ARTICLES
https://doi.org/10.1038/s41588-021-00922-y

nature genetics

OPEN

Genomic insights into the origin, domestication and diversification of *Brassica juncea*

Lei Kang[1,13], Lunwen Qian[1,2,13], Ming Zheng[3,13], Liyang Chen[4,13], Hao Chen[1], Liu Yang[1], Liang You[1], Bin Yang[1,5], Mingli Yan[6], Yuanguo Gu[7], Tianyi Wang[4], Sarah-Veronica Schiessl[8], Hong An[9], Paul Blischak[10], Xianjun Liu[11], Hongfeng Lu[4], Dawei Zhang[6], Yong Rao[5], Donghai Jia[7], Dinggang Zhou[6], Huagui Xiao[5], Yonggang Wang[7], Xinghua Xiong[1], Annaliese S. Mason[8,12], J. Chris Pires[9], Rod J. Snowdon[8], Wei Hua[3]✉ and Zhongsong Liu[1]✉

Despite early domestication around 3000 BC, the evolutionary history of the ancient allotetraploid species *Brassica juncea* (L.) Czern & Coss remains uncertain. Here, we report a chromosome-scale de novo assembly of a yellow-seeded *B. juncea* genome by integrating long-read and short-read sequencing, optical mapping and Hi-C technologies. Nuclear and organelle phylogenies of 480 accessions worldwide supported that *B. juncea* is most likely a single origin in West Asia, 8,000–14,000 years ago, via natural interspecific hybridization. Subsequently, new crop types evolved through spontaneous gene mutations and introgressions along three independent routes of eastward expansion. Selective sweeps, genome-wide trait associations and tissue-specific RNA-sequencing analysis shed light on the domestication history of flowering time and seed weight, and on human selection for morphological diversification in this versatile species. Our data provide a comprehensive insight into the origin and domestication and a foundation for genomics-based breeding of *B. juncea*.

Brassica juncea (L.) Czern & Coss is a diverse and important agricultural species[1]. An allotetraploid (AABB, $2n=36$), *B. juncea* derived from interspecific hybridization between the diploid progenitors *Brassica rapa* (AA, $2n=20$) and *Brassica nigra* (BB, $2n=16$)[2]. Four subspecies have been proposed based on crop use and morphology: *juncea* (seed mustard), *integrifolia* (leaf mustard), *napiformis* (root mustard) and *tumida* (stem mustard)[3]. *B. juncea* has a wide geographic range as native plants, adapted crops and introduced weeds, spanning the continents of Asia, Europe, Africa, America and Australia[4]. *B. juncea* is an important oilseed crop in India, Bangladesh, China and Ukraine, and is recently also gaining importance in Canada and Australia[5]. Meanwhile, it is grown as a condiment in Europe, North America, Argentina and China. Root mustard is distributed in Mongolia and northeastern China, whereas leaf mustards are most common in China and Southeast Asia[5,6].

Brassica juncea is regarded as one of the earliest domesticated plants, with mustard mentioned as a condiment in Sanskrit and Sumerian texts from as early as 3,000 BC[7]. However, its center of origin is uncertain. Based on biogeographic explorations, Vavilov[8] proposed Central Asia (Afghanistan and its contiguous regions) as the primary center of the origin of *B. juncea*, and Asia Minor, central/western China and eastern India as secondary centers of diversity. By contrast, many investigators[9–12] proposed that *B. juncea* first evolved in the Middle East where its progenitor species, *B. rapa* and *B. nigra*, are sympatric. Whether *B. juncea* has a monophyletic or polyphyletic origin is controversial. Early morphological studies proposed a single origin[13,14], whereas more detailed investigations implementing chemotaxonomy[15], nuclear DNA markers[16,17] and chloroplast (CP) genomic markers[18] suggested a polyphyletic origin. Recently, a single origin was proposed once again based on genome re-sequencing, using 109 *B. juncea* accessions[19,20]. More comprehensive studies would accelerate our understanding of either the center of origin of *B. juncea*, or the number of origin and/or domestication events that gave rise to this important crop species.

Population genomics offers an opportunity to improve our understanding of the origin and domestication of crop plants[21]. To obtain a comprehensive overview of the origin, domestication and diversification of *B. juncea*, we first generated a chromosome-scale de novo assembly of a genome of the yellow-seeded *B. juncea* var. Sichuan Yellow (SY), using PacBio long reads combined with BioNano optical mapping and Hi-C chromatin interaction maps. Subsequently, we re-sequenced 480 *B. juncea* accessions from 38 countries, leading to the identification of around 4.53 million SNPs and 0.97 million insertion–deletion polymorphisms (InDel; <50 bp). Our combined analysis of CP, mitochondrial (MT) and nuclear genome data supports a single origin of *B. juncea* in West Asia, followed by at least three independent domestication events, and the evolution of new forms through spontaneous gene mutations and introgressions during its eastward spread. We furthermore scanned

[1]College of Agronomy, Hunan Agricultural University, Changsha, China. [2]Collaborative Innovation Center of Grain and Oil Crops in South China, Hunan Agricultural University, Changsha, China. [3]Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences, Key Laboratory of Biology and Genetic Improvement of Oil Crops, Ministry of Agriculture and Rural Affairs, Wuhan, China. [4]Novogene Bioinformatics Institute, Beijing, China. [5]Guizhou Institute of Oil Crops, Guizhou Academy of Agricultural Sciences, Guiyang, China. [6]Hunan Key Laboratory of Economic Crops Genetic Improvement and Integrated Utilization, School of Life Science, Hunan University of Science and Technology, Xiangtan, China. [7]Xinjiang Academy of Agricultural Sciences, Urumqi, China. [8]Department of Plant Breeding, Justus Liebig University Giessen, Giessen, Germany. [9]Division of Biological Sciences, University of Missouri, Columbia, MO, USA. [10]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [11]College of Life Sciences, Resources and Environment Sciences, Yichun University, Yichun, China. [12]Plant Breeding Department, University of Bonn, Bonn, Germany. [13]These authors contributed equally: Lei Kang, Lunwen Qian, Ming Zheng, Liyang Chen. ✉e-mail: huawei@caas.cn; zsliu48@hunau.net

Table 1 | Summary statistics for the *Brassica juncea* var. Sichuan Yellow genome assembly

Genomic feature	SY
Estimated genome size (Mb)	1056.53
Total assembly size (bp)	933,496,244
Longest scaffold (bp)	76,001,744
Scaffold N50 (bp)	59,341,207
Contig N50 (bp)	1,926,153
Missing bases (%)	4.76
Sequences anchored to chromosome (%)	92.91
Annotated protein-coding genes (*n*)	82,723
TE proportion (%)	50.36

for selective sweeps, performed genome-wide association studies (GWAS) for flowering time and seed weight, and illuminated the domestication history and artificial selection of genes implicated in morphological diversification among diverse *B. juncea* subspecies. Our results provide a comprehensive picture of the origin and domestication history of this versatile and economically important crop species.

Results

Chromosome-scale genome of a yellow-seeded *Brassica juncea*. Yellow-seeded *B. juncea* is grown widely as a condiment and oilseed. For de novo assembly of the SY genome, we integrated four sequencing and assembly technologies: PacBio long-read sequencing, Illumina short-read sequencing, BioNano optical mapping and Hi-C data (Supplementary Fig. 1 and Supplementary Table 1). The SY genome size was estimated to be 1056.53 Mb by *k*-mer analysis (Table 1 and Supplementary Fig. 2), close to the 1,068 Mb estimated by flow cytometry[22]. PacBio reads (~93×) were first assembled using FALCON[23], followed by contig correction using Illumina reads (~130×) to generate a V.1 assembly (Supplementary Table 2). Using 202-fold coverage of BioNano data, we then generated an optical consensus map, which was implemented to assemble 1,897 super-scaffolds with an N50 of 5.87 Mb (assembly V.2). These contigs were categorized and ordered into 18 chromosome-scale scaffolds using a 15,543-marker high-density linkage map (Supplementary Fig. 3a and Supplementary Table 3). Finally, we used Hi-C data to confirm the pseudo-chromosomes and manually adjusted 165 mis-joined contigs by Juicebox[24] (Supplementary Fig. 3b,c and Supplementary Table 2). The final SY assembly captured 933.5 Mb of genome sequence, with 867.3 Mb (~92.9%) anchored into chromosomes (Fig. 1 and Supplementary Table 4), which is superior to previous assemblies of stem[19] and Indian[25] mustard in terms of genome size, contiguity and anchorage. We simultaneously assembled the CP (153,465 bp) and MT (219,803 bp) genomes of SY (Supplementary Figs. 4 and 5).

The high quality of the SY assembly was validated (Methods) by BUSCO and CEGMA scores of more than 98.5% (Supplementary Table 6), by alignment of over 95% identity with 81 randomly selected BACs and 2,567 paired BAC-end sequences[26] (Supplementary Fig. 6 and Supplementary Tables 7 and 8), by high long terminal repeat (LTR) Assembly Index (LAI)[27] of 10.73 among the assembled *Brassica* genomes (Supplementary Table 9), by high consistency with our genetic and optical maps (Supplementary Figs. 3a and 7), by consistent syntenic gene ordering (Supplementary Fig. 8) using genome-ordered graphical genotypes[28], and by the good collinearity of SY to those of *B. rapa*[29] and *B. nigra*[30] and other previously reported *Brassica* genomes[19,25,31] (Supplementary Fig. 9).

The SY assembly contained 50.36% TEs (Table 1 and Supplementary Table 10), slightly more than the published genomes of *B. juncea* T84-66 (43.5%)[19] and Varuna (45.8%)[25] and *B. rapa* (37.51%)[32], but less than *B. nigra* (53.73%)[30]. In accordance with previous *Brassica* genomes[19,25,29–33], LTR/*gypsy* retroelements were the predominant TE family (Supplementary Table 10). We distinguished the chromosomal centromeric from pericentromeric regions by specific repeats[30,34–37] (Fig. 1, Extended Data Fig. 1 and Supplementary Table 11), and remarkably lower recombination frequencies (Supplementary Fig. 3a). The centromere and pericentromeric regions were enriched for LTR/*copia* and LTR/*gypsy* elements, respectively (Fig. 1 and Supplementary Table 12).

Among 92,878 predicted gene models (Supplementary Note and Supplementary Table 13), 95.5% were functionally annotated in public databases (Supplementary Table 14). Alignment to known proteins and expression in at least one tissue type showed 82,723 gene models were high-confidence (HC) genes (Supplementary Table 15), with an average coding sequence length of ~1.13 kb and an average of five exons per gene, similarly to predictions in other *Brassica* genomes (Supplementary Table 13). A total of 5,756 genes (6.96% of the HC genes) encoded putative transcription factors belonging to 58 different families (Supplementary Table 16). We also identified 2,525 tRNAs, 8,363 rRNAs, 1,951 microRNAs and 4,691 small nuclear RNAs (Supplementary Table 17).

Population structure and genomic variation. To explore genetic variation in *B. juncea*, we re-sequenced 480 accessions representing the four subspecies from 38 countries (Fig. 2a and Supplementary Table 18) with an average depth of 15× and 97.7% of the SY genome. Using this dataset, we identified 4,529,618 high-quality SNPs and 967,266 InDels (Supplementary Table 19) based on four parameters (Methods), corresponding to 4.85 SNPs and 1.04 InDels per kb (Supplementary Table 20). A total of 946,661 SNPs (20.9%) and 50,955 InDels (5.27%) were located in coding regions. Among them, 345,138 SNPs (7.62%) caused codon changes, elongated transcripts or premature stop codons, while 27,420 InDels (2.83%) led to frameshift mutations. The SNP distribution varied across the genome depending on genome context and gene density, but was generally higher toward the telomeric chromosome regions (Supplementary Fig. 10). The A subgenome of *B. juncea* had higher nucleotide diversity ($\pi = 2.05 \times 10^{-3}$) than the B subgenome ($\pi = 1.45 \times 10^{-3}$; Supplementary Fig. 11). Moreover, linkage disequilibrium (LD) decayed faster in the A subgenome than in the B subgenome (Supplementary Fig. 12), indicating a higher degree of genetic recombination in the A subgenome of *B. juncea*.

Next, we investigated the genetic structure of the *B. juncea* population for clusters (*K*) from 2 to 10 based on 4.53 million SNPs among the 480 *B. juncea* accessions. When $K = 6$, clusters maximized the marginal likelihood (Supplementary Fig. 13). To better clarify the relationships within the population, 90 genetically admixed accessions with main genetic components of less than 60% were excluded from further analysis. Both phylogenetic and principal component analyses (PCAs) of the remaining 390 samples indicated three distinct clades (Fig. 2b,c). Clade I consisted only of root mustard from Northeast Asia. Clade II consisted of seed mustard from West Asia, Central Asia and Northwest China along the Steppe Route, a trans-Eurasian trading route predating the Silk Road[38]. Clade III included oilseed and vegetable mustards from the Indian subcontinent and southern China, corresponding to the South Silk Road connecting East and Central Asia[39].

Our phylogenetic and genetic clustering analyses resolved six *B. juncea* genetic groups (G1–G6), which largely corresponded to morphologically distinct crops (Supplementary Fig. 14 and Supplementary Table 21). G1, the root mustard group, showed the slowest LD decay, especially in the B subgenome, and strong genetic divergence from the other five groups (pairwise $F_{ST} \geq 0.33$;

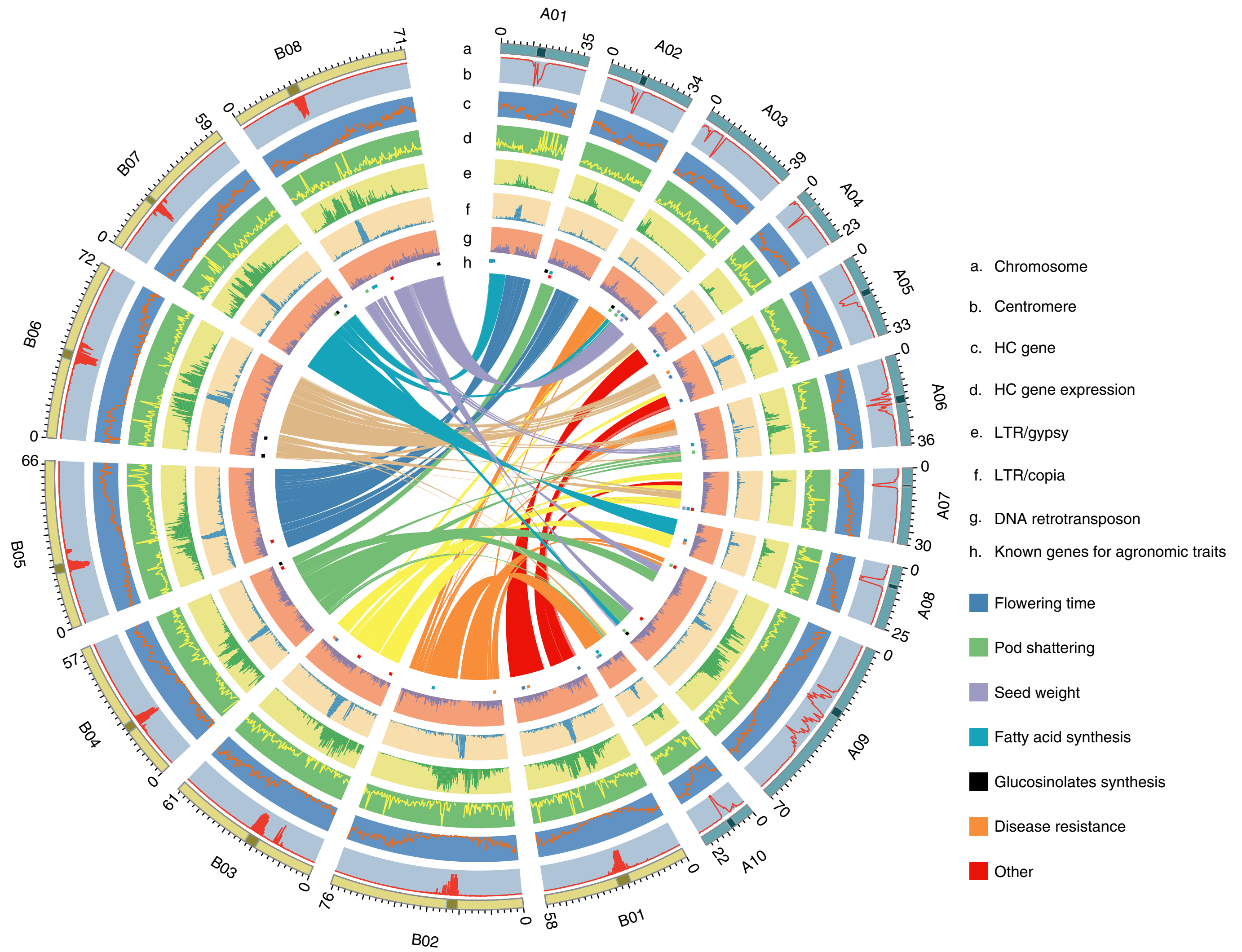


Fig. 1 | Chromosomal features and functional and synteny landscape of the yellow-seeded *B. juncea* var. SY genome. Tracks from outer (**a**) to inner (**h**) rings indicate the following: **a**, Chromosome size with units in Mb; **b**, Density of centromere-specific repeats in 5-Mb bins; **c**, Density of HC genes in 5-Mb bins; **d**, Expression of HC genes from nine tissues, calculated as the fragments per kilobase of transcript per million mapped reads (FPKM) in 5-Mb bins and normalization of FPKM by $\log_{10}$(FPKM + 1). **e**, LTR/Gypsy density (Gypsy length/5 Mb). **f**, LTR/Copia density (Copia length/5 Mb). **g**, DNA retrotransposon density (DNA retrotransposon length/5 Mb). **h**, Location of known genes (Supplementary Table 5) for major phenotypic traits. Lines in the center linking different chromosomal regions show the syntenic relationships between the A and B subgenomes.

Fig. 2e and Supplementary Tables 22 and 23). G2 comprised yellow-seeded mustard, and almost 60% of the G2 accessions with known geographic origins were from northwestern China; other G2 accessions sourced from the former Soviet Union, Canada and Europe were documented introductions from China[40–42]. G3 spanned wide geographic origins from Tibet, central and western Asia to Europe. G3 clustered close to but distinctly from G2 ($F_{ST} = 0.07$; Fig. 2d). G4 comprised mainly accessions from southwestern China and clustered closest to the G5 group. The G5 group, including 96 leaf, 14 stem and 10 seed mustards originating from southern China to Japan[43] and the USA[9,41], showed the highest nucleotide diversity ($\pi = 1.54 \times 10^{-3}$) and the greatest LD decay (Fig. 2d,e). The 59 accessions forming the group G6 were almost all from South Asia. G6 showed a similarly slow LD decay to G1, and it also exhibited the lowest nucleotide diversity ($\pi = 0.93 \times 10^{-3}$), consistent with a narrow genetic base of Indian mustard[44]. All genotypes belonging to G2 and G3 in Clade II and to G4 and G6 in Clade III are grown for seed use, whereby G2 and G3 differentiate less strongly from G4 (pairwise $F_{ST} = 0.25$ and 0.24, respectively) than from G6 (pairwise $F_{ST} = 0.42$ and 0.39, respectively; Fig. 2e and Supplementary Table 22).

Domestication and spread of *Brassica juncea*. To delineate domestication and spread, we further constructed A and B subgenome phylogenies of *B. juncea* and its progenitors (Supplementary Table 24). Both subgenome phylogenetic trees confirmed six groups of *B. juncea* and that the G1 group was the closest to the progenitor species, although G4 and G6 had the opposite positions (Fig. 3a and Supplementary Figs. 15 and 16). These nuclear phylogenies support the hypothesis that *B. juncea* originated monophyletically[19].

We assembled 478 CP and 10 MT genomes to study cytoplasmic relationships between *B. juncea* and its progenitors (Supplementary Tables 18 and 25). Based on the assembled CP genomes, we found two InDel variants and divided the *B. juncea* CP genomes into three types (CPs 1–3; Extended Data Fig. 2a and Supplementary Table 18). Meanwhile, we classified the MT genomes into three types (MTs 1–3) using an InDel and a SNP locus[45] (Extended Data Fig. 2b and Supplementary Table 18). These three MT types corresponded

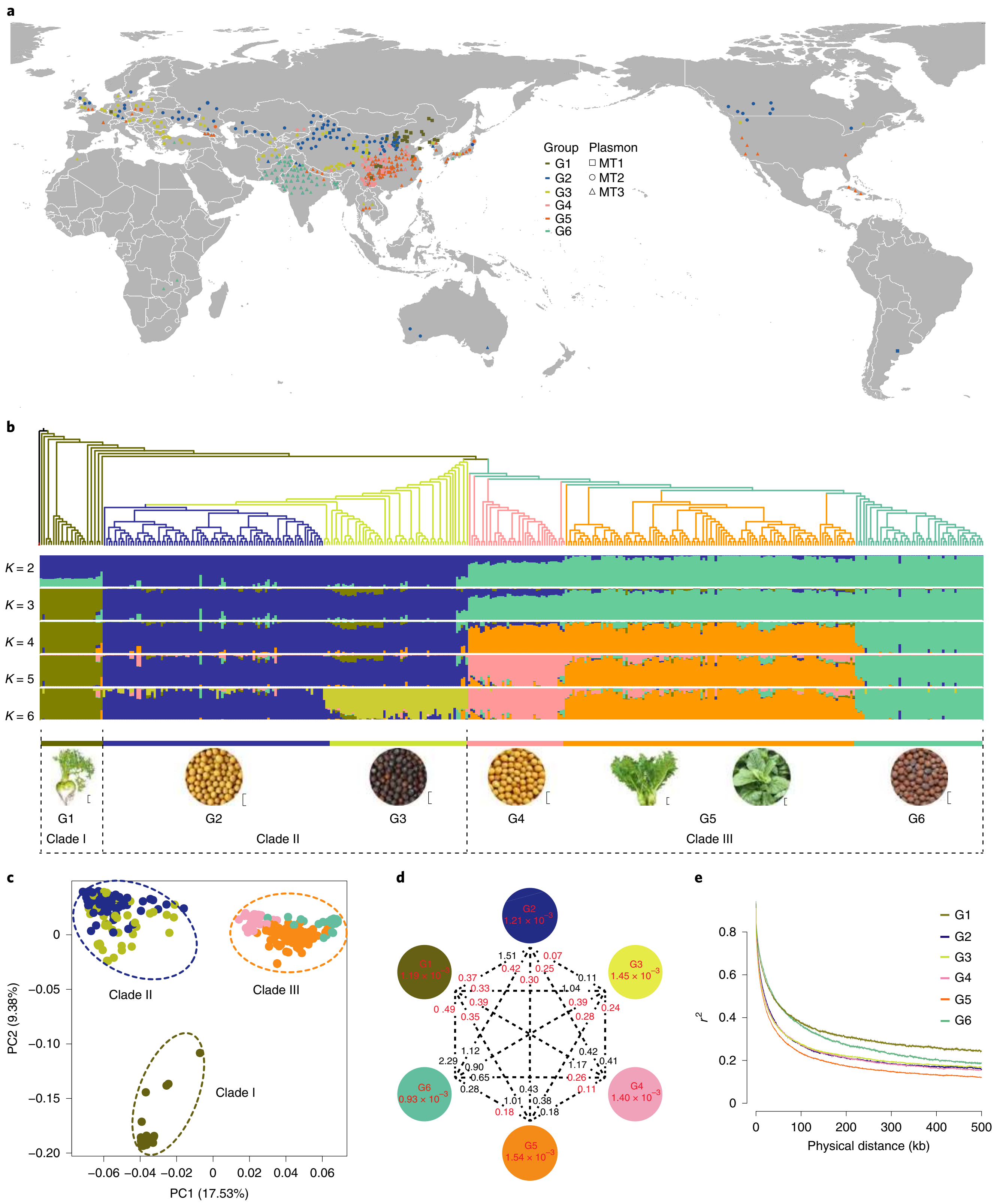


Fig. 2 | Geographic distribution, population structure and genomic diversity of *Brassica juncea* accessions. a, Geographic distributions of 480 *B. juncea* accessions. The geographic map was drawn using R ggplot2. **b**, The maximum-likelihood phylogeny of 390 *B. juncea* accessions with over 60% genetic components to the group and model-based clustering with *K* from 2 to 6. The five other *Brassicaceae* species used to root the phylogenetic tree are shown as a single branch. Branch colors indicate different groups based on the population structure. Scale bars, 5 cm for G1 and G5; 5 mm for G2, G3, G4 and G6. **c**, PCA plots showing three divergent clades of 390 *B. juncea* accessions. **d**, Nucleotide diversity (π), population divergence (F_{ST}) and genetic distance (D) across the six groups. The value in each circle represents a measure of nucleotide diversity for each group; values in red on each line indicate pairwise population divergence between groups, while values in black on each line indicate pairwise genetic distances among groups. **e**, Group-specific LD decay plots.

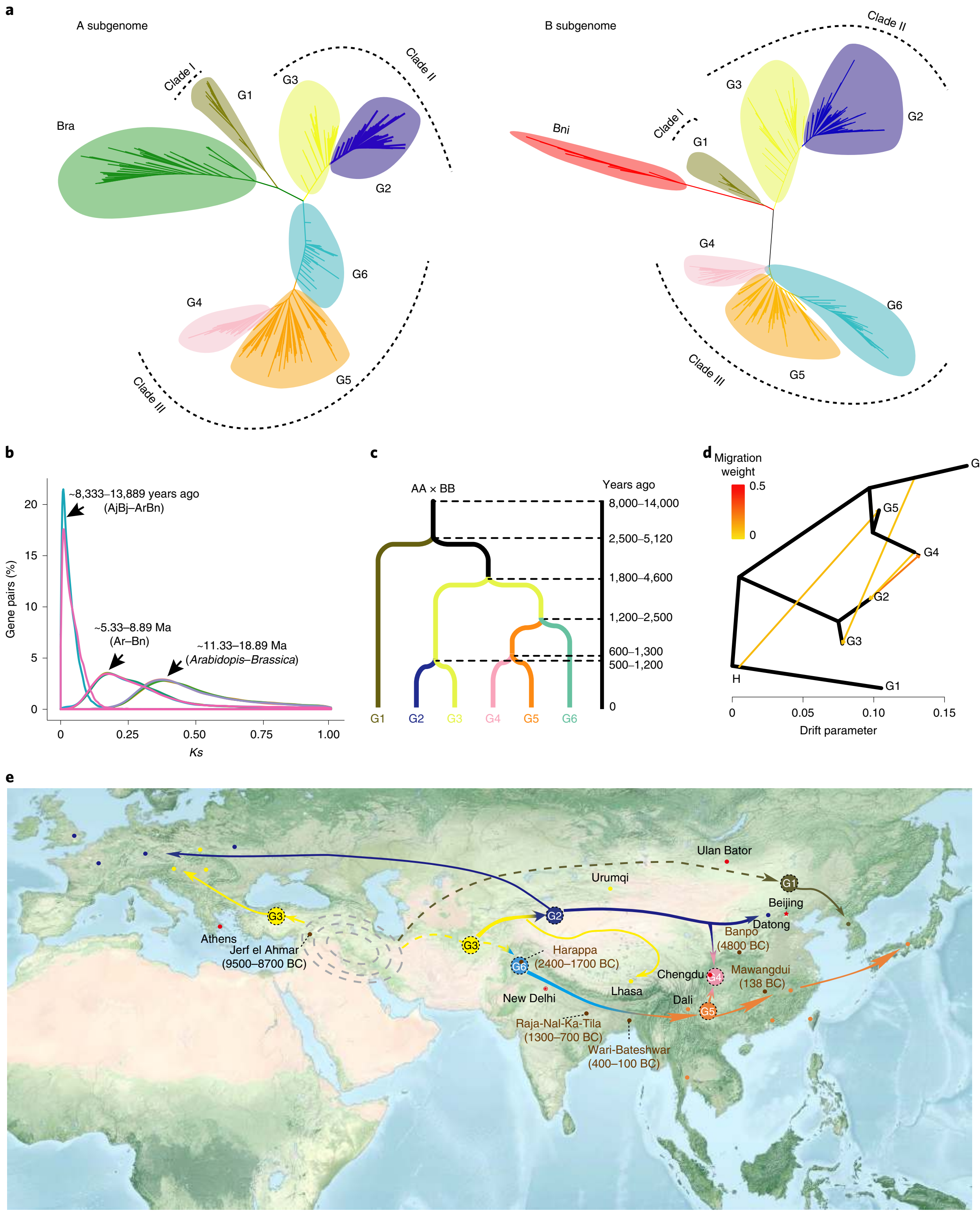


Fig. 3 | Speciation and demographic history of *Brassica juncea*. **a**, Maximum-likelihood phylogenies of the subgenomes of 390 *B. juncea* accessions compared to 68 *B. rapa* accessions (left), and 11 *B. nigra* accessions (right). **b**, Estimates of molecular divergence between *B. juncea* (AjBj) and its pseudo-ancestor (ArBn, pooled by two progenitors, *B. rapa* and *B. nigra*). **c**, Divergence time for six groups was estimated using SMC++. **d**, Detection of gene flows among *B. juncea* groups by TreeMix analysis. Arrows represent the direction of migrations. Horizontal branch length is proportional to the amount of genetic drift that has occurred on the branch. Scale bar shows ten times the average standard error of the entries in the sample covariance matrix. **e**, Putative spread routes of *B. juncea*. Archaeological evidence showing that seed cakes or carbonized mustard seeds were excavated from Jerf el Ahmar (9500–8700 BC)[54], Banpo site (about 4800 BC)[55], Harappa (2400–1700 BC)[59], Raja-Nal-ka-Tila site (1300–700 BC)[60], Wari-Bateshwa (400–100 BC)[61] and Mawangdui site (about 138 BC)[64]. The geographic map was adapted from NASA (https://visibleearth.nasa.gov/images/147190/explorer-base-map/147191w/). Ma, million years ago.

to the three specific CP classifications, and were subsequently named plasmotypes I–III. All G1 accessions carried plasmotype I, whereas all G6 and most (94.2%, 113/120) G5 accessions harbored plasmotype III. The remaining three groups contained all three plasmotypes, with plasmotype II predominating (G2 91.3%, G3 71.2%, G4 70.0%; Supplementary Table 21). In the CP phylogeny, most (467/478) of the *B. juncea* accessions were rooted in the *B. rapa* lineage (Supplementary Fig. 17), consistent with the conclusion that *B. rapa* is the maternal ancestor of *B. juncea*[46,47]. CP and MT phylogenies (Supplementary Figs. 17 and 18) and PCR analysis (Extended Data Fig. 2) indicated that plasmotype I of *B. juncea* descended from *B. rapa* and evolved into plasmotype II and III via insertion/deletions and a base substitution. From the perspective of cytoplasmic inheritance, *B. juncea* shows a single origin.

The progenitor species of *B. juncea* are sympatric in the Middle East[48]. Wild *B. juncea* forms have been observed to grow on the plateaus in Asia Minor and southern Iran[10,49–52]. The group G3, including Turkish accessions, possessed not only all three plasmotypes (Fig. 2a and Supplementary table 21) but also higher nucleotide diversity (Fig. 2d), implying that the place where the G3 accessions were collected is a plausible center of origin. Collectively, these data support that *B. juncea* most likely originated in West Asia (the Middle East).

Importantly, we estimated that *B. juncea* formed ~8,000–14,000 years ago by natural hybridization between both progenitors (Fig. 3b). A demographic history model of the *B. juncea* groups favors at least three independent evolutionary routes (Fig. 3c). Four gene flows were detected among the six groups by Treemix and *D*-statistic analyses: from root mustard (G1) to leafy mustard (G5), from Indian mustard (G6) to West and Central Asia mustard (G3), from northwestern China (G2) to southwestern China yellow-seeded mustard (G4) and, with a lower weight, in the reciprocal direction from G4 to G2 (Fig. 3d and Supplementary Table 26).

Root mustard first diverged from wild *B. juncea*, approximately 2,500–5,120 years ago (Fig. 3c). We speculate that root mustard was domesticated in Mongolia and northeastern China according to its current geographic distribution and historical records[53], although how it spread into East Asia remains elusive (Fig. 3e). Additionally, wild *B. juncea* was domesticated into the seed mustard (G3), and a diverse range of *B. juncea* accessions developed (Fig. 3c,e and Supplementary Table 18). The G3 mustard spread eastward from northern Afghanistan along the Steppe Route and entered Tibet via the Hexi corridor. During the dissemination process of G3, a new yellow-seed mustard (G2) evolved about 500 years ago from spontaneous gene mutations[56,57], probably in Xinjiang[58] (Fig. 3e). In parallel, the G3 mustard spread from southern Afghanistan into the Indian subcontinent[12] where it was domesticated into Indian mustard (G6), which is supported by archaeological excavations[59]. Indian mustard then spread eastward[60,61] to form a new type of broad-leaf mustard (var. *rugosa*)[13], probably around 300 BC[62]. These broad-leaf mustards spread further east into southwestern China, where they were grown as vegetables and oilseed before the sixth century AD[63]. Historical records documented the subsequent derivation of stem mustard from broad-leaf mustard in the Sichuan Basin in the eighteenth century[6]. Accordingly, we observed very low genetic diversity in stem mustard and a closer relationship to leaf mustard (G5) than G4 accessions from the same geographic region (Supplementary Table 27).

The G4 group inherited yellow-seed color and plasmotype II from G2, and early maturity from G5. Migration weight, *f*-branch and f_{d} values showed more genetic components were introgressed into the B subgenome than into the A subgenome from G2 to G4 (Extended Data Fig. 3), which can explain the opposite position of G4 and G6 in the A and B subgenome phylogenies (Fig. 3a). The proportions of introgressed fragments from G2 detected in the G4 accessions varied from 0.07 to 0.26, with an average of 0.159 (Supplementary Fig. 19 and Supplementary Table 28). The five largest introgressed genomic blocks (relative IBD rate > 0.7; Methods) included the regions from 49.8 to 50.8 Mb on chromosome A09 and from 39.8 to 41.8 Mb on chromosome B08, which carry *Arabidopsis thaliana TT8* (TRANSPARENT TESTA 8) orthologous genes (BjuA09g45700S and BjuB08g18790S) that are non-functional in yellow-seed *B. juncea*[56,57]. Therefore, we concluded that G4 is a genetic admixture from the natural hybridization of G2 with G5, implying that the combination of gene mutations by natural hybridization played a significant role in the domestication and spread of yellow-seeded *B. juncea*.

Ecogeographic adaptation of *Brassica juncea* flowering time. We observed flowering time variation across 390 *B. juncea* accessions grown under four contrasting environments: 94 to 194 d in Guiyang, 71 to 200 d in Xiangtan, 29 to 78 d in Kunming and 25 to 65 d in Urumqi (Supplementary Fig. 20 and Supplementary Table 29). The flowering time of 390 accessions was positively correlated across different environments ($r^2 = 0.46$ to 0.95; Supplementary Fig. 21). The broad-sense heritability of flowering time reached 0.74 (Supplementary Table 29). Most of the root mustards and some leaf mustards did not flower in Kunming, indicating vernalization failure due to insufficiently low temperatures.

We identified 43 and 38 putative selective sweeps in G6/G1 and G6/G2, respectively, containing 63 flowering time candidate genes (Fig. 4a and Supplementary Table 30). Of these genes, 30 and 7 have known roles in the photoperiod and vernalization pathways, respectively. We also scanned selective sweeps for flowering time by comparing G1 with group G2, G3, G4 or G5 and identified 42 candidate genes for flowering time (Supplementary Fig. 22). Simultaneously, a total of 56 candidate genes showed significant association to flowering time across the four environments by GWAS analysis (Supplementary Fig. 23 and Supplementary Table 31). Of these genes, 12 also detected by the selective-sweep scan were investigated in more detail as potential contributors to domestication (Supplementary Fig. 24).

Notably, two SNPs in the region of BjuA10g14550S (*SRR1*, SENSITIVITY TO RED LIGHT REDUCED 1) and five SNPs in BjuB05g31990S (*VIN3*, VERNALIZATION INSENSITIVE 3) were found to be significantly associated with flowering time (Fig. 4b,e and Supplementary Table 31). SRR1 is a pioneer protein involved in the regulation of the circadian clock and phytochrome B signaling[65], while *VIN3* is a crucial gene involved in vernalization[66]. We found strong LD between *SRR1* on chromosome A10 and *VIN3* on B05 (Extended Data Fig. 4a). The combinations of both *SRR1* and *VIN3* haplotypes were consistent with the haplotypes of either gene (Extended Data Fig. 4b,c). *SRR1*-A10-Hap1 and *VIN3*-B05-Hap1 were present in late-flowering or non-flowering accessions of the G1 group, which was domesticated in cold, long-day environments. *SRR1*-A10-Hap2 and *VIN3*-B05-Hap2 were present mostly in accessions from G2 and G3 with moderate flowering time. These seed mustard groups were domesticated under long-day conditions with large diurnal temperature variations (20–30 °C). Finally, *SRR1*-A10-Hap3 and *VIN3*-B05-Hap3 were present in the earliest-flowering accessions, mainly from G4, G5 and G6 (Fig. 4c,d,f,g and Supplementary Table 32). These results demonstrate the coevolution of *SRR1* and *VIN3* during the domestication of *B. juncea*, and support the conclusion that *B. juncea* underwent three independent domestication events.

Furthermore, a 4,597-bp insertion was found in the exon of *SRR1*. All *SRR1*-A10-Hap3 accessions have this insertion, whereas it is carried only by some (50/118) *SRR1*-A10-Hap2 accessions (Supplementary Fig. 25a,b). Comparing flowering time, we found that *SRR1*-A10-Hap2 accessions with the insertion flower earlier than those without the insertion, suggesting that this gene lost its function because of the premature termination codon produced by the insertion (Supplementary Fig. 25b,c). A 13-bp insertion in

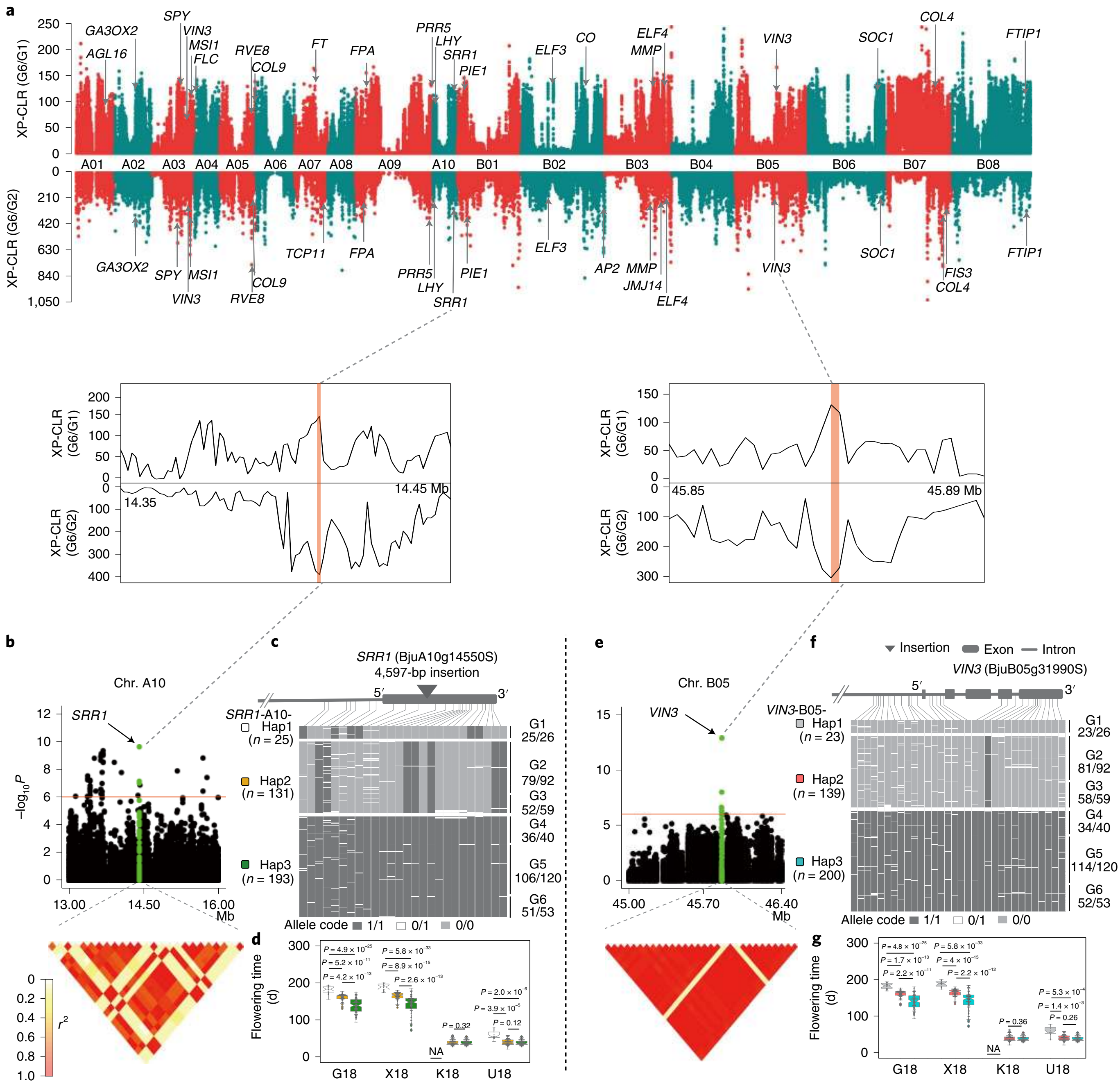


Fig. 4 | Genome-wide screening of selective sweeps and GWAS for flowering time in *Brassica juncea*. **a**, Genome-wide distribution of selective sweeps identified through comparisons between G1 or G2 with G6 using XP-CLR (cross-population composite likelihood-ratio test) values (sliding window = 10 kb, step = 1 kb). The flowering time candidate genes in the selective regions are labeled. **b**,**e**, Local Manhattan plot showing the 14.35–14.45 Mb and 45.85–45.89 Mb regions on chromosomes A10 and B05, respectively. The green plots represent the position of these SNPs in *SRR1* (BjuA10g14550S) and *VIN3* (BjuB05g31990S). Two and five SNPs in the gene regions of *SRR1* and *VIN3* were significantly associated with flowering time, respectively. Heat maps spanning the SNP markers in LD with the most strongly associated SNPs in *VIN3* and *SRR*1 gene regions. The red lines indicate the significance threshold ($-\log_{10}P = 6.0$). **c**,**f**, Three haplotypes with a frequency greater than 0.01 were identified in the *SRR1* and *VIN3* gene regions, respectively. Box plot showed three haplotypes corresponding to flowering time in *SRR1* and *VIN3* gene regions, respectively. **d**,**g**, Box plots for flowering time based on the haplotypes (Hap.) for *SRR1* (**d**) and *VIN3* (**g**) under four different environments. Box edges represent the 0.25 and 0.75 quartiles, with the median values shown by bold lines. Whiskers extend to data no more than 1.5 times the interquartile range, and remaining data are indicated by dots. *P* values were calculated using two-sided *t*-tests. NA, data missing (G1 group did not flower in Kunming).

the third intron and 6-bp deletion in the fifth exon of *VIN3* were detected in *VIN3*-B05-Hap1 and *VIN3*-B05-Hap2 (Supplementary Fig. 26a). *VIN3*-B05-Hap3 accessions have the highest relative expression level and flower earliest, while *VIN3*-B05-Hap1 and *VIN3*-B05-Hap2 accessions flower latest and show a moderate, but not significantly different, gene expression level (Supplementary Fig. 26b) because these two haplotypes differ at only a single SNP (Supplementary Fig. 27).

In addition, we identified 15 genes significantly associated with flowering time by both GWAS and selective-sweep scan

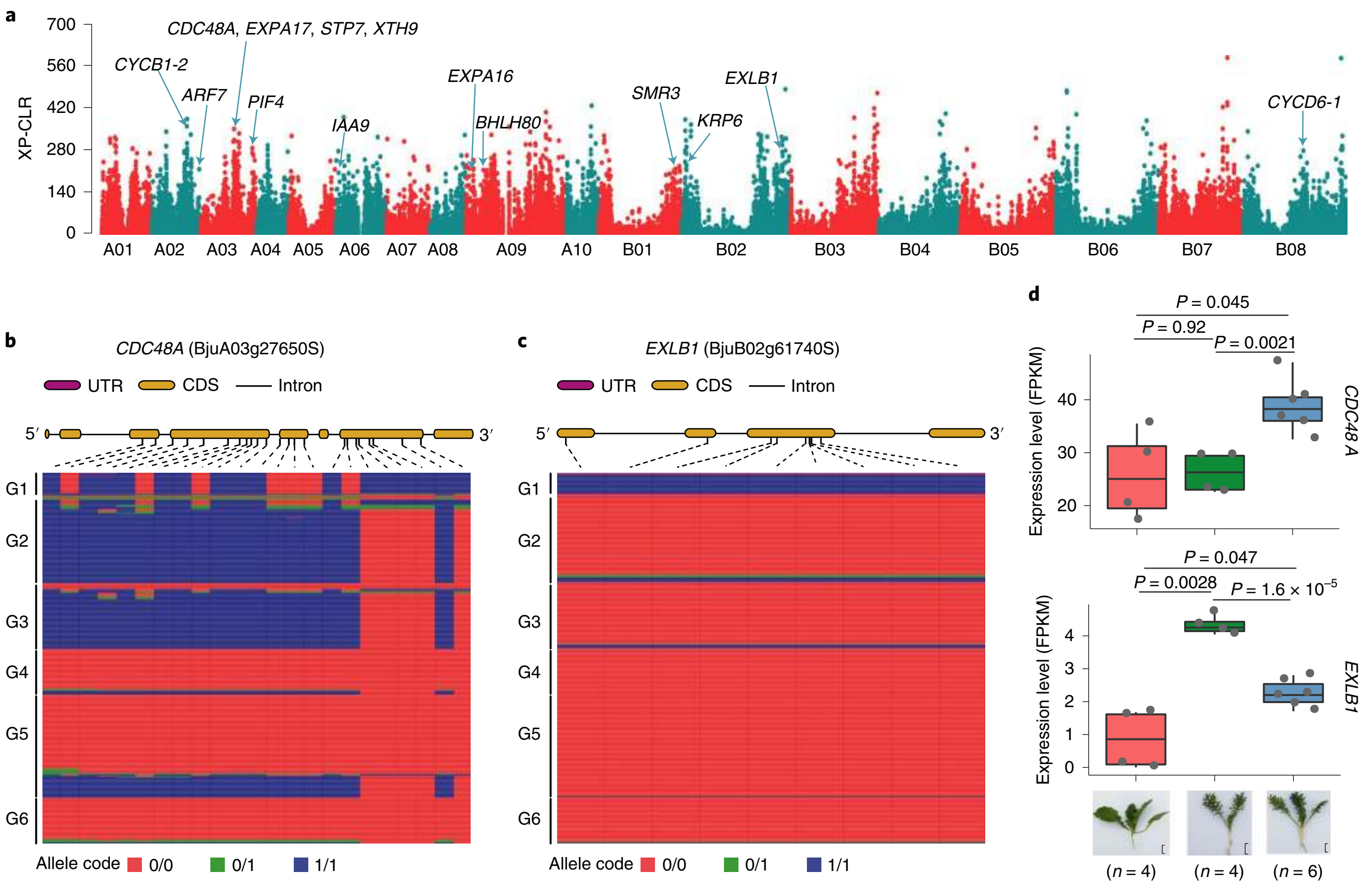


Fig. 5 | Identification of candidate genes for root enlargement in root mustard (*Brassica juncea ssp. napiformis*). **a**, Genome-wide distribution of selective sweeps related to tuber root formation in *B. juncea*. **b**, Haplotypes for the candidate gene *CDC48A* (BjuA03g27650S). **c**, Haplotypes for the candidate gene *EXLB1* (BjuB02g61740S). **d**, Expression levels of *CDC48A* and *EXLB1* in non-root and root mustard (before and 2 weeks after root enlargement) were estimated based on FPKM values. Box edges represent the 0.25 and 0.75 quartiles, with the median values shown by bold lines. Whiskers extend to data no more than 1.5 times the interquartile range, and remaining data are indicated by dots. *P* values were calculated using two-sided *t*-tests. Scale bars, 2 cm.

(Supplementary Table 33). These genes included transcription factors, SUVR and WD-40 repeat proteins, and gibberellic acid signaling, which warrant further investigation.

Genetics of morphological diversification in *Brassica juncea*. Domestication and artificial selection of *B. juncea* imparted major morphotype changes, including the increase in seed size, root expansion and stem swelling. We aimed to identify selective sweeps and genomic regions associated with each of these traits in the *B. juncea* panel.

Seed size is a primary agronomic trait that contributes to seed yield in condiment and oilseed mustards. We observed significant variation in thousand seed weight (TSW), ranging from 0.29 to 2.48 g, 0.52 to 2.94 g, 0.66 to 3.16 g and 0.96 to 4.30 g across the four environments, respectively (Supplementary Fig. 21 and Supplementary Table 29). A high broad-sense heritability of 0.92 was calculated for TSW (Supplementary Table 29). Significant positive correlations were detected across the environments, with r^2 values of 0.44–0.82 (Supplementary Fig. 21).

We identified 33 and 51 putative selective sweeps in G5/G2 and G6/G2, respectively, which contained 65 candidate genes for TSW. Among these genes, 19 overlapped between G5/G2 and G6/G2 (Supplementary Table 34). We detected 22 significantly associated candidate genes using GWAS (Supplementary Fig. 28 and Supplementary Table 35), of which 7 were also detected by selective sweeps (Supplementary Fig. 28). The two genes detected by both approaches, BjuA04g00760S (*CYP78A9*, CYTOCHROME P450 78A9) and BjuB05g28000S (*CAM7*, CALMODULIN 7; Extended Data Fig. 5b,e and Supplementary Table 35), were previously shown to regulate seed weight in *Brassica napus*[67] and *Gossypium hirsutum*[68]. Four haplotypes were detected in *CYP78A9*. *CYP78A9*-A04-Hap4 was present in 7 G3 accessions with the highest TSW, whereas *CYP78A9*-A04-Hap1 was present in 11 G5 vegetable accessions with the lowest TSW under four environments. *CYP78A9*-A04-Hap2 was mainly present in accessions from G1, G2 and G3, while *CYP78A9*-A04-Hap3 was present mainly in accessions from G4, G5 and G6. We also detected four haplotypes for *CAM7*. *CAM7*-B05-Hap1 corresponded to the G1 root mustard types with the lowest TSW, whereas *CAM7*-B05-Hap4 corresponded to 10 G2 oilseed accessions which had the highest TSW across environments. The accessions with *CAM7*-B05-Hap2 and *CAM7*-B05-Hap3 corresponded well to those with *CYP78A9*-A04-Hap2 and *CYP78A9*-A04-Hap3, respectively (Extended Data Fig. 5f,g and Supplementary Table 36).

Interestingly, Hap2 of *CYP78A9* and *CAM7* was sensitive to environments. For example, the G2 and G3 accessions of *CYP78A9*-A04-Hap2 produced heavier seeds under long-day than under short-day conditions (Supplementary Fig. 29 and Supplementary Table 36). However, they showed delayed flowering under short-day environments and produced lighter seeds than the G4, G5 and G6 accessions of *CYP78A9*-A04-Hap3. The significant increase in TSW of G2 and G3 accessions under long-day environments is a major factor causing opposing phenotypes in accessions with these two haplotypes under long-day and short-day conditions. Quantitative PCR with reverse transcription (RT–qPCR) analysis showed that both *CYP78A9* and *CAM7* were upregulated in the large-seeded accession '7981' (TSW, 2.65–4.30 g) compared to the small-sized seeds accession 'SY' (TSW, 1.40–2.46 g; Supplementary Fig. 30). Collectively, these results implicate *CYP78A9* and *CAM7* as causal genes for TSW in *B. juncea*. Haplotype analysis suggests that selection of these genes for local photoperiod adaptation induced diversification of seed size in *B. juncea*.

Meanwhile, we detected 30 genes significantly associated with TSW by both GWAS and selective-sweep scan (Supplementary Table 37). These genes included transcription factors, hormone signaling pathways, lipid transporters and ribosomal proteins, which require further investigations.

To investigate selection signatures putatively related to the domestication of root mustard, we compared the root mustard genomes to those of seed and leaf mustards using selective-sweep scan. In total, 2,803 sweep regions were identified in root mustard, covering 21.85 Mb with 5,756 genes (Supplementary Table 38). Fourteen candidate genes implicated in the formation of storage roots were identified (Fig. 5a and Supplementary Table 39), with putative functions in auxin signaling, sugar transport, cell division, cell expansion and cell wall modification. Of these, *CDC48A4* (BjuA03g27650S), participating in cell division and growth[69], was found to have three haplotypes corresponding to the three independent domestication events (Fig. 5b). Its expression was upregulated during root enlargement in root mustard (Fig. 5d). The root and non-root mustards carried distinctly different haplotypes of the expansin gene *EXPB1* (BjuB02g61740S; Fig. 5c). Its expression was downregulated after root enlargement in root mustard (Fig. 5d), which is consistent with the expression patterns of *EXPB1* in *Raphanus sativus*[70] and *Ipomoea batatas*[71] during storage root development. We observed similar expression patterns in another expansin gene, *EXPA16* (BjuA09g18260S), and the cell elongation gene *XTH9* (BjuA03g32220S) after root enlargement in root mustard (Supplementary Table 39).

Stem mustard is characterized by its enlarged edible stem with a diameter of > 20 cm, much bigger in diameter than leaf mustard (usually <5 cm[72]). We compared genomes of stem and leaf mustards and identified a total of 5,018 selective sweeps, spanning 46.51 Mb (Extended Data Fig. 6 and Supplementary Table 40). Twelve candidate genes selected during stem mustard breeding (Supplementary Table 41) are implicated in cell division, cell expansion, regulation of auxin signaling and glucose transport, functions with reported roles in storage organ formation in *Brassica*[73]. Expression of BjuA05g02460S, orthologous to *GRF7* (GROWTH-REGULATING FACTOR 7) regulating leaf and stem development[74], was upregulated during stem swelling (Extended Data Fig. 6b,d), while the genes encoding auxin-responsive protein, *IAA33* (BjuA10g12920S), and the auxin-response factor, *MP* (also known as *ARF5*, (BjuB03g51870S), were downregulated after stem swelling (Extended Data Fig. 6c,d and Supplementary Table 41). This result contrasts with reports in turnip (*B. rapa* ssp. *rapa*)[75], where expression of auxin-response genes did not change significantly during hypocotyl expansion. Overall, a greater subgenomic prevalence of selective sweeps related to root and stem swelling suggests that the A subgenome has undergone stronger selection than the B subgenome (Supplementary Tables 38 and 40). This finding is consistent with the high morphotype diversity of *B. rapa*[73], which putatively provides a better selective substrate than the narrower range of variation present in *B. nigra*.

Discussion

SY is a yellow-seeded landrace of *B. juncea* and represents a new form evolved from hybridization between two big gene pools. Therefore, SY is different from previously sequenced stem[19] and Indian[25] mustard. The chromosome-scale reference genome of SY, in conjunction with re-sequencing of 480 accessions, captured major genetic variation and allowed detailed reconstruction of the evolutionary and domestication history of this diverse ancient crop species. Plant genomics, together with archaeological evidence and historical written records, likely indicated a monophyletic origin of *B. juncea* in West Asia 8,000–14,000 years ago and at least three subsequent independent domestication events in the last 500–5,000 years: seed mustard near Central Asia, oilseed mustard in the Indian subcontinent and root mustard in East Asia. As *B. juncea* spread eastward, yellow-seeded (Oriental) mustard arose in Northwest China, stem mustard in the Sichuan Basin and probably broad-leaf mustard in eastern India, by selection acting on via spontaneous mutations. Hybridization of leaf mustard with yellow-seeded and root mustard gave rise to early-maturing yellow-seeded mustard in the Yunnan–Kweichow Plateau and lobed-leaf mustard (var. multisection Bailey) in eastern China, respectively. We also identified underlying genes and causal alleles for morphological variants such as root and stem swelling, flowering time and seed size variation associated with domestication and diversification. Our results not only elucidate the complex evolutionary and domestication history of *B. juncea*, but also pave the way for future research and breeding of this morphologically diverse condiment, oilseed, leaf, stem and root vegetable species.

Online content

Any methods, additional references, Nature Research 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/s41588-021-00922-y.

Received: 13 July 2020; Accepted: 23 July 2021;
Published online: 6 September 2021

References

1. Vaughan, J. G. & Hemingway, J. S. The utilization of mustards. *Econ. Bot.* **13**, 196–204 (1959).
2. Nagaharu, U. Genomic analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jpn. J. Bot.* **7**, 389–452 (1935).
3. Gladis, T. & Hammer, K. The *Brassica* collection in Gatersleben: *Brassica juncea*, *Brassica napus*, *Brassica nigra* and *Brassica rapa*. *Feddes Rep.* **103**, 469–507 (1992).
4. Spect, C. E. & Diederichsen, A. *Brassica* in *Mansfeld's Encyclopedia of Agricultural and Horticultural Crops* (ed. Hanelt, P.) **3**, 1453–1456 (Springer Press, 2001).
5. Dixon, G. R. Origins and Diversity of *Brassica* and its relatives in *Vegetable Brassicas and Related Crucifers* (ed. Dixon, G.R.) 1–34 (CABI Press, 2007).
6. Chen, S. R. The origin and differentiation of mustard varieties in China. *Cruciferae Newsl.* **7**, 7–10 (1982).
7. Hemingway, J. The mustard species: condiment and food ingredients use and potential as oilseed crops in *Brassica Oilseeds: Production and Utilization* (eds. Kimber, D. S. & McGregor, D. I.) 373–383 (CAB Press, 1995).
8. Vavilov, N. I. Phytogeographic basis of plant breeding. *Chronica Bot.* **13**, 14–56 (1951).
9. Bailey, L. H. The cultivated *Brassicas*. Second paper. *Gentes Herb.* **2**, 211–267 (1930).
10. Mizushima, U. & Tsunoda, S. A plant exploration in *Brassica* and allied genera. *Tohoku J. Agric. Res.* **17**, 249–277 (1967).
11. Sun, V. G. Breeding plants of *Brassica*. *J. Agron. Assoc. China* **71**, 141–152 (1970).
12. Hinata, K. & Prakash, S. Ethnobotany and evolutionary origin of Indian oleiferous *Brassicae*. *Indian J. Genet.* **44**, 102–112 (1984).
13. Prain, D. The mustards cultivated in Bengal. *Agric. Ledger* **5**, 1–80 (1898).
14. Sinskaia, E. N. The oleiferous plants and root crops of the family Cruciferae. *Bull. Appl. Bot. Genet. Plant Breed.* **19**, 555–648 (1928).
15. Vaughn, J. G., Hemmingway, J. S. & Schofield, H. J. Contributions to a study of variation in *Brassica juncea* Czern and Coss. *J. Linn. Soc.* **58**, 435–447 (1963).

16. Song, K. M., Osborn, T. C. & Williams, P. H. *Brassica* taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs). 1. Genome evolution of diploid and amphidiploid species. *Theor. Appl. Genet.* **75**, 784–794 (1988).
17. Chen, S. et al. Evidence from genome-wide simple sequence repeat markers for a polyphyletic origin and secondary centers of genetic diversity of *Brassica juncea* in China and India. *J. Hered.* **104**, 416–427 (2013).
18. Kaur, P. et al. Polyphyletic origin of *Brassica juncea* with *B. rapa* and *B. nigra* (*Brassicaceae*) participating as cytoplasm donor parents in independent hybridization events. *Am. J. Bot.* **101**, 1157–1166 (2014).
19. Yang, J. et al. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nat. Genet.* **48**, 1225–1232 (2016).
20. Yang, J. et al. Chinese root-type mustard provides phylogenomic insights into the evolution of the multi-use diversified allopolyploid *Brassica juncea*. *Mol. Plant* **11**, 512–514 (2018).
21. Larson, G. et al. Current perspectives and the future of domestication studies. *Proc. Natl Acad. Sci. USA* **111**, 6139–6146 (2014).
22. Johnston, J. S. et al. Evolution of genome size in *Brassicaceae*. *Ann. Bot.* **95**, 229–235 (2005).
23. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
24. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
25. Paritosh, K. et al. A chromosome-scale assembly of allotetraploid *Brassica juncea* (AABB) elucidates comparative architecture of the A and B genomes. *Plant Biotechnol. J.* **19**, 602–614 (2021).
26. Liu, X. et al. Genome-wide identification, localization and expression analysis of proanthocyanidin-associated genes in *Brassica*. *Front. Plant Sci.* **7**, 1831 (2016).
27. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
28. He, Z. & Bancroft, I. Organization of the genome sequence of the polyploid crop species *Brassica juncea*. *Nat. Genet.* **50**, 1496–1497 (2018).
29. Belser, C. et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* **4**, 879–887 (2018).
30. Perumal, S. et al. A high-contiguity *Brassica nigra* genome localizes active centromeres and defines the ancestral *Brassica*. *Genome Nat. Plants* **6**, 929–941 (2020).
31. Song, M. J. et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* **6**, 34–45 (2020).
32. Zhang, L. et al. Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. *Hortic. Res.* **5**, 50 (2018).
33. Chalhoub, B. et al. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**, 950–953 (2014).
34. Lim, K. B. et al. Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassica rapa*. *Mol. Cells* **19**, 436–444 (2005).
35. Lim, K. B. et al. Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related *Brassica* species. *Plant J.* **49**, 173–183 (2007).
36. Schelfhout, C. J., Snowdon, R., Cowling, W. A. & Wroth, J. M. A PCR-based B-genome-specific marker in *Brassica* species. *Theor. Appl. Genet.* **109**, 917–921 (2004).
37. Wang, G. et al. ChIP-cloning analysis uncovers centromere-specific retrotransposons in *Brassica nigra* and reveals their rapid diversification in *Brassica* allotetraploids. *Chromosoma* **128**, 119–131 (2019).
38. Christian, D. Silk roads or steppe roads? The silk roads in world history. *J. World Hist.* **11**, 1–26 (2000).
39. Wu, X. M. et al. Genetic diversity in oil and vegetable mustard (*Brassica juncea*) landraces revealed by SRAP markers. *Genet. Resour. Crop Evol.* **56**, 1011–1022 (2009).
40. Pustovoit, V. S. Indian mustard in *Handbook of Selection and Seed Growing of Oil Plants* (ed. Pustovoit, V. S.) 149–205 (Israel Program for Scientific Translations, 1973).
41. Musil, A. F. Distinguishing the species of *Brassica* by their seed. USDA Misc. Publ. No. 643 1–35 (1948).
42. Oram, R. N. et al. Breeding Indian mustard *Brassica juncea* (L.) Czern for cold-pressed, edible oil production: a review. *Aust. J. Agric. Res.* **56**, 581–596 (2005).
43. Hoshikawa, K. Mustard in *The Origin and Propagation of Cultivated Plants* (ed. Hoshikawa, K.) 92–93 (Ninomiya Syoten Press, 1998).
44. Chauhan, J. S., Singh, K. H., Singh, V. V. & Kumar, S. Hundred years of rapeseed-mustard breeding in India: accomplishments and future strategies. *Indian J. Agr. Sci.* **81**, 1093–1109 (2011).
45. Hatono, S., Nishimura, K., Murakami, Y., Tsujimura, M. & Yamagishi, H. Complete mitochondrial genome sequences of *Brassica rapa* (Chinese cabbage and mizuna), and intraspecific differentiation of cytoplasm in *B. rapa* and *Brassica juncea*. *Breed. Sci.* **67**, 357–362 (2017).
46. Li, P. et al. A phylogenetic analysis of chloroplast genomes elucidates the relationships of the six economically important *Brassica* species comprising the triangle of U. *Front. Plant Sci.* **8**, 111 (2017).
47. Chang, S. et al. Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in *Brassica*. *BMC Genomics* **12**, 497 (2011).
48. Tsunoda, S. Eco-physiology of wild and cultivated forms in *Brassica* and allied genera in *Brassica Crops and Wild Allies* (eds. Tsunoda, S. et al.) 109–120 (Japan Scientific Societies Press, 1980).
49. Olsson, G. Species crosses within the genus *Brassica* I. Artificial *Brassica juncea* Coss. *Hereditas* **46**, 171–222 (1960).
50. Tsunoda, S. & Nishi, S. Origin, differentiation and breeding of cultivated *Brassica*. *Proc. XII Int. Congr. Genet.* **2**, 77–88 (1968).
51. Kayaçetin, F. Morphological characterization and relationships among some important wild and domestic Turkish mustard genotypes (*Brassica spp*.). *Turk. J. Bot.* **43**, 499–515 (2019).
52. Dönmeza, A. A., Aydına, Z. U. & Wang, X. W. Wild *Brassica* and its close relatives in Turkey, the genetic treasures. *Hort. Plant J.* **7**, 97–107 (2021).
53. Wang, S. M. & Shu, G. G. in *Explanations of Cucurbits and Vegetable Crops* 11, 1576–1588 (The Commercial Press, 1937)..
54. Willcox, G. Charred plant remains from a 10th millenium B.P. kitchen at Jerf el Ahmar (Syria). *Veget. Hist. Archaeobot.* **11**, 55–60 (2002).
55. Institute of Archaeology of Chinese Academy of Sciences. Xian Banpo country. 223 (Special issue of Archaeology, Archaeology Press, 1963).
56. Liu, X. et al. Inheritance, mapping, and origin of yellow-seeded trait in *Brassica juncea*. *Acta Agron. Sin.* **35**, 839–847 (2009).
57. Liu, Z. et al. Domestication and molecular mechanism underlying yellow seed in *Brassica juncea* Czern & Coss. 131 (14th International Rapeseed Congress, Saskatoon, Canada, 2015).
58. Vavilov, N. I. *Origin and Geography of Cultivated Plants* (translated by Love, D.) (Cambridge Univ. Press, 1992).
59. Hutchinson, J. B. India: local and introduced crops. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **275**, 129–141 (1976).
60. Pokharia, A. K. et al. Neolithic Early historic (2500–200 BC) plant use: the archaeobotany of Ganga Plain, India. *Quatern. Int.* **443**, 223–237 (2017).
61. Rahman, M., Castillo, C. C., Murphy, C., Rahman, S. M. & Fuller, D. Q. Agricultural systems in Bangladesh: the first archaeobotanical results from Early Historic Wari-Bateshwar and Early Medieval Vikrampura. *Archaeol. Anthropol. Sci.* **12**, 37 (2020).
62. Prakash, S., Wu, X. & Bhat, S. R. History, evolution and domestication of *Brassica* crops. *Plant Breed. Rev.* **35**, 19–84 (2012).
63. Jia, S. X. & Shu, Q. M. Y. *Important Arts for the Peoples' Welfare* (Shanghai Classics Publishing House, 2009).
64. Liu, Z. M. The origin and development of cultivated rice in China. *Acta Genet. Sin.* **2**, 23–29 (1975).
65. Staiger, D. et al. The *Arabidopsis SRR1* gene mediates phyB signaling and is required for normal circadian clock function. *Genes Dev.* **17**, 256–268 (2003).
66. Sung, S. & Amasino, R. M. Vernalization in *Arabidopsis thaliana* is mediated by the PHD finger protein VIN3. *Nature* **427**, 159–164 (2004).
67. 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).
68. Cheng, Y. et al. GhCaM7-like, a calcium sensor gene, influences cotton fiber elongation and biomass production. *Plant Physiol. Biochem.* **109**, 128–136 (2016).
69. Rancour, D. M., Park, S., Knight, S. D. & Bednarek, S. Y. Plant UBX domain-containing protein 1, PUX1, regulates the oligomeric structure and activity of *Arabidopsis* CDC48. *J. Biol. Chem.* **279**, 54264–54274 (2004).
70. Xie, Y. et al. Comparative proteomic analysis provides insight into a complex regulatory network of taproot formation in radish (*Raphanus sativus* L.). *Hortic. Res.* **5**, 51 (2018).
71. Noh, S. A. et al. Down-regulation of the *IbEXPB1* gene enhanced storage root development in sweet potato. *J. Exp. Bot.* **64**, 129–142 (2013).
72. Shi, H. et al. Cell division and endoreduplication play important roles in stem swelling of tuber mustard (*Brassica juncea* Coss. var. *tumida* Tsen et Lee). *Plant Biol.* **14**, 956–963 (2012).
73. Cheng, F. et al. Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. *Nat. Genet.* **48**, 1218–1224 (2016).
74. Wang, F. et al. Genome-wide identification and analysis of the growth-regulating factor family in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). *BMC Genomics* **15**, 807 (2014).
75. Liu, M. et al. What makes turnips: anatomy, physiology and transcriptome during early stages of its hypocotyl-tuber development. *Hortic. Res.* **6**, 38 (2019).

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

Methods

Genome sequencing and assembly. *Genome sequencing.* High-molecular-weight DNA was isolated from fresh young leaves of *B. juncea* ssp. *juncea* var. SY. A SMRTbell library constructed with Sequel 1.0 reagents was sequenced on the PacBio Sequel. Illumina paired-end libraries of 350 bp in length were prepared following the manufacturer's protocol. Hi-C libraries were performed as previously described[76]. Hi-C libraries were controlled for quality and sequenced on the Illumina HiSeq X Ten platform. Total RNA samples were extracted from root, stem, leaf, flower bud, siliques (7 and 15 d post-anthesis (DPA)), pod wall (20 DPA) and seed (20 DPA). RNA-sequencing (RNA-seq) libraries were made using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB) following the manufacturer's recommendations and also sequenced on the Illumina X Ten platform.

Optical mapping. High-molecular-weight DNA extracted using the BioNano Plant Tissue DNA isolation kit (BioNano Genomics) was digested by Nt.BspQI and labeled with IrysPrep Labeling mix. The labeled DNA sample was loaded on the IrysChip and imaged using the BioNano Irys System.

Construction of a high-density Brassica juncea genetic map. A set of 172 recombinant inbred lines were derived from the cross SY × Purple Leaf Mustard (PM). Genomic DNA extracted from recombinant inbred line individual plants were digested with MseI. The fragments between 330 and 550 bp were gel excised and eluted. The pooled libraries were amplified and sequenced on a HiSeq 2000 platform. After stringent filtering, a total of 51,018 SNPs were identified in 21,210 genotyping-by-sequencing tags using the UNEAK pipeline[77]. To map the reads, the published *B. juncea* 'T84-66' genome (http://brassicadb.cn/#/SearchJBrowse/?Genome=Bju15/) was used as the reference. Genotyping of recombinant inbred lines was performed using a hidden Markov mode[78] and the genetic map was constructed using MSTMap[79].

De novo genome assembly. The genome size of SY was estimated by Jellyfish (v.2.2.9)[80] using the *k*-mer of 17. After low-quality PacBio subreads shorter than 500 bp or with a quality score lower than 0.8 were filtered out, clean PacBio subreads were error-corrected and assembled into contigs by FALCON[23] with the parameters --max_diff 100, --max_cov 100 and --min_cov 3, and then connected to scaffolds using Sspace-longread (v.1.1)[81]. After filling gaps using PacBio reads with PBJelly (v.1.9.1)[82], gap-closed scaffolds were polished by Quiver[83] and Pilon[84] software with PacBio reads and Illumina data, respectively.

Scaffolding by integrating BioNano optical map. High-quality labeled molecules were pairwise aligned, clustered and assembled into contigs following the BioNano Genomics assembly pipeline. The BioNano Solve (V3.1) pipeline module 'HybridScaffold' was used to perform the hybrid assembly between the initial scaffold sequences and BioNano-assembled genome maps with the one-enzyme method. Using 202-fold coverage of BioNano data, we then generated an optical consensus map, which was implemented to assemble 1,897 super-scaffolds with an N50 of 5.87 Mb (assembly v2). Visualization of alignments between genome sequences and BioNano optical maps was performed by BioNano Access software (v1.5.1).

Pseudo-chromosomes assembly using the high-density genetic map. For pseudo-chromosomes assembly, markers of the high-density *B. juncea* genetic map were aligned to SY assembly V.2 by BWA (v. 0.7.8)[85] mem. We set a threshold of at least three linked markers to order and orientate the contigs. Contigs which showed conflicts to the genetic map were called as potential mis-joins and checked based on marker continuity. A total of 35 mis-joins were found in 2,329 contigs and split to give 2,364 contigs after correction (Supplementary Table 2). Subsequently, the software Chromonomer (v.1.07, http://catchenlab.life.illinois.edu/chromonomer/manual/) was used to construct the initial pseudo-chromosomes of SY, with default parameters, following the internationally agreed nomenclature for *Brassica* chromosomes (http://www.brassica.info/resource/maps/lg-assignments.php).

Pseudo-chromosomes validation using Hi-C. To avoid artificial bias, the following type of reads were removed: (1) reads with ≥10% unidentified nucleotides (N); (2) reads with >10 bp aligned to the adaptor, allowing ≤10% mismatches; (3) reads with >50% bases having phred quality < 5. The filtered Hi-C reads were aligned to the initial pseudo-chromosome genome by BWA (v0.7.8)[85] with default parameters. Reads were excluded from subsequent analysis if they did not align within 500 bp of a restriction site. Only uniquely mapped reads and valid paired-end ditags were used to validate the pseudo-chromosome sequences. The scaffolds of assembly V3 were used to make the Hi-C map by HiCPlotter[86], and the interaction matrix of each chromosome was visualized with heat maps at the 25-kb resolution. A total of 165 mis-joined contigs were identified and manually broken using Juicebox[24] according to the discrete chromatin interaction pattern. Of these, 150 mis-joined contigs, which lacked sufficient linked markers (three or more per contig or subcontig), were corrected and ordered by Hi-C contact map. Next, 13 mis-joins showing conflicts between the results of Hi-C data and the high-density map were broken, then re-clustered and ordered according to the Hi-C contact signal. Two remaining unanchored contigs that could not be anchored by the genetic map were repositioned to their pseudo-chromosome based on the Hi-C data.

Assessment of SY genome quality. The 1,440 conserved protein models in the BUSCO embryophyta_odb9 dataset (https://busco.ezlab.org/frame_wget.html) and the 248 conserved protein models in the CEGMA dataset (http://korflab.ucdavis.edu/dataseda/cegma/) were searched against the SY genome by using the BUSCO (v2)[87] and the CEGMA (v. 2.5)[88] programs with default parameters. Eighty-one BAC sequences and 2,567 BAC-end sequences from the PM BAC library were aligned to the SY genome by LASTZ[89] with parameters (M = 254, K = 4,500, L = 3,000, Y = 15,000; --seed = match12 --step = 20 --identity = 85). LTRharvest[90] (with parameters --similar 85.00 --vic 10 --seed 30 --seqids yes --motif TGCA --motifmis 1 --minlenltr 100 --maxlenltr 3,500 --mindistltr 1,000 --maxdistltr 20,000 --mintsd 4 --maxtsd 20) and LTR_FINDER[91] (with parameters: --w 2 --l 100 --L 3,500 --d 1,000 --D 20,000 --M 0.3) were used to de novo predict the candidate LTR-RTs (full-length LTRs retrotransposon) in the SY assembly sequences. LTR_retriever[92] was then used to combine and refactor all the candidates to get the final full-length LTR-RTs. LAI[27] was calculated based on the formula: LAI = (intact LTR-RTs length/total LTR-RTs length) × 100%. As recommended by the steering group of the Multinational *Brassica* Genome Project (https://www.brassica.info/), the consistency of syntenic gene ordering was evaluated by exploiting the linkage mapping information depicted by the genome-ordered graphical genotypes[28]. Protein sequences of annotated HC genes from *B. juncea* vars. SY, T84-66 (ref. [19]) and Varuna[25], both progenitors *B. rapa*[29] and *B. nigra*[30], and previously reported *B. napus* cv. ZS11 (ref. [31]) were reciprocal aligned using BLASTP with an *E*-value cutoff of 1e^{-5}. The reciprocal best hit for each alignment was used to build whole-genome synteny between SY and the other five *Brassica* subgenomes by MCScanX[93].

Detailed procedures for the SY genome annotation are provided in the Supplementary Note.

Genome blocks and centromere detection. We first constructed the three subgenomes (LF, MF1 and MF2) following methods described previously[94]. Then, we defined the genomic blocks in SY based on the syntenic relationship of the *B. juncea* and *A. thaliana* genomes[95]. We aligned the A subgenome centromeric repeat sequences (CentBrs, CRB and TR805)[34,35] and the B subgenome centromeric repeat sequences (CRB, pBNBH35 and CLs)[30,35–37] to the SY assembly using BLAST (*E*-value 1e^{-5}). The pericentromeric regions of A subgenome were detected using peri-centromere-specific retrotransposons and the tandem repeat sequence TR238 (ref. [35]), whereas the pericentromeric regions of B subgenome contained more LTR/*gypsy* elements[30]. Then, the densities of centromeric repeat sequences were calculated to detect the centromere locations.

Re-sequencing, reads mapping and SNP calling. A panel of 480 mustard accessions (Supplementary Table 18) were self-pollinated over multiple generations before re-sequencing. Genomic DNA extracted from fresh leaves was used for 350-bp Illumina libraries preparation. Sequencing protocols were the same as mentioned above. A total of 7.01 Tb (~14.48 Gb per sample) of clean data was generated after removing reads with ≥10% unidentified nucleotides (N), >10 nucleotides aligned to the adaptor or of which >50% bases had Phred quality scores less than 5. The paired-end reads were mapped to the SY genome using BWA (v0.7.8)[85] with the command 'mem --t 4 --k 32 --M'. Duplicated reads were removed with SAMtools (v.0.1.19)[96]. The genomic variants for each accession were then identified with the HaplotypeCaller module and the GVCF model by Genome Analysis Toolkit[97] (GATK) software. All the GVCF files were merged. The high-quality SNPs and InDels were created in the HaplotypeCaller module filtered with the following four parameters: depth for individual ≥ 4, genotype quality for individual ≥ 5, minor allele frequency (MAF) ≥ 0.05, with missing rate ≤ 0.1 and heterozygous rate < 0.1. The identified SNPs and InDels were further annotated with ANNOVARtool (v2013-05-20)[98], and divided into the following groups: variations occurring in intergenic regions, within 1 kb upstream (downstream) of transcription start (stop) sites, in coding sequences and in introns.

Population structure and phylogenetic analyses. The population genetic structure was examined using the program ADMIXTURE (v1.23)[99] with *K* values (the putative number of populations) from 2 to 10. The *K* = 6 value was chosen because clusters maximized the marginal likelihood. To better clarify the relationships of *B. juncea* accessions, 390 accessions with the genetic components of larger than 0.6 were retained for the further analysis. To construct maximum-likelihood phylogeny, we screened 30,609 synonymous SNPs to reduce influences of natural or artificial selection. Phylogenetic tree analysis was performed using IQ-TREE (v1.6.6)[100], based on the best model (GTR + F + ASC + R7) determined by the Bayesian information criterion. Bootstrap support values were calculated using the ultrafast bootstrap approach (UFboot) with 1,000 replicates. Five known closely related species *A. thaliana*, *Crambe hispanica*, *Cardamine hirsuta*, *Eutrema halophilum* and *Eutrema salsugineum* were used as outgroups. The phylogenetic tree was visualized by the online tool EvolView (https://www.evolgenius.info//evolview/). PCAs were done by GCTA[101]. The population relatedness and migration events were inferred using TreeMix[102]. We ran the tree with the group 1 as the root group and made this the base tree topology. Then we ran TreeMix using introducing migration events from 1 to 6. To detect admixture, we computed *D*-statistics[103] based on ABBA and BABA SNP frequency differences. For a triplet of taxa P1, P2 and P3, and an outgroup O, that follows the phylogeny of (((P1, P2), P3), O), a *D* statistic significantly different from zero indicates P3 exchanged gene with P1 (*D* value 0) or P2 (*D* value >0). Then, the *f-branch* statistic calculated introgressions among the six groups by the software package Dsuite[104]. The f_d statistic[105] was used to calculate the fraction of introgression in G4 from G2 in 100-SNP windows, which signifies gene flow when $0 < f_d < 1$.

Nucleotide diversity (π) and fixation index (F_{ST}) were calculated by vcftools[106] and pairwise genetic distance was calculated by Arlequin (v.3.5.2.2)[107]. To estimate and compare the pattern of LD among different groups, the squared correlation coefficient (r^2) between pairwise SNPs was computed using the PopLDdecay (v.3.40)[108] software. Parameters in the program were --MaxDist 500 --MAF 0.05 --Miss 0.1. The average r^2 value was calculated for pairwise markers in a 500-kb window and averaged across the whole genome.

To construct subgenome trees, we selected 390 *B. juncea* accessions with genetic components greater than 60% in each group and 68 *B. rapa* and 11 *B. nigra* samples (Supplementary Table 24). We selected 14,264 and 10,629 synonymous SNPs for the A and B subgenomes, respectively, filtered with the following processes: depth for individual ≥ 4, missing rate ≤ 0.1, MAF > 5%. The maximum-likelihood phylogeny for each subgenome was constructed using IQ-TREE (v1.6.6)[100] based on the optimal models (TVM + F + ASC + R6) following the same pipeline implemented as that for the *B. juncea* phylogeny.

Pairwise identity-by-descent detection. To investigate genome-wide introgression between G4 and G2, we identified haplotypes in the G4 accessions that were identical by descent (IBD) with individuals from both the original source of diversification, the G5 leaf mustard, and the source of introgression, the G2 yellow-seeded mustard, following an approach described previously[109]. To estimate the frequency of shared haplotypes along individual chromosomes, each chromosome was divided into bins of 10 kb with a sliding window of 5 kb, and the number of recorded IBD tracts between G4 and the two groups (G2 and G5) was computed per bin. As the total number of pairwise comparisons differed between the groups, these numbers were normalized from 0 (no IBD detected) to 1 (IBD shared by all individuals within the group). The normalized IBD between G4 and the G2 (nIBD$_{G2}$) and the normalized IBD between G4 and the G5 (nIBD$_{G5}$) were then used to calculate the rIBD (nIBD$_{G2}$ − nIBD$_{G5}$). Finally, the putative introgression segments from the G2 to each of the G4 accessions were identified.

Estimation of divergence time and demographic history. With genome-scale characterization of the divergence of orthologous genes, we managed to date the divergence between *B. rapa* A genome and *B. juncea* A subgenome, between *B. nigra* B genome and *B. juncea* B subgenome, and between *Brassica* and *Arabidopsis*. The synonymous divergence (K_S) values for *A. thaliana*, *B. rapa*, *B. nigra*, and A and B subgenomes of *B. juncea* were calculated using the K_A/K_S Calculator (v2.0)[110]. The divergence time between species was calculated as $K_S/2\mu$, where μ is the mutation rate ($1.5 \times 10^{-8} \sim 9 \times 10^{-9}$ per synonymous site[111]).

SMC++ (v1.13)[112] was used to estimate the divergence time and historical *Ne* among different groups of *B. juncea*. For normalizing the population size, we selected seven different samples from each group. Generations were calculated by the upper and lower mutation rates of 1.5×10^{-8} and 9×10^{-9} per synonymous site for each generation[111], and the generation time was 1 year.

Organellar genome analysis. The CP genomes were assembled by NOVOPlasty[113] using genome re-sequencing data. After manually correcting the orientation of the two inverted repeats, the assembled CP genomes were annotated by GeSeq[114]. The InDel variants in CP genomes of *B. juncea* were identified through sequence alignment and confirmed by PCR (Extended Data Fig. 2a). The maximum-likelihood phylogeny of CP genomes was constructed based on high-quality variants (variants with >20% missing calls and MAF < 0.01) using RAxML (v8.0.17)[115] with the GTRGAMMAI model. A bootstrap of 1,000 repetitions was used to assess the reliability of the phylogeny reconstructed. The MT genomes were assembled by Celera Assembler[116] with default parameters using PacBio reads of ten *B. juncea* accessions. For the mitotype analysis, an InDel and a reported SNP locus[45] were identified by sequence alignment and confirmed by PCR (Extended Data Fig. 2b). Phylogenetic tree analysis of MT genomes was performed through IQ-TREE (v1.6.6)[100] using the best model (HKY + F) determined by the Bayesian information criterion with 1,000 bootstrap replicates.

Measurement and statistical analysis of agronomic traits. The 390 *B. juncea* accessions were grown in four locations: Guiyang (Guizhou, E106.72/N026.58, short-day, mild-winter), Xiangtan (Hunan, E112.90/N027.86, short-day, mild-winter), Kunming (Yunnan, E102.72/N025.04, long-day, subtropical) and Urumqi (Xinjiang, E087.60/N043.80, long-day, continental steppe with large diurnal temperature differences) in 2018 (designated G18, X18, K18 and U18, respectively). The field trials were conducted with two replications. The flowering time was recorded as days to flowering by 25% plants. Open pollinated seeds were harvested and dried. The mean weight of a thousand seeds from the three replications was used for further analysis. Statistical analyses of phenotypic data were performed with the R packages Hmisc (v4.1.1)[117] and Psych (v1.8.4)[118].

GWAS analysis. Only SNPs with MAF ≥ 0.05 and missing rate ≤ 0.1 in a population were used to carry out GWAS. This resulted in 4,423,439 SNPs that were used in GWAS for 390 *B. juncea* accessions. We performed GWAS using GEMMA (the genome-wide efficient mixed-model association) program[119] under the mixed-linear model. The top three PCs were used for population-structure correction. The genetic relationship between individuals was modeled as a random effect using the kinship (K) matrix. Significant *P*-value thresholds ($P < 10^{-6}$ and 10^{-5} for flowering time and TSW, respectively) were set to control the genome-wide type I error rate.

Selective-sweep analysis. The XP-CLR score were calculated using the XP-CLR[120] package with sliding windows of 10 kb that had a 5-kb overlap between adjacent windows. The top 5% regions were assigned to candidate selective regions, and genes in these regions were considered as candidate genes.

Transcriptome analysis. Total RNA was isolated from a sampled organ with two biological replicates at a specific developmental stage to investigate expression of the genes associated with formation of special organs for enlarged roots and tuber stems. As above, RNA-seq libraries were constructed and sequenced on an Illumina X Ten. The clean reads were mapped against the SY genome using TopHat (v2.0.12)[121] software. The number of reads mapped was counted using HTSeq (v0.6.1)[122] and then FPKM values were calculated for each gene. Transcripts of less than one per million mapped reads were ignored. Analysis of differential gene expression between two samples was performed using the DESeq R package (v1.18.0)[123]. Genes with an adjusted *P* value < 0.05 found by DESeq were assigned as differentially expressed. Procedures for the RT–qPCR analysis are provided in the Supplementary Note.

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

Data availability

The genome sequence and annotation data for *B. juncea* var. SY, the re-sequencing data for 480 *B. juncea* accessions and transcriptome data are accessible under NCBI BioProject no. PRJNA615316. For Functional annotation of the SY genome, the SwissProt (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz/), NR (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz/) and KEGG (release 53; https://www.genome.jp/kegg/brite.html) databases were used. Seeds of accessions used, phenotype data and sequences of the CP and MT genomes reported here are available from the corresponding authors upon request. Source data are provided with this paper.

References

76. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
77. Wu, Z. K. et al. Evaluation of linkage disequilibrium pattern and association study on seed oil content in *Brassica napus* using ddRAD sequencing. *PLoS ONE* **11**, e0146383 (2016).
78. Xie, W. B. et al. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc. Natl Acad. Sci. USA* **107**, 10578–10583 (2010).
79. Wu, Y., Bhat, P. R., Close, T. J. & Lonardi, S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.* **4**, e1000212 (2008).
80. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
81. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211 (2014).
82. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
83. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
84. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
85. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler Transform. *Bioinformatics* **25**, 1754–1760 (2009).
86. Akdemir, K. C. & Chin, L. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol.* **16**, 198 (2015).

87. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
88. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
89. Kent, W. J. BLAT—The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
90. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
91. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
92. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
93. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
94. Cheng, F. et al. Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. *Plant Cell* **25**, 1541–1554 (2013).
95. Schranz, M. E., Lysak, M. A. & Mitchell-Olds, T. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci.* **11**, 535–542 (2006).
96. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
97. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
98. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
99. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
100. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
101. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
102. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
103. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
104. Malinsky, M., Matschiner, M. & Svardal, H. Dsuite—Fast *D*-statistics and related admixture evidence from VCF files. *Mol. Ecol. Resour.* **21**, 584–595 (2021).
105. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).
106. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
107. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
108. Zhang, C. et al. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
109. Bosse, M. et al. Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nat. Commun.* **5**, 4392 (2014).
110. Zhang, Z. et al. KaKs_Calculator: calculating K_a and K_s through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259–263 (2006).
111. Qi, X. et al. Genomic inferences of domestication events are corroborated by written records in *Brassica rapa*. *Mol. Ecol.* **26**, 3373–3388 (2017).
112. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
113. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e8 (2016).
114. Tillich, M. et al. GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).
115. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
116. Gennady, D. et al. Consensus generation and variant detection by Celera Assembler. *Bioinformatics* **24**, 1035–1040 (2008).
117. Harrell, F. Hmisc: Harrell Miscellaneous. https://CRAN.R-project.org/package=Hmisc (2018).
118. Revelle, W. Psych: procedures for personality and psychological Research. https://CRAN.R-project.org/package=psych (2018).
119. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
120. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
121. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
122. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **5**, 621–628 (2008).
123. Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136–138 (2010).

Acknowledgements

This research was supported by National Natural Science Foundation of China (U20A2029), National Key Research and Developmental Program of China (2017YFD0101702) and Key Projects of Hunan Provincial Education Department (18A086). H.A. and J.C.P. were supported by the US Department of Energy (DOE HDTRA 1-16-1-0048) and the US National Science Foundation (NSF IOS 1339156). X. Wu at OCRI, CAAS, W. Qian at Xi'nan University, J. Zhang and D. Liu at Sichuan Academy of Agricultural Science and Y. Yuan at Tibet Academy of Agricultural Science kindly provided a part of the accessions used in this study. We thank Y. Liu at Johns Hopkins University and Q. Yang at Huazhong Agricultural University for constructive discussion and feedback on this manuscript.

Author contributions

Z.L. and W.H. conceived and designed the study. L.K., L.Q., M.Z., L.C., H.C. and H.L. performed data analysis. L. Yang, L. You, B.Y., X.L. and X.X. managed the fieldwork and prepared the samples. B.Y., M.Y., Y.G., D. Zhang, Y.R., D.J., D. Zhou, H.X. and Y.W. measured the agronomic traits. L.Q. and T.W. performed GWAS analysis. L.K., M.Z., L.C. and H.C. performed RNA-seq data analysis. H.A. and P.B. carried out the *f*-branch analysis. L.K., L.Q., M.Z. and Z.L. wrote the manuscript. S.-V.S., H.A., P.B., A.S.M., J.C.P. and R.J.S. revised the manuscript and gave suggestions and comments. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

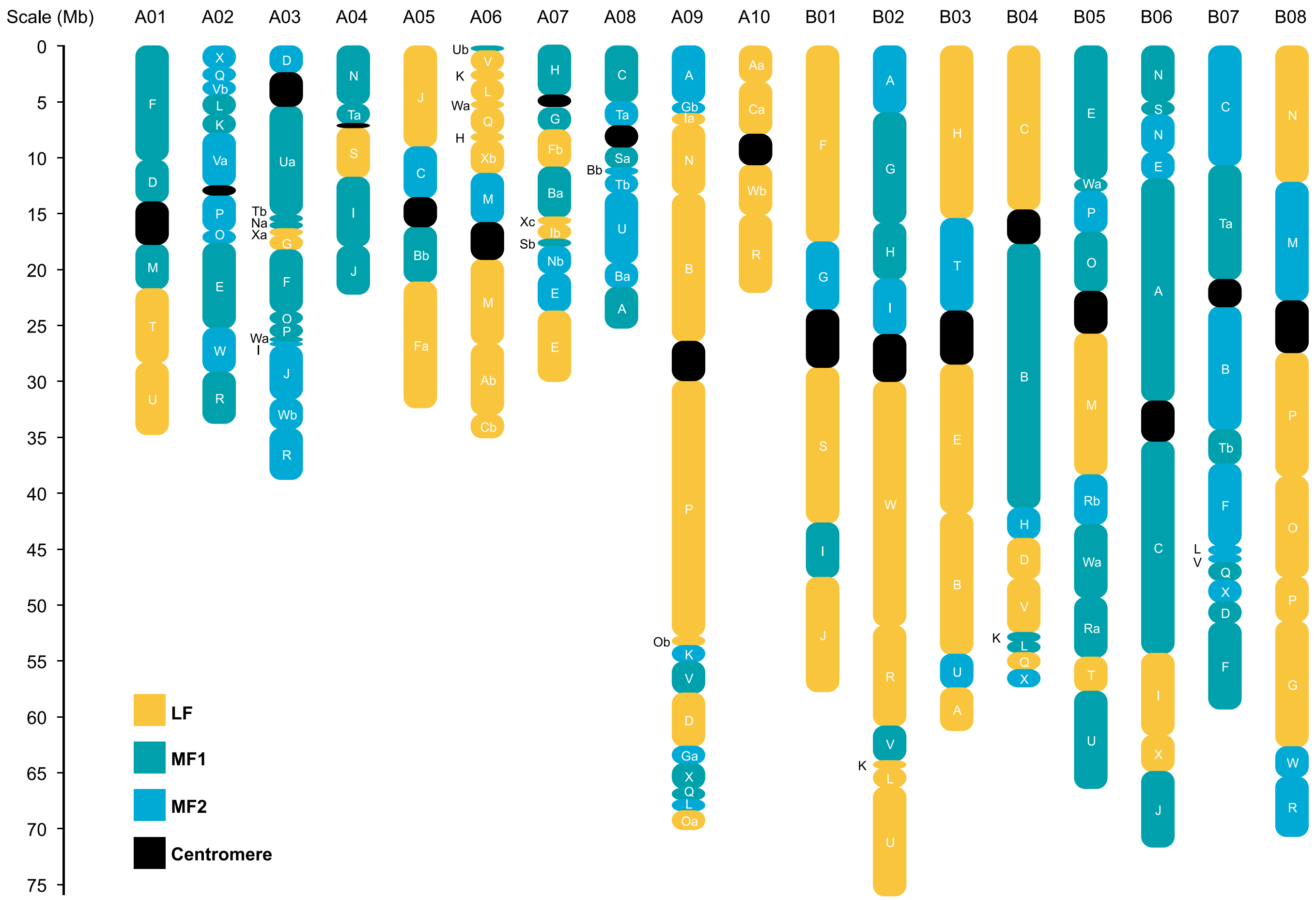
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00922-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00922-y.

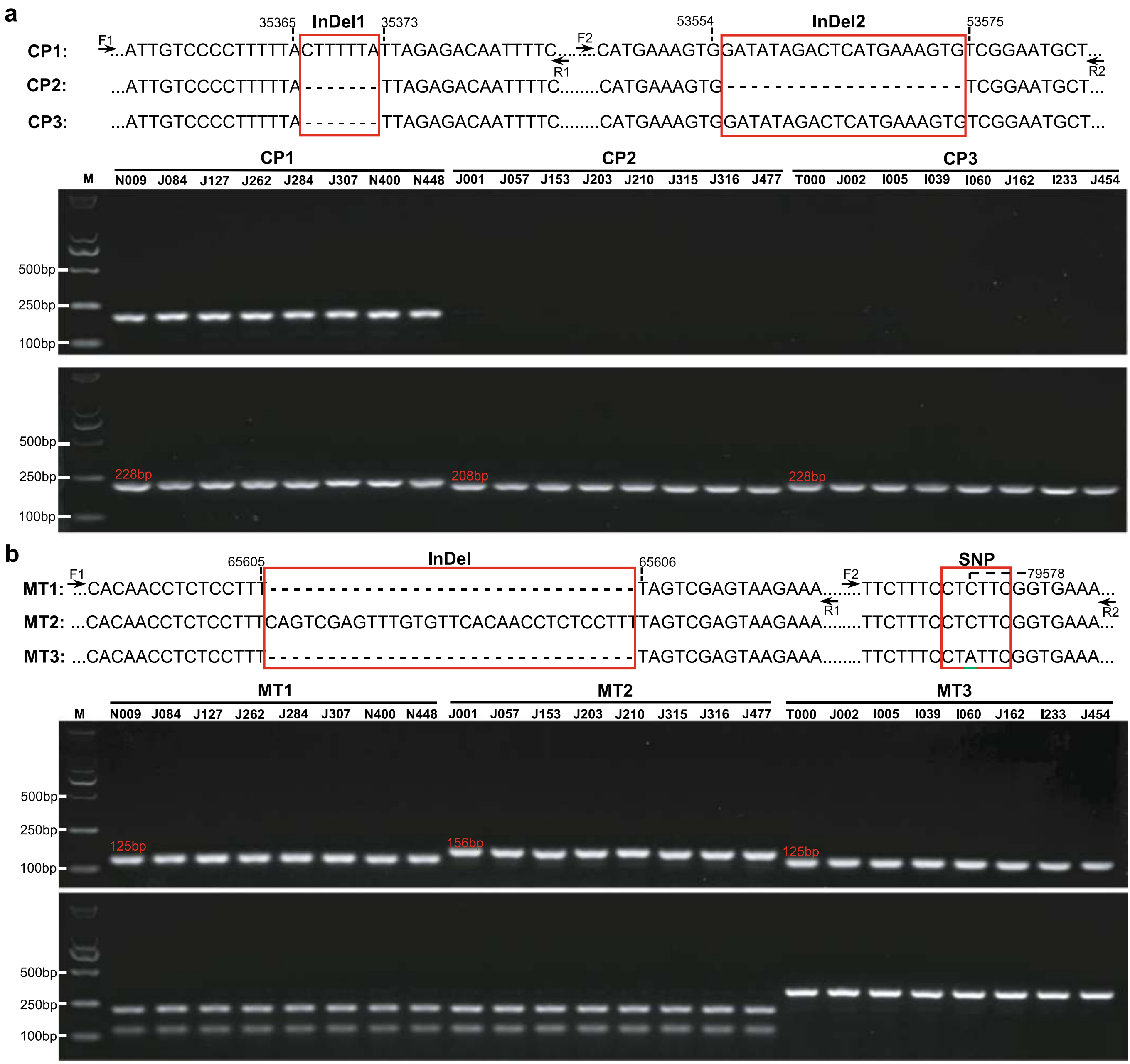
Correspondence and requests for materials should be addressed to Wei Hua or Zhongsong Liu.

Peer review information *Nature Genetics* thanks Caroline Belser and Xiaowu Wang for their contribution to the peer review of this work.

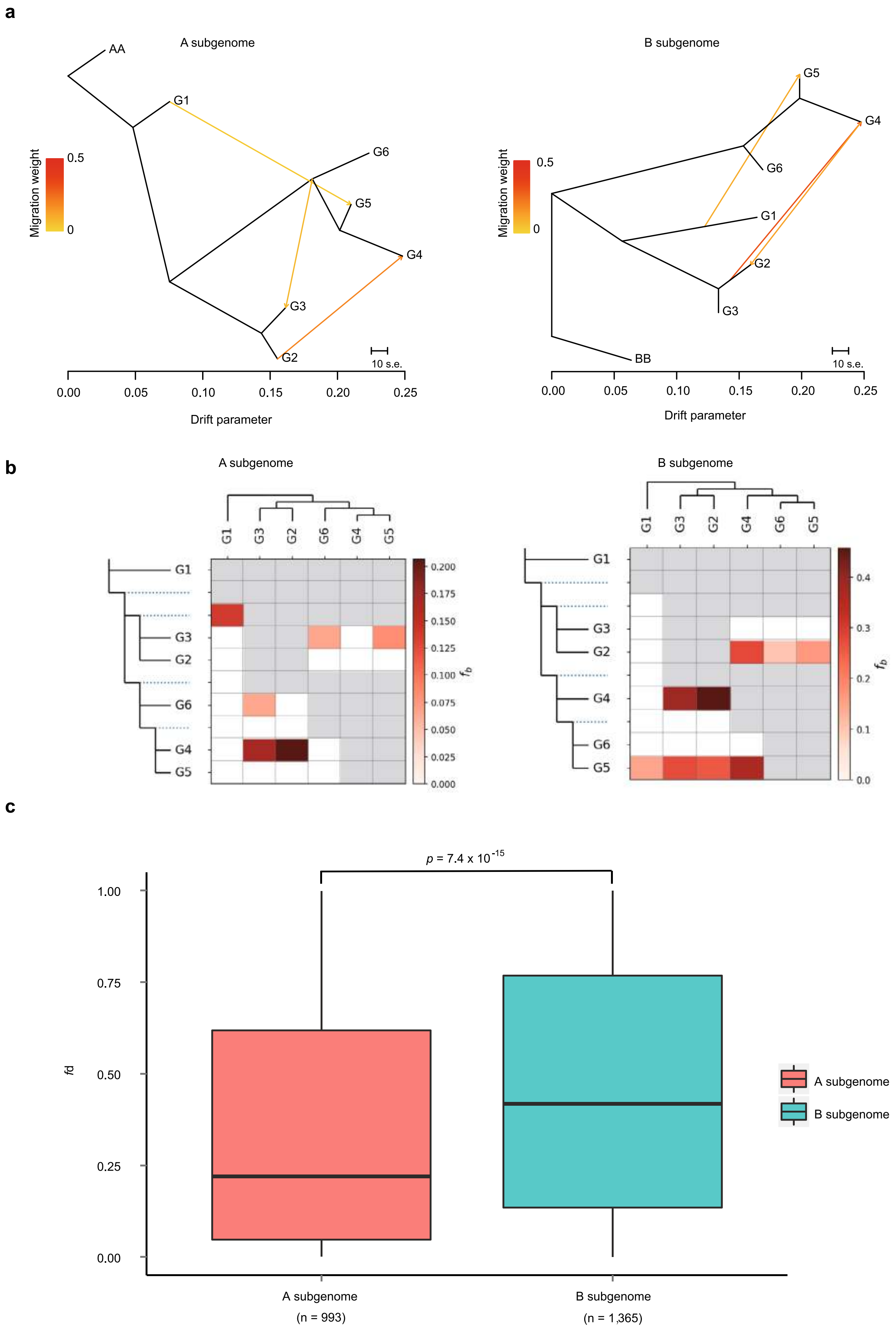
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Distribution of genomic blocks along the eighteen chromosomes of the *Brassica juncea* var. Sichuan Yellow genome. Genome blocks on eighteen chromosomes were assigned to the subgenomes LF (orange), MF1 (dark cyan), and MF2 (deep sky blue). The 24 conserved genomic blocks are defined and labelled from A to X (colored) based on the syntenic relationship of the *B. juncea* and *A. thaliana* genomes. The centromeres in the SY genome are shown as black. The orientation of chromosomes is according to international standards such that the centromeres are toward the top of the chromosome.

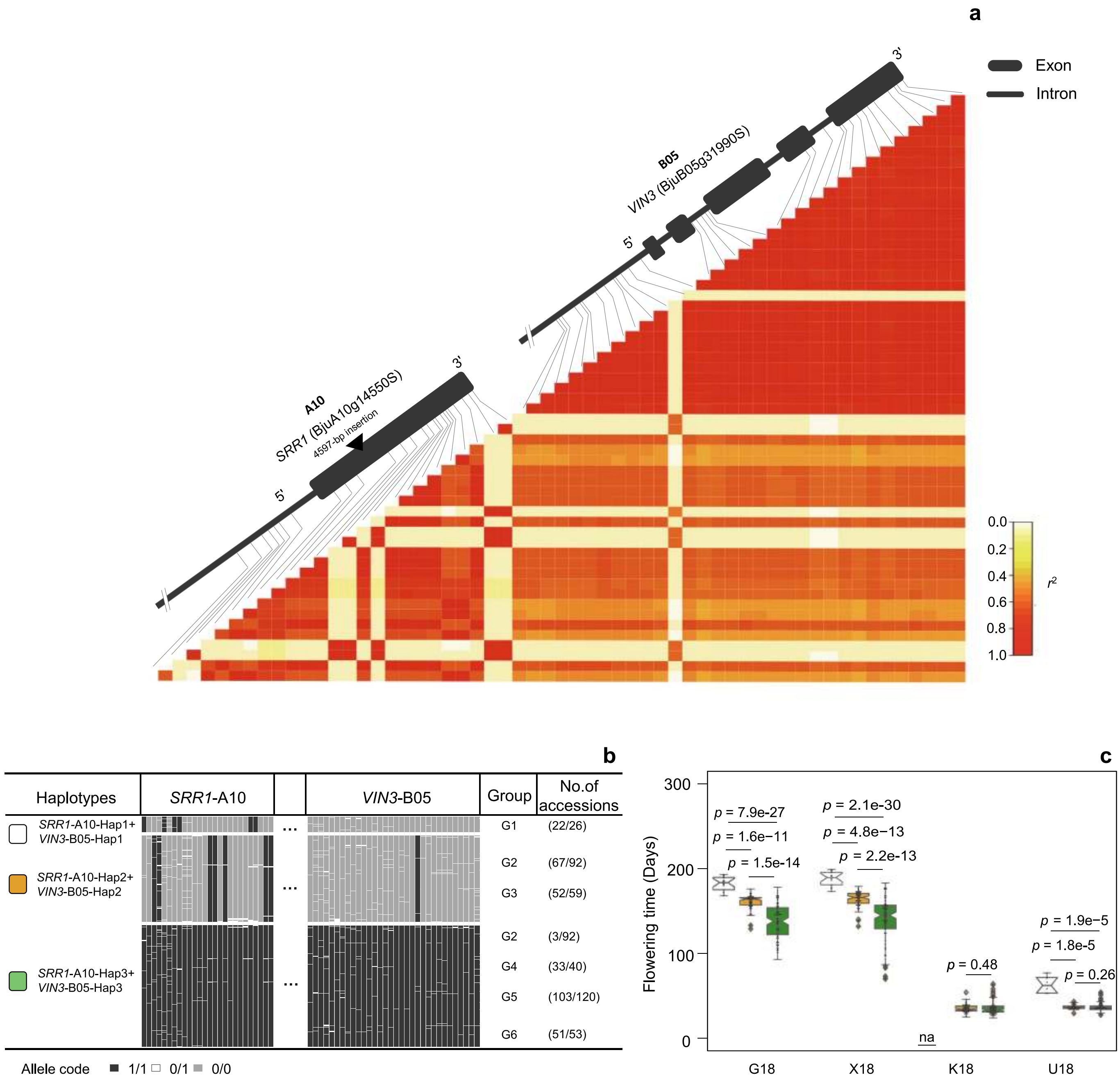

a
35365
InDel1
35373
53554
InDel2
53575
F1
F2
R1
R2
CP1: ...ATTGTCCCCTTTTTACTTTTTATTAGAGACAATTTTC........CATGAAAGTGGATATAGACTCATGAAAGTGTCGGAATGCT...
CP2: ...ATTGTCCCCTTTTTA-------TTAGAGACAATTTTC........CATGAAAGTG--------------------TCGGAATGCT...
CP3: ...ATTGTCCCCTTTTTA-------TTAGAGACAATTTTC........CATGAAAGTGGATATAGACTCATGAAAGTGTCGGAATGCT...
CP1
CP2
CP3
M
N009 J084 J127 J262 J284 J307 N400 N448
J001 J057 J153 J203 J210 J315 J316 J477
T000 J002 I005 I039 I060 J162 I233 J454
500bp
250bp
100bp
500bp
250bp
100bp
228bp
208bp
228bp
b
65605
InDel
65606
SNP
79578
F1
F2
R1
R2
MT1: ...CACAACCTCTCCTTT--------------------------------TAGTCGAGTAAGAAA........TTCTTTCCTCTTCGGTGAAA...
MT2: ...CACAACCTCTCCTTTCAGTCGAGTTTGTGTTCACAACCTCTCCTTTTAGTCGAGTAAGAAA........TTCTTTCCTCTTCGGTGAAA...
MT3: ...CACAACCTCTCCTTT--------------------------------TAGTCGAGTAAGAAA........TTCTTTCCTATTCGGTGAAA...
MT1
MT2
MT3
M
N009 J084 J127 J262 J284 J307 N400 N448
J001 J057 J153 J203 J210 J315 J316 J477
T000 J002 I005 I039 I060 J162 I233 J454
500bp
250bp
100bp
125bp
156bp
125bp
500bp
250bp
100bp

Extended Data Fig. 2 | Three types of *Brassica juncea* chloroplast and mitochondrial genomes. a, Three *B. juncea* chloroplast genome types were identified by sequence alignment. PCR validation of the two InDels in the chloroplast genomes of *B. juncea* accessions. **b**, Three *B. juncea* mitotypes were shown by sequence alignment. PCR validation of the InDel and the SNP in the mitochondrial genomes of *B. juncea* accessions. The amplified DNA was treated with the restriction enzyme *Ear*I. All the PCR experiments were repeated independently for three times with similar results. The primers used for PCR were listed in Supplementary table 42. Source data for the gels were provided as a Source Data file.

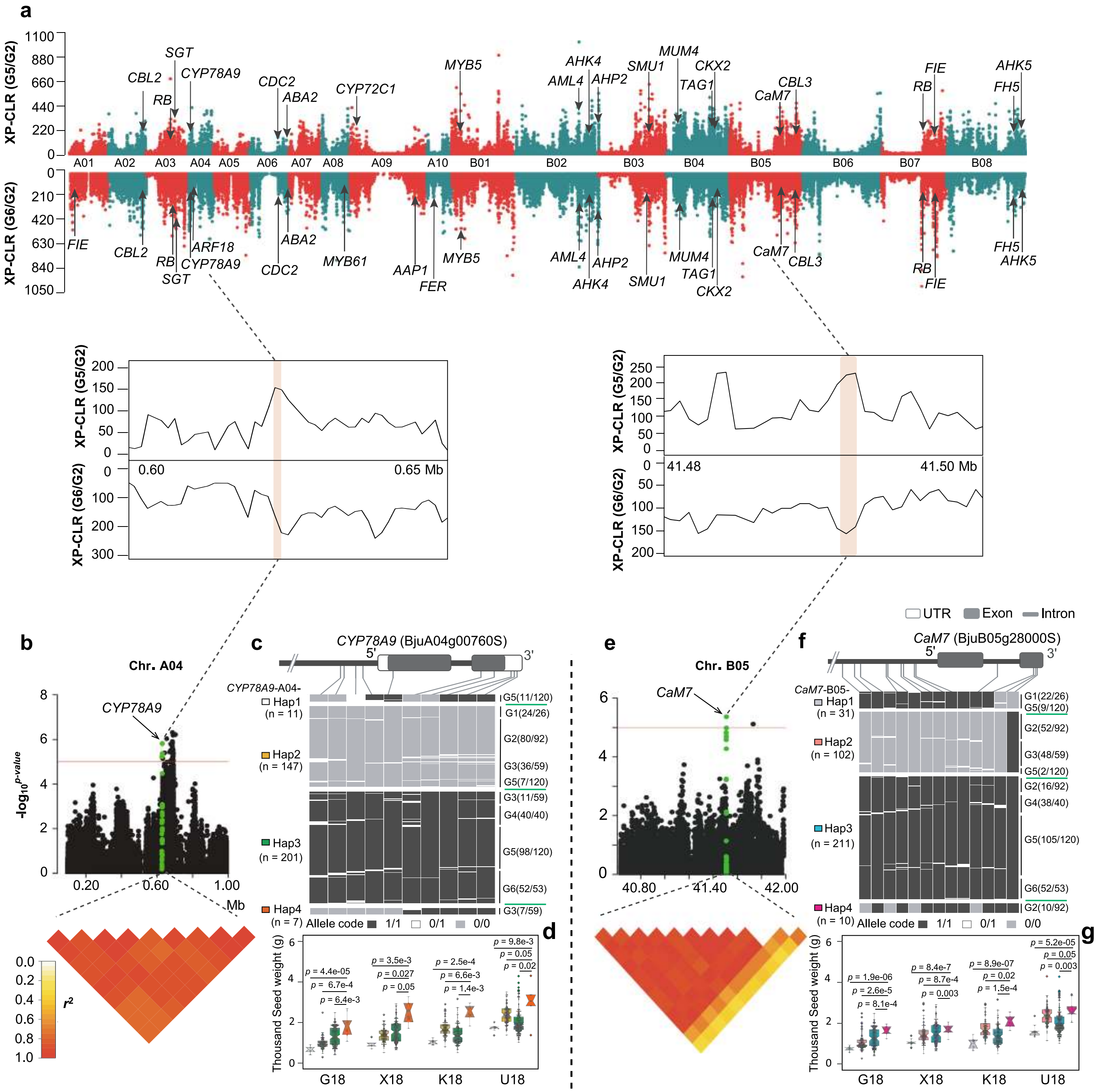


Extended Data Fig. 3 | See next page for caption.

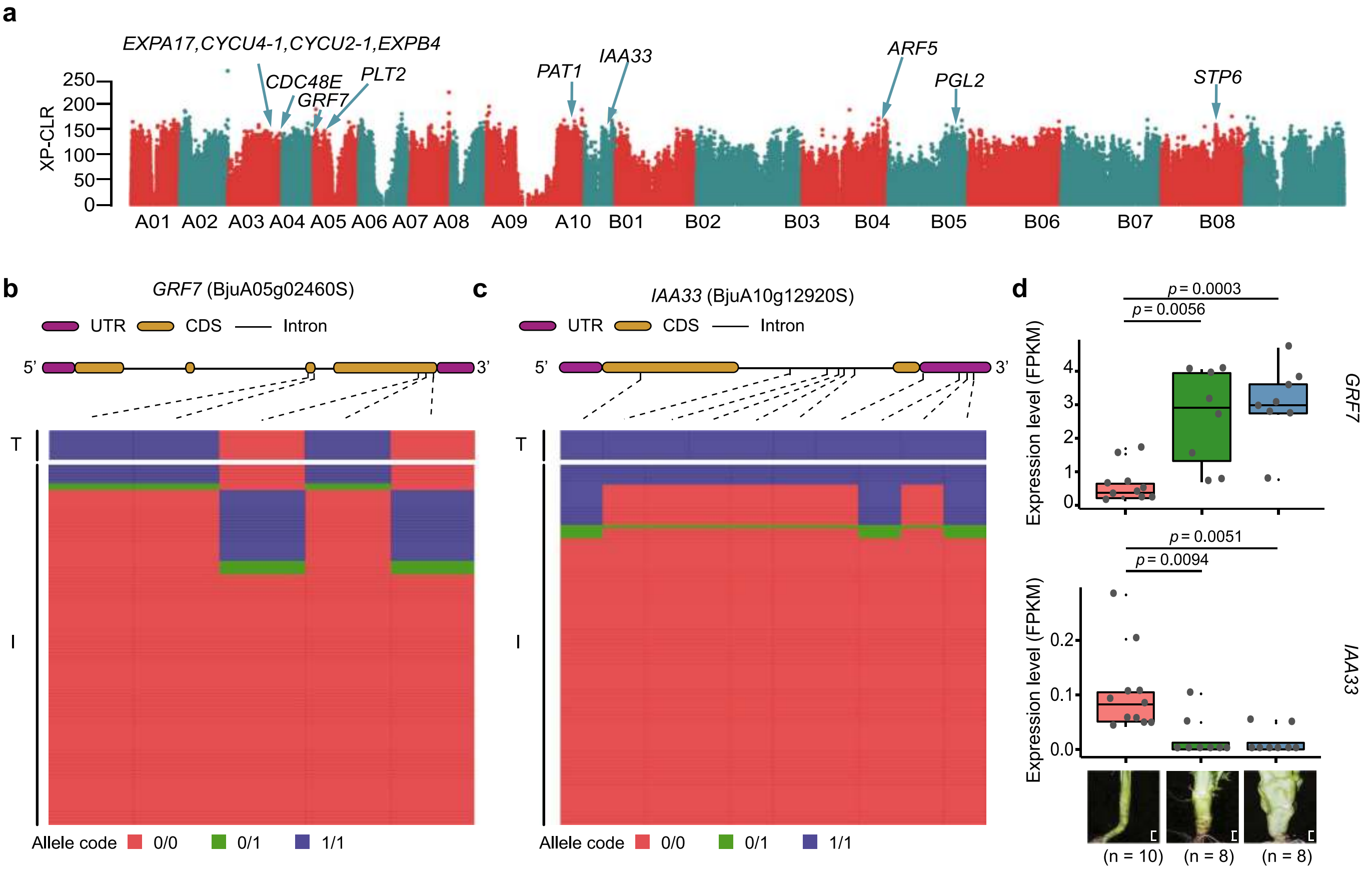
Extended Data Fig. 3 | Estimation of introgressions among the six groups of *Brassica juncea*. a, Treemix analysis. Migration arrows are colored according to their weight. Horizontal branch length is proportional to the amount of genetic drift that has occurred on the branch. Scale bar shows ten times the average standard error of the entries in the sample covariance matrix. **b**, *f*-branch values. **c**, *fd* values from G2 to G4. The center lines in box plots indicate the median values, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Whiskers extend to data no more than 1.5 times the interquartile range. *p*-value was calculated using two-sided *t*-test.



Extended Data Fig. 4 | Co-evolution analysis of the flowering time genes *SRR1* (BjuA10g14550S) and *VIN3* (BjuB05g31990S) in *Brassica juncea*. a, LD analysis between *SRR1* and *VIN3* genes. **b**, The combinations of both *SRR1* and *VIN3* haplotypes (*SRR1*-A10-Hap1 + *VIN3*-B05-Hap1, *SRR1*-A10-Hap2 + *VIN3*-B05-Hap2, and *SRR1*-A10-Hap3 + *VIN3*-B05-Hap3). **c**, Boxplot showing comparison between these three haplotypes corresponding to accessions across four environments. Box edges represent the 0.25 and 0.75 quantiles with the median values shown by bold lines. Whiskers extend to data no more than 1.5 times the interquartile range, and remaining data are indicated by dots. *p*-value was calculated with two-sided *t*-test. na, data missing (G1 group did not flower in Kunming).



Extended Data Fig. 5 | Genome-wide selective sweep scan and GWAS for seed weight in *Brassica juncea*. **a**, Genome-wide distribution of selective-sweep signals identified through comparisons between G5 or G6 with G2 using XP-CLR values (sliding window = 10 kb, step = 1 kb). The thousand seed weight candidate genes in the selection regions are labeled. **b** and **e**, Local Manhattan plot showing the 0.60 - 0.65 Mb and 41.48 - 41.50 Mb region on chromosomes A04 and B05, respectively. The green plots represent the position of these SNPs in *CYP78A9* (*BjuA04g00760S*) and *CaM7* (*BjuB05g28000S*). Three and one SNPs in *CYP78A9* and *CaM7* are significantly associated with thousand seed weight, respectively. The heatmaps span the SNP markers that show linkage disequilibrium (LD) with the most strongly associated SNPs. The grey dashed lines indicate the significance threshold (-$\log_{10}{}^{p}$ = 5.0). **c** and **f**, Comparison of conserved SNPs specific to six groups in *CYP78A9* and *CaM7* gene region, respectively. Two haplotypes with frequency greater than 0.01 were identified in *CYP78A9* and *CaM7* gene region, respectively. **d** and **g**, Comparison in thousand seed weight between accessions of three haplotypes in *CYP78A9* and *CaM7* gene region, respectively. Box edges represent the 0.25 quantile and 0.75 quantile with the median values shown by bold lines. Whiskers extend to data no more than 1.5 times the interquartile range, and remaining data are indicated by dots. *p*-value was calculated with two-sided *t*-test.



Extended Data Fig. 6 | Identification of candidate genes for stem tuber formation in stem mustard. **a**, Genome-wide distribution of selective sweeps in stem mustard for stem tuber formation. **b**, Haplotypes for the candidate gene *GRF7* (BjuA05g02460S) in stem mustard (T) and leaf mustard (I). **c**, Haplotypes for the candidate gene *IAA33* (BjuA10g12920S) in stem mustard (T) and leaf mustard (I). **d**, The expression levels of *GRF7* and *IAA33* in non-stem mustard, stem mustard (one week after the stem swelled, three weeks after the stem swelled) (from left to right) were estimated based on FPKM values. In box plots, the center lines indicate the median values and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Whiskers extend to data no more than 1.5 times the interquartile range, and remaining data are indicated by dots. *p*-value was calculated using two-sided *t*-test. Scale bars, 2 cm.

nature research

Corresponding author(s): Zhongsong Liu and Wei Hua

Last updated by author(s): May 24, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection

(1) DNA sequencing: PacBio reads were collected from single-molecule real-time (SMRT) cells on PacBio Sequel instruments; Hi-C data and pair-end reads were collected from the Illumina HiSeq X Ten platform and Illumina HiSeq2000 platform; BioNano data were collected from BioNano Irys System using genomic DNA digested by Nt.BspQI.

(2) RNA-seq data were generated using Illumina HiSeq X Ten and its system software.

Data analysis

(1) Genome assembly: We used Jellyfish (v.2.2.9) to estimated the genome size. We used FALCON (falcon-kit==0.7), Sapace-longread (v.1.1), PBJelly (v.1.9.1), Chromonomer (v.1.07), QUIVER (v2.0.0), Plion (V1.22), BioNano Solve (V3.1), BWA (v.0.7.8), Juicebox (V1.9.8) for genome assembly. NOVOPlasty (V3.7) and Celera Assembler (V7.0) were used for chloroplast assembly and mitochondrial assembly, respectively.

(2) Genome annotation: We used Repeatmasker (v4.0.5), LTR_FINDER (V 1.0.5), RepeatScout (V1.0.5), RepeatModeler (v1.0.11), RepeatProteinMask (version 3.3.0,) Tandem Repeats Finder (TRF, version 4.07b), Augustus (version 2.5.5), Genscan (version 1.0), GlimmerHMM (version 3.0.1), Geneid (V1.4), SNAP (2013.11.29), Tophat (version 2.0.8), Cufflinks (version 2.1.1), PASA (v2.3.3), EvidenceModeler (V 1.1.1), tRNAscan-SE (V2.0), INFERNAL (v 1.1.3), WUblast (version 2.0), GeneWise (V2.4.1), InterProScan (v 5.32-71.0), BLAST (V2.2.28) for genome annotation.

(3) Assessment of genome completeness: We used BUSCO (v. 2) and CEGMA (v. 2.5) to evaluate the genetic integrity of the genome. LTR_retriever (v 2.8.7) was used to evaluate the integrity of the genome repeat sequence. LASTZ (V 1.02.00) software was used for BAC and BES evaluation. Heatmaps were generated using the JUICER-pre command, and visualized using Juicebox (V 1.9.8). BLASTP (v 2.2.6) was used to align the protein sequence of SY with previously reported Brassica genome, and MCScanX was used to build genome synteny between SY and other five Brassica subgenomes.

(4) Population structure analysis: We used BWA (v. 0.7.8), SAMtools (v.0.1.19), GATK (sentieon-genomics-201711), ANNOVARtool (version 2013-05-20) for SNP calling and annotation. ADMIXTURE (v.1.23) was used for population genetic structure analysis. IQ-TREE v1.6.6 was used

to construct phylogenetic tree. GCTA (V 1.26.0) was used for principal component analysis. The population relatedness and migration events were inferred using TreeMix (V 1.13). Arlequin (v.3.5.2.2) was used to calculate Nucleotide diversity (π) and fixation index (Fst). PopLDdecay (v.3.40) was used to calculate the squared correlation coefficient (r2) between pairwise SNPs. Ka/Ks Calculator (v. 2.0) and SMC++ (V1.13) were used for estimation of divergent time.

(5) GWAS and Selective sweep analysis: We performed GWAS using GEMMA (V 0.98.1) program under the mixed-linear model; XP-CLR (V 1.0) was used for selective sweep analysis.

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

Data

Policy information about availability of data

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

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

The genome sequence and annotation data for B. juncea var. Sichuan Yellow, the re-sequencing data for 480 B. juncea accessions, and transcriptome data are accessible under NCBI BioProject numbers PRJNA615316. Functional annotation of the SY genome used the SwissProt (ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz), NR (ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) and KEGG (release 53, https://www.genome.jp/kegg/brite.html) databases. Seed of accessions used, phenotype data, and sequences of the chloroplast and mitochondrial genomes reported here are available from the corresponding authors upon request. Source data are provided with this paper.

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

A panel of 480 mustard accessions were collected from 38 countries. It represents the four subspecies and samples all regions where mustard is spread. The 480 samples were clustered into six groups based on population structure. To better clarify the relationships of Brassica juncea accessions, 390 accessions with the genetic components of larger than 0.6 were retained for the further analysis.
Sample size for linkage map analysis was determined based on 172 recombinant inbred lines individuals.

Data exclusions

For population genomic analyses, we excluded 90 Brassica juncea accessions with the genetic components of less than 0.6 based on the population structure analysis.

Replication

Phenotyping was performed in four locations with two replicates at each location.
For RNA-Seq experiment, we used two biological replicates.
The PCR experiments were repeated independently for three times with similar results.
For all experiments, at least two replication were repeated and succeed.

Randomization

The 390 samples were clustered into six groups based on the result of population structure, PCA and phylogenetic tree.

Blinding

The investigators were blinded to the group allocations.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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