**q**), see Extended Data Fig. 2f for the legends. The gray dashed line represents the Bonferroni-corrected significance threshold for eQTL (two-sided $P = 1.83 \times 10^{-5}$ for *BnaA02.MYB34* and two-sided $P = 1.80 \times 10^{-5}$ for *BnaA09.GTR2b*).

https://doi.org/10.1038/s41588-024-01957-7

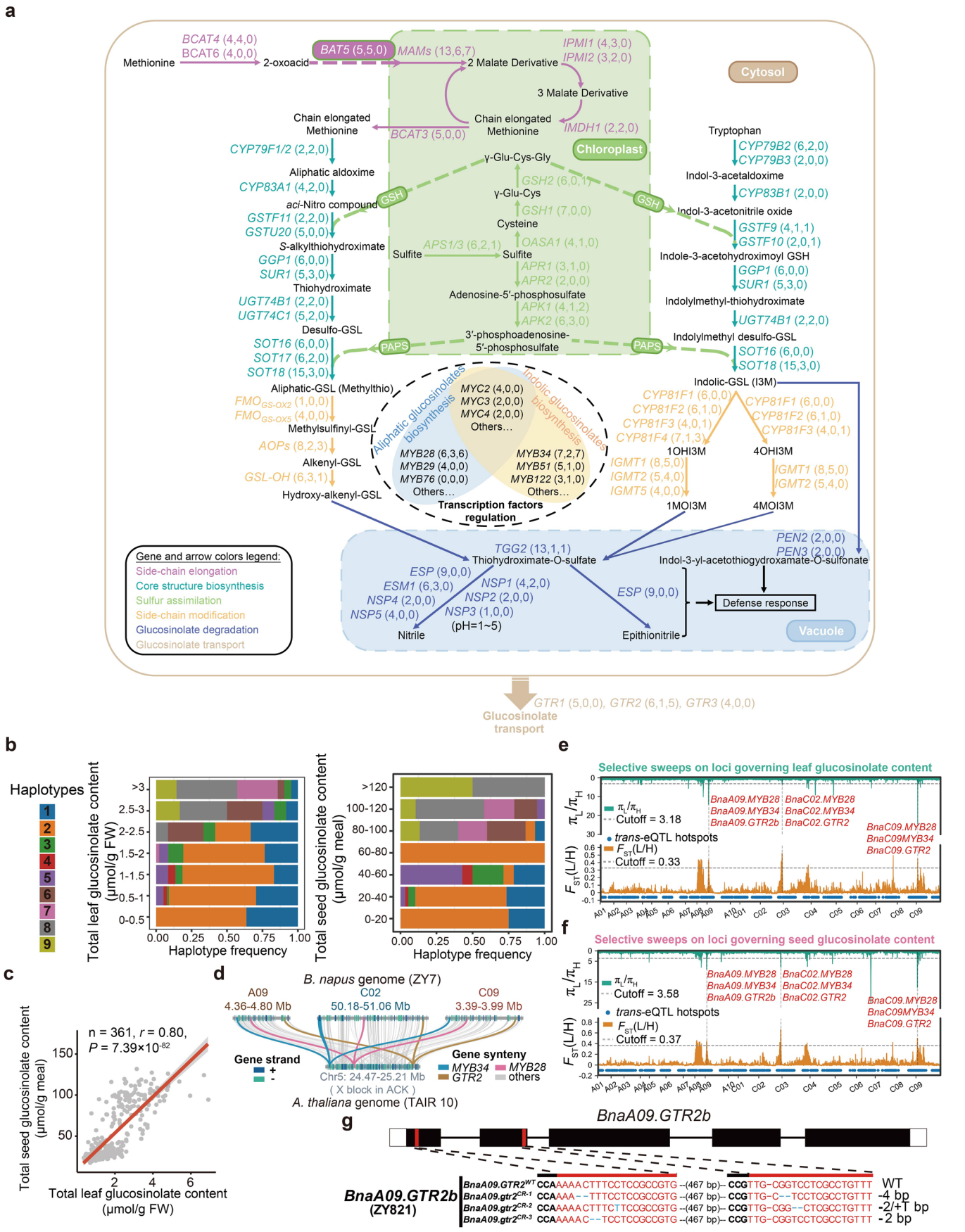


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41588-024-01957-7

Extended Data Fig. 6 | A TE insertion increasing silique length by *cis*- *and trans*-regulating expression of downstream genes. (**a**) Local Manhattan plot of SV-GWAS highlighting a 3.7-kb insertion (red triangle arrow) significantly associated with silique length. See Extended Data Fig. 2d for the legends. The gray dashed line represents the Bonferroni-corrected significance threshold (two-sided $P = 1.84 \times 10^{-5}$). (**b**) Local Manhattan plot of TWAS showing an association between gene expression in siliques at 18 DAP and silique length in which *BnaA09.CYP78A9* exhibits a most significant association. See Fig. 4c for the legends. The gray dashed line represents the Bonferroni-corrected significance threshold (two-sided $P = 1.80 \times 10^{-5}$). (**c**) Colocalization analysis of the eQTL regulating *BnaA09.CYP78A9* expression in silique at 18 DAP (x axis) and GWAS QTL of silique length (y axis), suggesting a causal SV (insertion) that is 3.7 kb transposable element (TE). See Fig. 4d for the legends. The horizontal and vertical gray dashed lines represent the Bonferroni-corrected significance threshold of GWAS (two-sided $P = 1.84 \times 10^{-5}$) and eQTL (two-sided $P = 1.83 \times 10^{-5}$), respectively. (**d**) Diagram showing that the 3.7-kb insertion is *cis*-eSV upstream of 5′-end of the target gene *BnaA09.CYP78A9*. Blue pentagon with IAA indicates auxin compounds. (**e**) Population allelic variation in *BnaA09.CYP78A9* expression level between the accessions with ($n = 162$) or without ($n = 21$) the insertion in the population. (**f**) Expression of the downstream auxin-responsive genes upregulated by the 3.7-kb insertion in *trans*-eQTL hotspot-197 ($n = 162$ with insertion and $n = 60$ without insertion). (**g**) Population allelic variation in silique length between the accessions with ($n = 162$) or without ($n = 21$) the insertion in the population. (**h**) Local Manhattan plots of eQTL on expression of *BnaA09.CYP78A9* and seven downstream auxin-responsive genes. See Extended Data Fig. 2f for the legends. The gray dashed line represents the Bonferroni-corrected significance threshold (two-sided $P = 1.83 \times 10^{-5}$). In (**e**–**g**), see Fig. 2i for the legends of boxplots and *P* values.

https://doi.org/10.1038/s41588-024-01957-7



Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41588-024-01957-7

Extended Data Fig. 7 | Summary of SV impact on glucosinolates biosynthesis and transport, and selective sweep on loci governing glucosinolate content and editing of *BnaA09.GTR2*. (**a**) Summary of SV impact on gene expression in the glucosinolate biosynthesis and transport pathways in *B. napus*. Bold arrows represent glucosinolate biosynthesis, degradation and transport reactions; dash arrows indicate the transport steps. The three numbers in brackets beside each gene, such as *MAMs* (13, 6, 7), represent: 13 is the total number of homologous genes that have SV-eQTL, 6 is the total number of homologous genes whose expression levels were significantly associated with glucosinolate contents in TWAS, and 7 is the total number of homologous genes that locate in SV-GWAS loci of glucosinolate contents. Genes were named based on function annotation and their orthologous/syntenic relationship with *Arabidopsis*. Pathway information from previous publications[28,30,88]. (**b**) The frequencies of eSV haplotypes (identified in Fig. 7a) determining leaf and seed total glucosinolate contents. (**c**) The correlation between total glucosinolate contents of leaves and seeds in a *B. napus* population. See Fig. 6g for the legends of statistical test and *P* value. (**d**) The linked key genes *MYB28*, *MYB34* and *GTR2* on three *B. napus* chromosomes A09, C02 and C09 are syntenic to an ancestral block in *A. thaliana* Chromosome 5. (**e**, **f**) Selective sweep loci governing glucosinolate content in leaves (**e**) and seeds (**f**). The values of π_L/π_H (the ratio of nucleotide diversity) and F_{ST} (genome differentiation) were estimated from SVs between the accessions with extremely high (H, top 20%) and extremely low (L, bottom 20%) glucosinolate contents. The horizontal gray dash lines are the genome-wide thresholds for selective sweeps. The vertical dash lines show the loci with both GWAS and selection signals containing *BnaMYB28*, *BnaMYB34* and *BnaGTR2*. (**g**) Characterization of *BnaA09.GTR2b* edited using CRISPR/Cas9 in ZY821. The protospacer adjacent motif (PAM) is highlighted in bold (CCA and CCG). CRISPR/Cas9 sgRNA-1 and sgRNA-2 targeting the first and second exons of *BnaA09.GTR2b*, respectively, are shown in red. The blue letters and hyphens indicate insertions and deletions in edited plants, respectively.

nature portfolio

Corresponding author(s): Shengyi Liu and Qing-Yong Yang

Last updated by author(s): Sep 6, 2024

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Oxford Nanopore Technology (ONT) platform produced Long-read sequencing data; Hiseq-PE150 sequencing system (Illumina) produced short-read sequencing data, including whole genome re-sequencing, RNA-Seq, Hi-C sequencing, ATAC-Seq and Chip-Seq data.
Data analysis	Genome assembling: Canu (version 1.8) (https://github.com/marbl/canu) wtdbg2 (version 2.5) (https://github.com/ruanjue/wtdbg2) Miniasm (version 0.3) (https://github.com/lh3/miniasm) Flye (version 2.4.1) (https://github.com/fenderglass/Flye) SMARTdenovo (version 1.0) (https://github.com/ruanjue/smartdenovo) Racon (version 1.3.1) (https://github.com/isovic/racon) BWA-MEM (version 0.7.15-r1140) (https://github.com/lh3/bwa) Pilon (version 1.22) (https://github.com/broadinstitute/pilon) Hi-C data processing: BWA-MEM (version 0.7.15-r1140) (https://github.com/lh3/bwa) Juicer_tools (version 1.7.6) (http://aidenlab.org/juicer/) Genome anchoring: 3D-DNA pipeline (version 180922) (https://github.com/aidenlab/3d-dna) Mummer4 (version 4.0.0beta2) (http://mummer4.github.io)

Gene annotation:
RepeatModeler (version 1.0.11) (http://www.repeatmasker.org/RepeatModeler)
LTR_FINDER (version 1.0.5) (https://github.com/xzhub/LTR_Finder)
RepeatMasker (version 4.0.9) (http://www.repeatmasker.org)
RepBase (version 20181026) (https://www.girinst.org/server/RepBase/index.php)
Tandem Repeat Finder-TRF (version 2.5) (https://tandem.bu.edu/trf/trf.html)
MAKER (version 2.31.10) (https://yandell-lab.org/software/maker.html)
Augustus (version 3.3.2) (http://bioinf.uni-greifswald.de/augustus/)
GlimmerHMM (version 3.0.4) (http://ccb.jhu.edu/software/glimmerhmm/)
Exonerate (version 2.2.0) (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate)
InterProScan5 (version 5.11-51.0) (https://github.com/ebi-pf-team/interproscan)
BLASTP (version 2.6.0+) (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

SNP/InDel calling:
SAMtools (version 1.6) (https://github.com/samtools/samtools)
Sentieon-DNAseq (version 201911) (https://github.com/Sentieon/sentieon-dnaseq)
GATK (version 3.6-0-g89b7209) (https://github.com/broadinstitute/gatk/)
VCFtools (version 0.1.15) (https://vcftools.github.io/index.html)

SV identification and panSV library construction:
NucDiff (version 2.0.3) (https://github.com/uio-cels/NucDiff)
NGMLR (version 0.2.8) (https://github.com/philres/ngmlr)
Sniffles (version 1.0.7) (https://github.com/fritzsedlazeck/Sniffles)
SURVIVOR (version 1.0.3) (https://github.com/fritzsedlazeck/SURVIVOR)
Paragraph (version 2.0) (https://github.com/Illumina/paragraph)

Phylogenetic tree construction:
TreeBeST (version 1.9.2) (https://github.com/Ensembl/treebest)
iTOL (version 5) (https://itol.embl.de/)

RNA-Seq data analysis:
Trimmomatic (version 0.36) (http://www.usadellab.org/cms/index.php?page = trimmomatic)
Hisat2 (version 2.1.2) (https://ccb.jhu.edu/software/hisat2/index.shtml)
SAMtools (version 1.6) (https://github.com/samtools/samtools)
StringTie (version 1.3.6) (http://ccb.jhu.edu/software/stringtie/)

Genome-wide association study:
Beagle (version 5.1) (http://faculty.washington.edu/browning/beagle/beagle.html)
GEMMA (version 0.94.1) (http://stephenslab.uchicago.edu/software.html)

Mapping and hotspot identification of SV-based expression quantitative trait loci:
''qqnorm'' function in R (version 4.1.2) (http://www.r-project.org/)
GEMMA (version 0.98.1) (http://stephenslab.uchicago.edu/software.html)
Hot_scan (version 05Oct2013) (https://github.com/itojal/hot_scan)
clusterProfiler (version 3.10.1) (http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html)

Analyses of ATAC-Seq and ChIP-Seq:
Trimmomatic (version 0.36) (http://www.usadellab.org/cms/index.php?page = trimmomatic)
Bowtie (version 2.3.2) (http://bowtie-bio.sourceforge.net/index.shtml)
Picard (version 2.19) (http://picard.sourceforge.net)
MACS2 (version 2.1.3.3) (https://github.com/jsh58/MACS)

SV annotation:
BEDtools (version 2.26.0-114-g4c407ce) (http://bedtools.readthedocs.org/)

Transcriptome-wide association study:
EMMAX (version beta-07Mar2010) (http://genetics.cs.ucla.edu/emmax/)

Colocalization analysis of GWAS and eQTL association:
COLOC (version 5.1.0) (https://chr1swallace.github.io/coloc/)
LocusCompare (version 0.2.1) (http://locuscompare.com/)

Analyses of the Best linear unbiased prediction (BLUP) values of phenotypic data:
R package 'lme4' (version 1.1-27) (http://lme4.r-forge.r-project.org/)

The custom scripts and codes used in this study are available at Zenodo (https://doi.org/10.5281/zenodo.13365025).

A website stored and displayed the data: BnaOmics Portal (https://BnaOmics.ocri-genomics.net).

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

Data

Policy information about availability of data

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

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

All sequencing and genome assembly data used in this study have been deposited into the National Center for Biotechnology Information (NCBI) database: the long-read sequencing data and genome assembled contigs data of six B. napus accessions (BioProjects PRJNA1155214 and PRJNA1149936), short reads data of genome resequencing of 366 accessions (BioProject PRJNA1156901), population RNA-Seq data of SAM, leaves and siliques (BioProjects PRJNA1149544, PRJNA1157560 and PRJNA1153365, respectively), and the data of Hi-C sequencing, ATAC-Seq and Chip-Seq (H3K27ac) of ZY821 and ZS11 accessions (BioProject PRJNA1155718). All above data also have been deposited into the National Genomics Data Center (https://ngdc.cncb.ac.cn/?lang=en) database under the GSA Bioproject PRJCA013095. The other data generated in the previous studies are publicly available under the GSA Bioproject PRJCA002836 for RNA-Seq data of 20 DAP and 40 DAP developing seeds; the NCBI BioProjects PRJNA526961, PRJNA546246 and PRJNA587046 for the ten genome assemblies; the NCBI BioProjects SRP067370, SRP125656 and SRP155312, the GSA BioProject PRJCA002835, and the ENA Projects PRJEB5974 and PRJEB6069 for short reads data of genome resequencing of 1,739 accessions. All genome assemblies, annotations and SV information are available at the BnaOmics Portal (https://BnaOmics.ocri-genomics.net).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No human participants involved.
Reporting on race, ethnicity, or other socially relevant groupings	No human participants involved.
Population characteristics	No human participants involved.
Recruitment	No human participants involved.
Ethics oversight	No human participants involved.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	Long-read genome assembly: 16 samples; Genome NGS resequencing: 366 samples; RNA-Seq: SAM (319 sapmles), leaf (154 sapmles), 18 DAP silique (229 sapmles). We constructed a reference library of 334,461 SVs from de novo assemblies of 16 representative B. napus accessions . Selection of these accessions was subject to consideration of balancing multiple factors among worldwide geographic distribution, species diversity and breeding history of key traits, including glucosinolates, erucic acid and oil content. We detected 258,865 SVs in 2,105 genome-resequenced B. napus accessions, where 366 samples were sequenced in this study and the others were downloaded from published study (All accessions listed in Supplementary Table 1). For RNA-Seq data of B. napus population, SAM, leaf and 18 DAP silique data were sequenced in this study, and developmental seeds (20DAP and 40 DAP) were sequenced in our previous study. The accessions used for population genetic analysis represent core germplasm of this species, and the sample size are sufficient for genetics analysis.
Data exclusions	For GWAS and TWAS, we excluded the samples without phenotype data or RNA-Seq data.
Replication	We planted more than 30 plants of each accession in each replicate, with at least three replicates in each location. At least ten plants 30 plants of each accession in each replicate were sampled for phenotyping. For experiments in multiple environments, the best linear unbiased prediction (BLUP) values were estimated by R package 'lme4' (version 1.1-27) as the final phenotypic data for further analysis. Replications of repeated experiments were evaluated by proper static analyses and confirmed to be successful. All the experiments were performed using independent biological replicates as indicated in the manuscript, figure legend and supplementary information data.

Randomization	All field experiments for phenotypic evaluations were conducted in a randomized block design with at least three replicates in the six growth seasons (2012-2018). The sampling process for DNA sequencing and phenotype data collection was randomly conducted.
Blinding	Blinding was not relevant for this study and the phenotypic data were collected without known genetic data.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

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

Antibodies

Antibodies used	Anti-Histone H3 (acetyl K27) antibodies (Cat#ab4729，Abcam), RRID:AB_2118291.
Validation	Anti-H3k27ac antibodies (Cat#ab4729，Abcam): according to the manufacturer, all batches of ab4729 are tested in Peptide Array against peptides to different Histone H3 modifications. Results show strong binding to Histone H3K27ac peptide. https://www.abcam.com/products/primary-antibodies/histone-h3-acetyl-k27-antibody-chip-grade-ab4729.html

Plants

Seed stocks	366 GWAS B. napus accessions were collected from all over the world (Information of all accessions are listed in Supplementary Table 1). BnaA09.GTR2 CRISPR/Cas9 editing lines and BnaA03.MAMf transgenic lines were described in manuscript. The seeds of all these plants were stocked in Prof. Shengyi Liu lab. Further information and requests for materials resources and reagents should be directed to and will be fulfilled by the corresponding author, Shengyi Liu (liusy@oilcrops.cn).
Novel plant genotypes	For gene editing of BnaA09.GTR2 by CRISPR/Cas9, we conducted clustered regularly interspaced short palindromic repeats/CRISPR-associated system 9 (CRISPR/Cas9) genome editing against BnaA09.GTR2. Two sgRNA target sequences specifically targeting the first and second exon regions of BnaA09.GTR2 in ZY821 genome were designed. CRISPR/Cas9 plasmid construction was conducted. In brief, to assemble two gRNAs, a single PCR fragment flanked by two sgRNA targets were amplified from the pCBC-DT1T2 vector with two pairs of partially overlapping primers, among which two forward and two reverse primers respectively contain one of the two target sites. Then the PCR fragment was purified and inserted into the binary vector pHSE401 by restriction-ligation reaction using BsaI restriction enzyme and T4 Ligase (New England Biolabs). The CRISPR/Cas9 binary vector pHSE401-2gRNA-BnA09.GTR was verified by sequencing was transformed into B. napus cv. ZY821 hypocotyls by the Agrobacterium-mediated method. Genomic DNA from individual transgenic plants was extracted for PCR analysis. The PCR products containing the target sites were amplified with specific primers BnaA09GTR2-F/BnaA09GTR2-R, and then cloned into pEASY-T3 vector (TransGen Biotech) for sequencing.
Authentication	The primers used in plasmid construction and transgenic identification for both BnaA09.GTR2 CRISPR/Cas9 editing and BnaA03.MAMf transgenic are listed in Supplementary Table 18. The mutants characterization of BnaA09.GTR2 edited using CRISPR/Cas9 in ZY821 were showed in Extended Data Figure 7G and the phenotype change (seed glucosinolates contents) were showed in Figure 7F. The phenotype change of transgenic BnaA03.MAMf (leaf glucosinolates contents) were showed in Figure 3J.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	The data of Chip-Seq (H3K27ac) of ZY821 and ZS11 accessions has been deposited into the National Center for Biotechnology Information (NCBI) database under PRJNA1155718 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1155718/).
Files in database submission	1) SAMC1006523 INPUT_ZS11_silique_1 ChIP-seq of silique wall (Brassica napus cv. Zhongshuang11) 2) SAMC1006524 INPUT_ZS11_silique_2 ChIP-seq of silique wall (Brassica napus cv. Zhongshuang11) 3) SAMC1006525 INPUT_ZY821_silique_1 ChIP-seq of silique wall (Brassica napus cv. Zhongyou821) 4) SAMC1006526 INPUT_ZY821_silique_2 ChIP-seq of silique wall (Brassica napus cv. Zhongyou821) 5) SAMC1006527 IP_H3K27Ac_ZS11_silique_1 ChIP-seq of silique wall (Brassica napus cv. Zhongshuang11) 6) SAMC1006528 IP_H3K27Ac_ZS11_silique_2 ChIP-seq of silique wall (Brassica napus cv. Zhongshuang11) 7) SAMC1006529 IP_H3K27Ac_ZY821_silique_1 ChIP-seq of silique wall (Brassica napus cv. Zhongyou821) 8) SAMC1006530 IP_H3K27Ac_ZY821_silique_2 ChIP-seq of silique wall (Brassica napus cv. Zhongyou821)
Genome browser session (e.g. UCSC)	We showed the detailed results of ChIP-seq (H3K27ac) of Brassica napus cv.Zhongshuang11 and Zhongyou821 in Figure 5J in our manuscript.

Methodology

Replicates	For ChIP-seq (H3K27ac) of silique wall, we carried out two biological replicates with high overlap of ChIP-seq peaks between replicates for each accessions (Zhongshuang11 and Zhongyou821).
Sequencing depth	All ChiP-seq libraries were sequenced with Illumina with PE 150. The depth, total number of reads and uniquely mapped reads of each experiment are as follows: Sample (Library) name / Depth / CleanReads / Unique mapped reads INPUT_ZS11_silique wall_1 6.93 46,170,692 43,842,732 INPUT_ZS11_silique wall_2 9.41 62,565,416 58,490,233 INPUT_ZY821_silique wall_1 7.22 47,996,984 44,919,788 INPUT_ZY821_silique wall_2 5.99 39,880,470 36,621,646 IP_H3K27Ac_ZS11_silique wall_1 7.14 47,481,862 25,581,088 IP_H3K27Ac_ZS11_silique wall_2 5.85 38,725,842 25,633,831 IP_H3K27Ac_ZY821_silique wall_1 6.55 43,318,306 27,692,226 IP_H3K27Ac_ZY821_silique wall_2 6.04 39,912,800 16,853,942
Antibodies	Anti-Histone H3 (acetyl K27) antibodies (Cat#ab4729, Abcam), RRID:AB_2118291.
Peak calling parameters	All clean reads were mapped to ZS11 and ZY821 genomes using Bowtie (version 2.3.2) with the default parameters. PCR duplicates were removed using Picard tools (version 2.19, http://picard.sourceforge.net). Peaks of ChIP-Seq were called using callpeak module of MACS2 software (version 2.1.3.3) with the parameters of "--shift -100 --extsize 200 --board -B -g 9.6e8" and "-B -g 9.6e8".
Data quality	Low-quality reads from raw data of ATAC-Seq and ChIP-Seq were filtered out using Trimmomatic (version 0.36) with parameters "ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36". Finally, the total of peaks at FDR 5% and above 5-fold enrichment of each sample are as follows: Zhongshuang11 Chip-seq silique wall-1: 17,580; Zhongshuang11 Chip-seq silique wall-2: 48,585; Zhongyou821 Chip-seq silique wall-1: 48,663; Zhongyou821 Chip-seq silique wall-2: 43,384.
Software	Reads filtering: Trimmomatic (version 0.36) (http://www.usadellab.org/cms/index.php?page = trimmomatic); Reads mapping: Bowtie (version 2.3.2) (http://bowtie-bio.sourceforge.net/index.shtml); PCR duplicates removing: Picard (version 2.19) (http://picard.sourceforge.net); Peak calling: MACS2 (version 2.1.3.3) (https://github.com/jsh58/MACS).