BMC Biology

RESEARCH ARTICLE Open Access

Genome sequencing sheds light on the contribution of structural variants to *Brassica oleracea* diversification

Ning Guo[1†], Shenyun Wang[2†], Lei Gao[3,4†], Yongming Liu[5†], Xin Wang[3], Enhui Lai[4,6], Mengmeng Duan[1], Guixiang Wang[1], Jingjing Li[5], Meng Yang[5], Mei Zong[1], Shuo Han[1], Yanzheng Pei[5], Theo Borm[7], Honghe Sun[1], Liming Miao[1], Di Liu[1], Fangwei Yu[2], Wei Zhang[2], Heliang Ji[8], Chaohui Zhu[9], Yong Xu[1], Guusje Bonnema[7*], Jianbin Li[2*], Zhangjun Fei[3,10*] and Fan Liu[1*]

Abstract

Background: *Brassica oleracea* includes several morphologically diverse, economically important vegetable crops, such as the cauliflower and cabbage. However, genetic variants, especially large structural variants (SVs), that underlie the extreme morphological diversity of *B. oleracea* remain largely unexplored.

Results: Here we present high-quality chromosome-scale genome assemblies for two *B. oleracea* morphotypes, cauliflower and cabbage. Direct comparison of these two assemblies identifies ~ 120 K high-confidence SVs. Population analysis of 271 *B. oleracea* accessions using these SVs clearly separates different morphotypes, suggesting the association of SVs with *B. oleracea* intraspecific divergence. Genes affected by SVs selected between cauliflower and cabbage are enriched with functions related to response to stress and stimulus and meristem and flower development. Furthermore, genes affected by selected SVs and involved in the switch from vegetative to generative growth that defines curd initiation, inflorescence meristem proliferation for curd formation, maintenance and enlargement, are identified, providing insights into the regulatory network of curd development.

Conclusions: This study reveals the important roles of SVs in diversification of different morphotypes of *B. oleracea*, and the newly assembled genomes and the SVs provide rich resources for future research and breeding.

Keywords: *Brassica oleracea*, Cauliflower, Cabbage, Structural variants, Curd development

* Correspondence: guusje.bonnema@wur.nl; jbli@jaas.ac.cn; zf25@cornell.edu; liufan@nercv.org
†Ning Guo, Shenyun Wang, Lei Gao and Yongming Liu contributed equally to this work.
[7]Plant Breeding, Wageningen University and Research, 6708 PB Wageningen, The Netherlands
[2]Jiangsu Key Laboratory for Horticultural Crop Genetic Improvement, Institute of Vegetable Crops, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China
[3]Boyce Thompson Institute, Ithaca, NY 14853, USA
[1]National Engineering Research Center for Vegetables, Beijing Academy of Agriculture and Forestry Sciences, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (North China), Beijing Key Laboratory of Vegetable Germplasm Improvement, Beijing 100097, China
Full list of author information is available at the end of the article

Background

Brassica oleracea includes several diverse dominant vegetable crops with a worldwide total production of nearly 100 million tons in 2018 (http://www.fao.org/faostat). The extreme diversity of this species is unique with morphotypes selected for the enlargement of distinct organs that represent the harvested product, e.g., inflorescences for cauliflower (*B. oleracea* var. *botrytis*) and broccoli (*B. oleracea* var. *italica*), leafy heads (terminal leaf bud) for cabbage (*B. oleracea* var. *capitata*), lateral leaf buds for brussels sprouts (*B. oleracea* var. *gemmifera*), leaves for kale (*B. oleracea* var. *alboglabra*), and tuberous stems for kohlrabi (*B. oleracea* var. *gongylodes*) [1, 2]. Reference genome sequences have been generated for different morphotypes of *B. oleracea* during the past several years, including kale [3], cabbage [4–6], cauliflower [7], and broccoli [8]. These genome sequences have greatly facilitated genetic variant analyses for a better understanding of the genetic diversity, population structure, and evolution and domestication of *B. oleracea.*

Structural variants (SVs) including insertions, deletions, duplications, and translocations are abundant throughout plant genomes and are more likely to cause phenotype changes than single nucleotide polymorphisms (SNPs) [9, 10]. Numerous SVs have been identified as causal genetic variants for important agronomic traits of various crops, such as the 4.7-kb insertion into the third exon of the *Or* gene leading to the orange curd in cauliflower [11], the 3.7-kb insertion in the upstream region of *BnaA9.CYP78A9* leading to the long siliques and large seeds of *Brassica napus* [12], and the 621-bp insertion in the promoter region of *BnaFLC.A10* contributing to the adaptation of rapeseed to winter cultivation environments [13]. Previous genome-wide variant analyses in *B. oleracea* focused on SNPs and small indels [14, 15] with genomic SVs largely ignored, mainly due to the limitations of using short sequencing reads in genetic variant identification. SV calling through mapping short sequencing reads to a reference genome is subject to high levels of both false negatives and false positives [16], especially for highly repetitive plant genomes such as those of *B. oleracea*. Therefore, to date, population dynamics of SVs in different *B. oleracea* morphotypes remain largely unexplored.

Recently, approaches by direct comparison of high-quality chromosome-level genome assemblies and/or mapping long reads generated using PacBio or Nanopore sequencing technologies to reference genomes have proven to be highly accurate for SV detection in large and complex plant genomes [17, 18]. In this study, we generated high-quality chromosome-scale genome assemblies for both cauliflower and cabbage using PacBio long reads and the high-throughput chromosome conformation capture (Hi-C) technology. Through direct genome comparison combined with long read mapping, we identified a total of 119,156 high-confidence SVs between these two genomes. We further generated and collected genome resequencing data of 271 *B. oleracea* accessions belonging to different morphotypes, and these data were used to genotype the 119,156 high-confidence SVs in these accessions. Allele frequencies of these SVs were investigated in different *B. oleracea* morphotypes, and mainly compared between cauliflower and cabbage populations. Together with gene expression analysis, we demonstrated the contribution of SVs to the regulation of cauliflower curd formation.

Results

De novo assembly of cauliflower and cabbage genomes

The inbred lines cauliflower Korso_1401 (hereafter Korso) and pointed cabbage OX-heart_923 (hereafter OX-heart) were selected for genome sequencing (Additional file 1: Figure S1). Approximately 70.0 Gb PacBio sequences were generated for each accession, covering about 120× of the Korso and OX-heart genomes, which had estimated sizes of 566.9 Mb and 587.7 Mb, respectively (Additional file 1: Figure S2). These PacBio reads were de novo assembled into contigs and errors in the assembled contigs were corrected using both PacBio long reads and Illumina short reads (~ 100 Gb for each accession). In addition, a genome map was assembled from 242.2 Gb cleaned BioNano optical map data for Korso and used to connect the assembled contigs. Furthermore, 285.8 and 453.0 million cleaned Hi-C read pairs, among which 58.6 and 140.0 million were valid, were used for pseudochromosome construction for Korso and OX-heart, respectively. The final genome assemblies of Korso and OX-heart comprised 615 and 973 contigs, respectively, with cumulative lengths of 549.7 Mb and 565.4 Mb, and N50 sizes of 4.97 Mb and 3.10 Mb (Additional file 2). A total of 544.4 Mb and 539.1 Mb, accounting for 99.0% and 95.3% of the Korso and OX-heart assemblies, respectively, were clustered into nine pseudomolecules. The Hi-C heatmaps (Additional file 1: Figure S3) and the good synteny between Korso and OX-heart assemblies and the broccoli HDEM assembly [8] (Additional file 1: Figure S4) supported their chromosome-scale structures.

Around 99.8% of the Illumina genomic reads could be mapped back to the Korso and OX-heart assemblies, with 99.6% of the assemblies covered by at least 5 reads. Based on the alignments, the estimated base error rates of the Korso and OX-heart assemblies were 1.23×10^{-5} and 5.6×10^{-5}, respectively (Additional file 3). BUSCO analysis [19] showed that 97.2% and 96.5% core conserved plant genes were completely assembled in Korso and OX-heart. In addition, up to 98.0% of the RNA-Seq reads could be mapped to the assemblies (Additional file 4). Together, these results demonstrated the high quality of the Korso and OX-heart assemblies.

Genome annotation and comparative genomics

Approximately 60.7% and 62.0% of the Korso and OX-heart assemblies were annotated as repetitive elements, respectively, similar to that (60.5%) in the *B. oleracea* var. *italica* HDEM genome assembly (Additional file 5). Full-length long terminal repeat retrotransposons (LTR-RTs) were then extracted from the Korso, OX-heart, and *B. rapa* (V3.0) [20] genomes (Additional file 6). Insertion time estimation of these intact LTR-RTs unraveled two LTR-RT bursts that occurred in Korso and OX-heart, around 0.2 and 1.5 million years ago (mya), respectively (Additional file 1: Figure S5). In contrast, in *B. rapa*, most of the LRT-RT formed recently, with more than 30% of the identified intact LTR-RTs younger than 0.2 mya, compared to 16.3% and 15.9% in Korso and OX-heart, respectively.

The high-quality Korso and OX-heart assemblies allowed us to precisely identify the centromere locations. The determined positions of centromeres on each chromosome in both genomes (Additional file 1: Figure S6) were consistent with the previously determined centromere locations using fluorescent in situ hybridization (FISH) analysis [21]. As expected, repetitive elements were enriched in the centromere regions. Different repeat families displayed clearly different patterns on the chromosomes, e.g., *Copia*-type LTRs were mainly in centromeres, while *Gypsy*-type LTRs were in pericentromeric regions (Fig. 1a).

A total of 60,640 and 62,232 protein-coding genes were predicted from Korso and OX-heart genomes, respectively, using an integrated strategy combining ab initio, transcript-based and homology-based predictions. Among these predicted genes, 70.9% and 76.4% were supported by transcriptome evidence, and 91.0% and 90.0% had homologs in other plant species.

Synteny analysis of Korso, OX-heart, *B. rapa* and *A. thaliana* genomes confirmed the whole genome triplication (WGT) and subsequent sub-genome divergence in *Brassica* species [4, 14, 22] (Additional file 1: Figures S7 and S8). Based on these syntenic relationships, we identified the triplicated regions within Korso and OX-heart genomes and divided them into three subgenomes based on their retained gene densities (Fig. 1). As previously reported in *B. rapa* [22] and *B. oleracea* [4], the three subgenomes of Korso and OX-heart, LF (the least fractionated), MF1 (the medium fractionated), and MF2 (the most fractionated), showed the same biased retention pattern of duplicated genes during diploidization [23] (Additional file 1: Figure S8a,b). Duplicated gene copies left in different subgenomes displayed diverged gene expression patterns, with the copies located in LF generally having higher expression levels than those in MF1 and MF2 (Additional file 1: Figure S8c,d).

We compared protein sequences of predicted genes from four *B. oleracea* accessions (cauliflower Korso, pointed cabbage OX-heart, broccoli HDEM and kale like

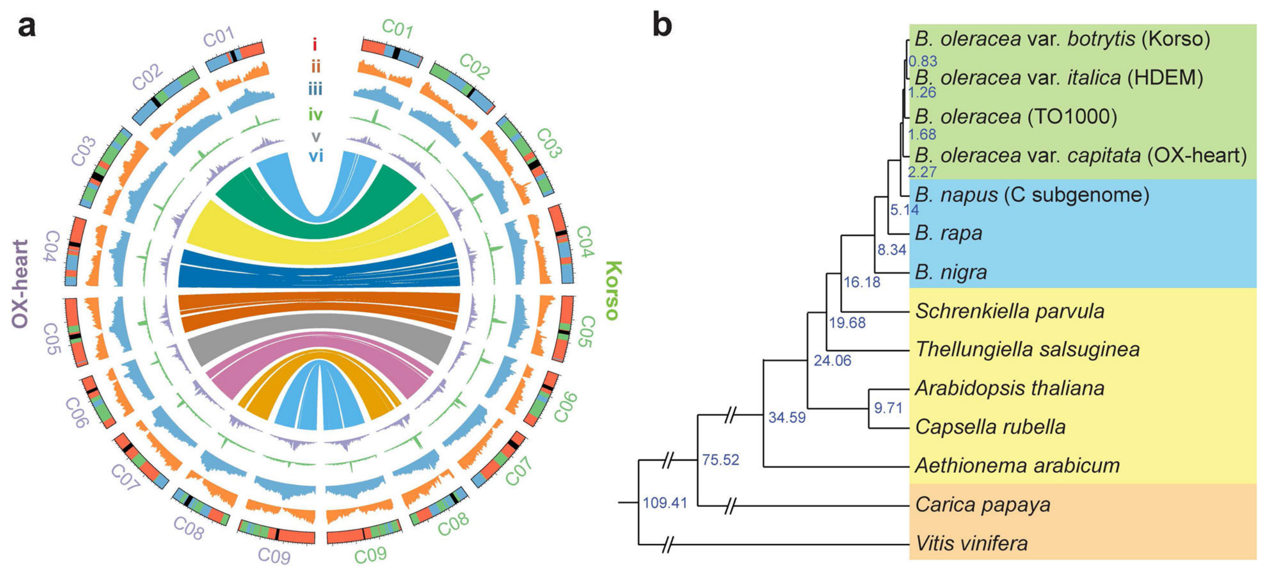

Fig. 1. Genomes of cauliflower Korso and point cabbage OX-heart. **a** Features of the Korso and OX-heart genomes. (i) Ideogram of the chromosomes. Red, green, blue, and black colors indicate the LF, MF1, and MF2 subgenomes and centromere regions, respectively. (ii) Gene density. (iii) Repeat density. (iv) *Copia*-type LTR density. (v) *Gypsy*-type LTR density. (vi) Synteny blocks between Korso and OX-heart genomes. **b** Phylogenetic tree of 14 plant species/varieties and their estimated divergence times (million years ago) based on 1638 single-copy orthologous genes

rapid cycling TO1000), three other *Brassica* species, *B. rapa*, *B. nigra*, and the C subgenome of *B. napus*, five other Brassiaceae species (*Aethionema arabicum*, *Arabidopsis thaliana*, *Capsella rubella*, *Thellungiella salsuginea*, and *Schrenkiella parvula*), and two outgroups (grape and papaya). A phylogenetic tree was constructed using 1638 single-copy orthologous genes, which indicated that cabbage and the common ancestor of cauliflower and broccoli diverged about 1.68 mya, the extant *B. oleracea* and the donor of *B. napus* C subgenome diverged about 2.27 mya, and *Brassica* diverged from other Brassiaceae species about 16.18 mya (Fig. 1b).

SVs between genomes of Korso and OX-heart

By taking advantage of the high-quality genome assemblies of Korso and OX-heart, we were able to identify high-confidence SVs through direct genome comparison combined with PacBio long read mapping. The Korso and OX-heart assemblies displayed very high collinearity indicating the balanced rearrangements (inversions and translocations) were not profound between them (Additional file 1: Figure S4). Therefore, in this study, we focused on the unbalanced SVs, mainly indels. A total of 119,156 SVs were identified between genomes of Korso and OX-heart, with sizes ranging from 10 bp to 667 kb with a clear bias to the relatively short ones, and these SVs were more likely to overlap with different types of repetitive sequences except the satellite sequences (Additional file 7 and Additional file 1: Figure S9).

SVs in gene bodies and promoter regions can affect the function or expression of the corresponding genes. The SV regions accounted for 14.5% and 15.0% of the total genome sizes of Korso and OX-heart, 10.0% and 11.3% of the gene regions, and 5.9% and 6.6% of the coding sequences, respectively, suggesting a functional constraint against the occurrence of SVs in genes, especially in coding regions, while no obvious restriction of SVs in promoter regions was detected (Additional file 8). More than half of the annotated genes in Korso (58.5%) and OX-heart (58.6%) were affected by at least one SV in their gene bodies or promoter regions, with a functional enrichment in diverse biological processes, such as cellular component organization, response to stress and stimulus, signal transduction, cell differentiation, embryo development, gene expression and epigenetic regulation, and flower and meristem development (Additional file 1: Figure S10). We detected several previously described SVs in *B. oleracea*, including the two indels in *BoFLC3* related to subtropical adaptation of broccoli [24], and the two indels in *BoFRIa* related to winter annual or biennial habit of cauliflower and cabbage [25].

Population dynamics of SVs in different *B. oleracea* morphotypes

Cabbage and cauliflower represent two extreme morphotypes of the *B. oleracea* species, and identifying genomic variations underlying the formation of their unique phenotypes (e.g., leafy head and curd) would provide novel insights into the molecular regulation of these important traits as well as important information for facilitating breeding. The high-quality SVs that we identified between Korso and OX-heart provided a valuable reference to investigate their dynamics in different morphologically diverged *B. oleracea* accessions. For this purpose, we performed genome resequencing of 163 *B. oleracea* accessions, including 89 cauliflower, 65 cabbage, and 9 broccoli accessions. We also collected resequencing data of an additional 108 *B. oleracea* accessions reported in Cheng et al. [14], including 15 cauliflower, 39 cabbage, 24 broccoli, 18 kohlrabi, four Chinese kale, two curly kale, two kale, two brussels sprout, and two wild *B. oleracea* accessions (Additional file 9). Among these 271 accessions, 211 were sequenced to a depth of more than 10×. The 119,156 high-quality reference SVs were genotyped in these 271 accessions based on the alignments of genome sequencing reads to the Korso and OX-heart genomes. To assess the accuracy of our SV genotyping, we genotyped the reference SVs in Korso and OX-heart by mapping their Illumina short reads to both these genomes, respectively. More than 86% of SVs could be genotyped, while only 0.1% were falsely genotyped (Additional file 10), suggesting high sensitivity and accuracy of our genotyping. The SV genotyping rate in each accession ranged from 41.3% to 80.2%, with 187 (69.0%) and 254 (93.7%) accessions having a genotyping rate greater than 70% and 60%, respectively (Additional file 1: Figure S11 and Additional file 9). In total, 89,882 (75.4%) SVs were successfully genotyped in more than 50% of the 271 accessions.

SV allele frequency variations among different groups of *B. oleracea* are mainly a result of domestication for different desirable traits and adaptation to different environments. As expected, SV loci with the homozygous Korso alleles were prevalent in cauliflower accessions, taking up an average of 82.3% of the genotyped SVs in each accession, whereas in cabbage accessions, the homozygous OX-heart alleles were prevalent, with an average frequency of 61.7% (Fig. 2a and Additional file 9). Phylogenetic and principal component analyses (PCA) using the SVs clearly divided cauliflower, cabbage, broccoli, and kohlrabi accessions into different groups (Fig. 2b, c), which were concordant with the patterns revealed by SNP data in our analysis based on the same 271 accessions (Additional file 1: Figure S12) and the previous report based on 119 accessions [14], further supporting that our SV detection and genotyping were highly reliable.

To identify SVs potentially related to the specific traits of cauliflower or cabbage, we extracted a total of 49,904 SVs with significantly different allele frequencies

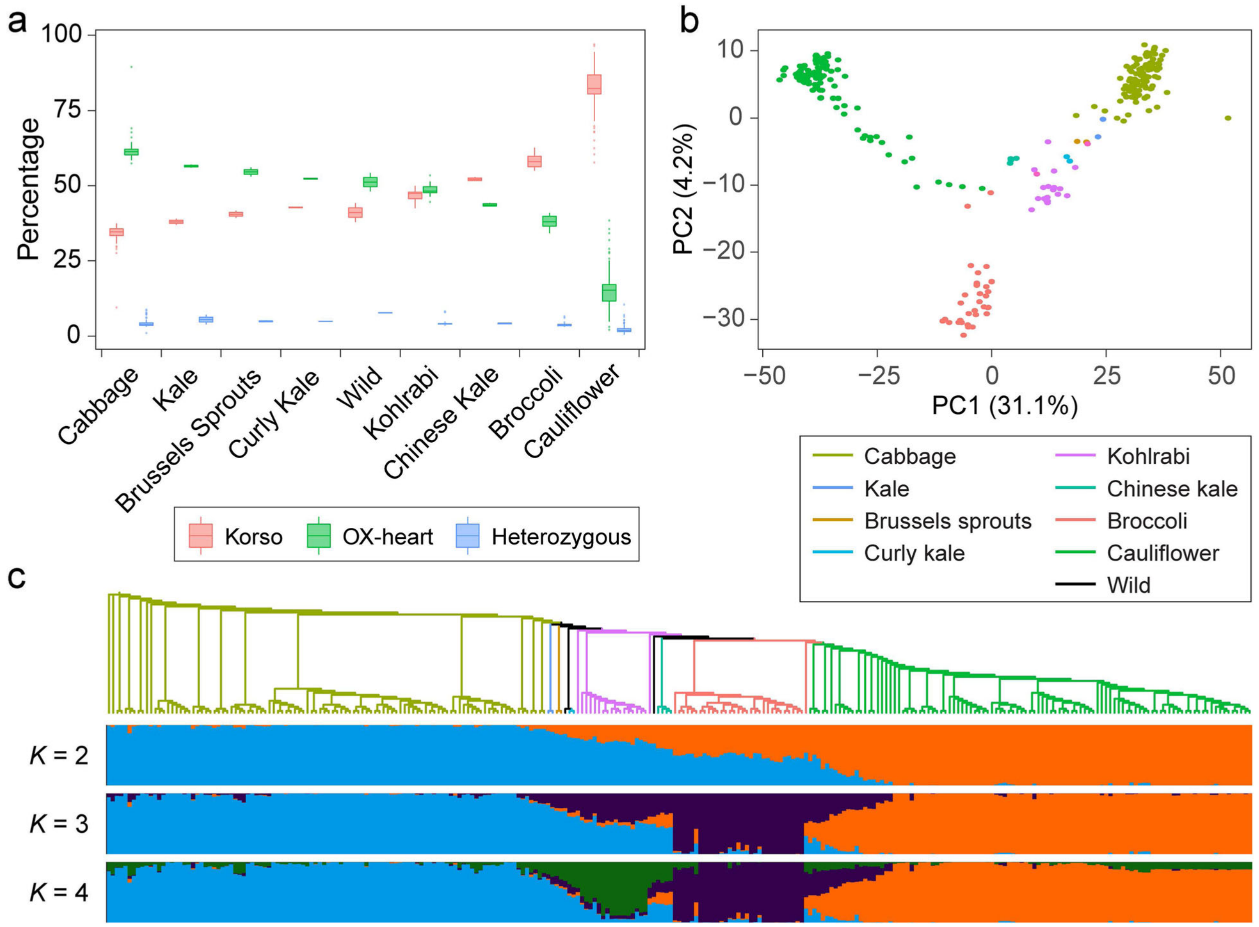

Fig. 2. SVs in different *B. oleracea* morphotypes. **a** Percentages of SVs with different genotypes in accessions of different morphotypes. **b** Principal component analysis of *B. oleracea* accessions based on SVs. **c** Maximum-likelihood tree and model-based clustering of the 271 *B. oleracea* accessions using SVs. Branch colors of the tree indicate different morphotypes as in **b**. *K*, number of ancestral kinships

between cauliflower and cabbage populations (Fig. 3a). Among these SVs, 49,285 (98.8%) had significantly higher allele frequencies of Korso genotypes in cauliflowers than in cabbages, while only 550 represented higher OX-heart allele frequencies in cauliflowers than in cabbages. These potentially selected SVs were distributed across the chromosomes without conspicuous hotspots (Additional file 1: Figure S13). Such prevalence of selected SVs across the genome is consistent with the relatively large divergence time (~ 1.68 mya) between the two highly specialized *B. oleracea* morphotypes and their independent evolution and domestication history (Fig. 1b).

In Korso and OX-heart genomes, 21,111 and 21,400 genes, respectively, overlapped with at least one selected SV in their gene bodies or promoter regions, with 6059 and 6344 overlapping with selected SVs in CDS regions. GO enrichment analyses of these genes with selected SVs revealed that those related to signal transduction, response to stimulus, cell differentiation, cell cycle, embryo development, cell growth and cell death, and flower development were significantly overrepresented (Fig. 3b), some of which showed potential associations with the distinct phenotypes of cauliflower and cabbage, such as flower development.

Selected SVs provide insights into the evolution of cauliflower curd formation

The curd of cauliflower is composed of a spirally iterative pattern of primary inflorescence meristems with floral primordia arrested in their development [26, 27]. The first insight in genetic control of the curd-like structure was achieved through characterization of the *Arabidopsis ap1* and *cal* double mutant with a cauliflower curd phenotype [28]. Subsequently, several studies indicated that the genetic nature of the cauliflower curd appears more complex [29–31]. Here, we retrieved a total of 294 genes harboring selected SVs in their promoters or gene regions and whose homologs in *Arabidopsis* have been reported to function in flowering time and

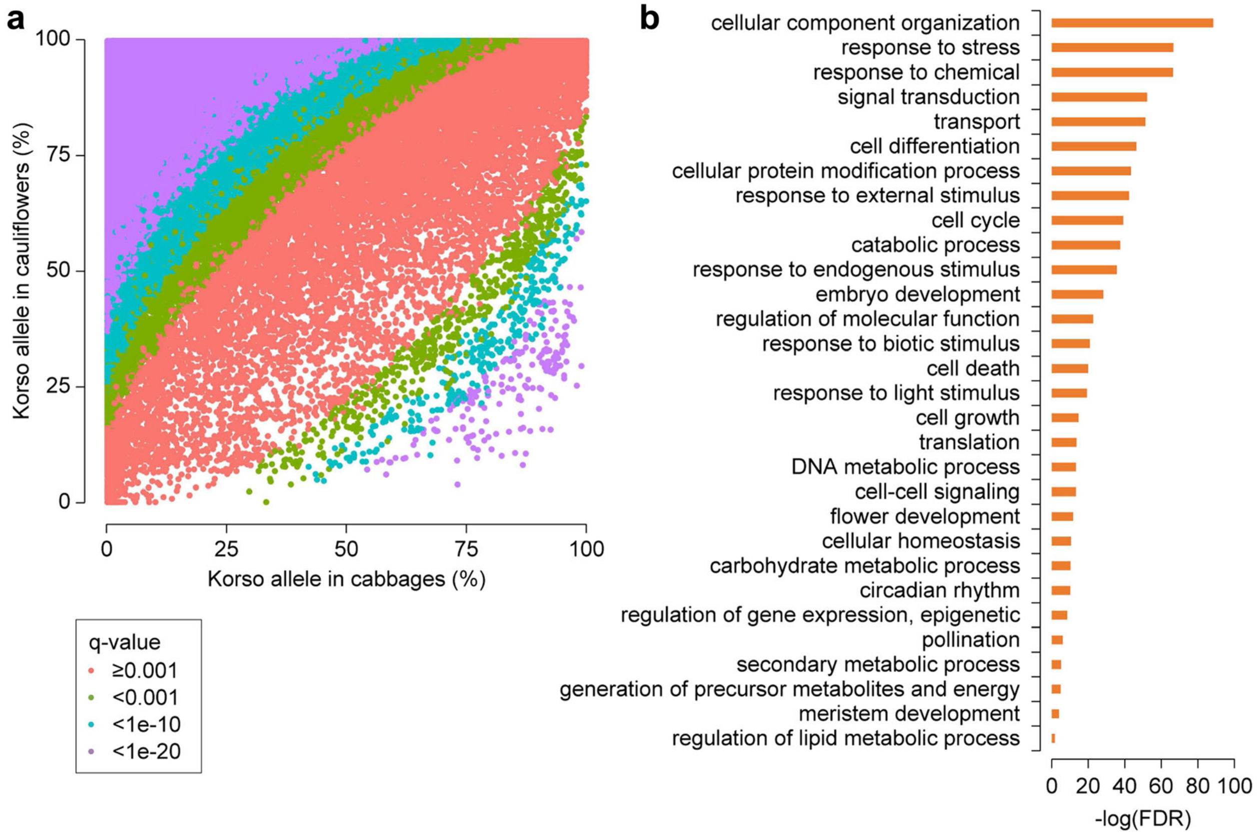

Fig. 3. SV divergence between cauliflower and cabbage. **a** Scatter plot showing Korso-allele frequencies of SVs in cabbage and cauliflower groups. *q*-value, Bonferroni-corrected *p* value of Fisher's exact test. **b** GO terms enriched in genes overlapping with SVs with significant allele frequency differences between cabbage and cauliflower groups

floral development, meristem maintenance and determination, organ size control, and shoot or inflorescence architecture (Additional file 11). In addition, RNA-Seq analysis of five stages from vegetative shoot apical meristem (SAM) to enlarged curd was conducted to reveal the potential roles of SVs in curd formation and development.

Transition from vegetative to generative development

The first stage of curd initiation corresponds with the switch from vegetative to generative development (Fig. 4a). Timely transition to the generative stage in cauliflower is essential for curd formation, while for cabbage a prolonged vegetative stage is needed for the proper development of the leafy head. The MADS box transcription factor FLC, a flowering time integrator in the vernalization and autonomous pathways, acts as a repressor of flowering [32, 33]. Several studies have demonstrated the roles of *FLC* paralogues in flowering time in diverse *B. oleracea* morphotypes [24, 34–37]. A 3371-bp insertion (SV_b_92666a) in the promoter of *BoFLC1.1* in Korso was found under strong differential selection, present in 99% and 88% of the cauliflower and broccoli accessions, respectively, while only in 9% of cabbage accessions (Fig. 4b). *BoFLC1.1* and its two tandem paralogs (*BoFLC1.2* and *BoFLC1.3*), as well as *BoFLC3* were all significantly down-regulated at the transition stage (Fig. 4c). The Korso allele of *BoFLC3* contains a 263-bp deletion (SV_w_24534) and a 49-bp insertion (SV_w_24533) in the first intron. The effect of the structure of the *FLC* first intron on flowering time has been reported in *Arabidopsis* and cruciferous crops [24, 38, 39]. We found that at these two SV loci, the Korso alleles were predominant in cauliflower (86.7% and 86.4%) and broccoli (96.9% and 92.9%), but rare in the cabbage accessions (9.7% and 8.7%) (Fig. 4b).

The FLC function is activated by FRI [40], which has been identified as a candidate gene in the QTL region for temperature-dependent timing of curd induction in cauliflower [41]. Two *FRI* homologs, *BoFRI1* and *BoFRI2*, were identified in both Korso and OX-heart genomes. A 743-bp deletion (SV_b_96002) in the promoter region of *BoFRI1* characterized the Korso allele. Most of the cauliflowers (98.0%) contained the homozygous Korso genotype, while the majority of cabbages (87.0%) harbored the homozygous OX-heart genotype

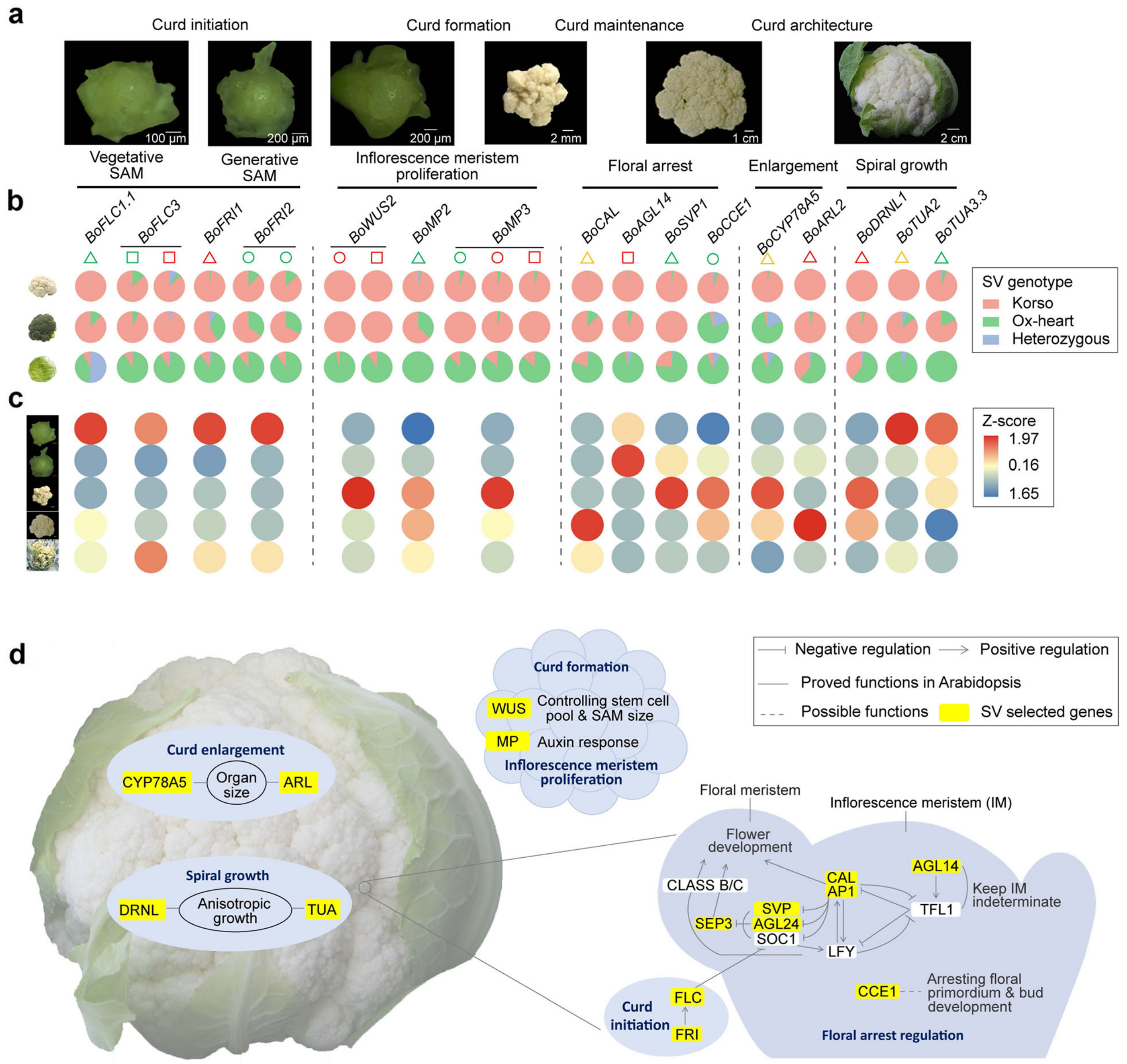

Fig. 4. SV contribution to cauliflower curd formation. **a** SAM and curd at different stages of curd development. **b** Korso-allele frequencies of SVs overlapping with candidate genes in cauliflower, broccoli, and cabbage. Triangles, circles, and squares indicate SVs in promoter, CDS, and intron regions, respectively, and their different colors, green, red, and yellow, indicate different types of SVs, insertion, deletion, and substitution, respectively, in Korso compare to OX-heart. **c** Heatmaps showing expression of candidate genes at different stages of curd development. The expression values (FPKM) were normalized by *Z*-score. **d** Proposed regulatory network for curd formation and development of cauliflower. Genes with yellow background had selected SVs between cauliflower and cabbage

(Fig. 4b). For *BoFRI2*, two insertions (12- and 21-bp, SV_w_31837, and SV_w_31838) were identified in its coding region, both displaying significant differences of genotype frequencies between cauliflower and cabbage (Fig. 4b). These two indels have been found to be related to winter annual or biennial habit of cauliflower and cabbage [25]. FES and SUF can form a putative transcription activator complex with FRI to promote *FLC* expression [42, 43]. The *B. oleracea* homologs *BoFES1.1* and *BoSUF4.2* harbored selected SVs in cauliflowers compared to cabbages, and their expression was significantly down-regulated from the vegetative to the transition stage in cauliflower, similar to that of *BoFLC1s* and *BoFLC3* (Additional file 11). Other genes involved in regulating *FLC* expression, including those involved in epigenetic modification such as the PRC1 and PRC2 complex components *BoVIN3*, *BoVIL2.3*, and *BoVRN1.1*, also harbored selected SVs (Additional file 11). Together, these results suggested that the FLC-related autonomous and vernalization pathways might be affected by the differential SVs between cauliflower and cabbage, possibly contributing to their different timing of switch to the generative stage.

Inflorescence meristem proliferation

The main process following curd initiation is the continuously regular spiral proliferation of undetermined inflorescence meristems that form the curd. Stem cell maintenance and meristem proliferation play key roles in this process. WUSCHEL acts as an auxin response rheostat to maintain apical stem cells in *Arabidopsis* [44]. We identified a 12-bp in-frame deletion (SV_w_83072) in the second exon and a 21-bp insertion (SV_w_83073) in the first intron of *BoWUS2* in Korso. All sampled cauliflower and broccoli accessions had the homozygous Korso genotypes for both SVs, while the Korso alleles were rare (4%) in cabbage (Fig. 4b). The expression of *BoWUS2* was significantly up-regulated from vegetative to curd formation, with the highest expression at the curd formation stage (Fig. 4c), implying that these two SVs could play roles in the curd formation.

MP/ARF5 together with ANT and AIL play key roles in auxin-dependent organ initiation and phyllotactic patterning [45, 46]. Selected SVs in promoters and gene regions of their homologs in *B. oleracea*, *BoMP2*, *BoMP3*, *BoANT*, *BoAIL5*, *BoAIL6*, and *BoAIL7*, were identified (Additional file 11). A 23-bp insertion (SV_w_71238) in the promoter of *BoMP2* in Korso is under strong selection in cauliflower (96.1% and 0% in cauliflower and cabbage accessions, respectively). An 11-bp insertion and a 23-bp deletion (SV_w_92482 and SV_w_92481) in the CDS and a 14-bp deletion (SV_w_92433) in the intron of *BoMP3* in Korso are under strong selection in cauliflower (95.6%, 96.1%, and 96.2% in cauliflower and 12%, 14%, and 13.3% in cabbage accessions, respectively) (Fig. 4b). Same as *BoWUS2*, the highest expression of *BoMP2* and *BoMP3* was also observed at the curd formation stage (Fig. 4c).

Curd maintenance and floral arrest

Cauliflower curd is composed of thousands of inflorescence meristems with floral meristems arrested in development. A large substitution (SV_b_70950) (~ 11.4 kb in OX-heart and ~ 7.7 kb in Korso) in the promoter region of the floral meristem identity (FMI) gene *BoCAL* was identified under strong selection. Almost all cauliflower (99.0%) and the majority of broccoli (87.5%) accessions shared the Korso allele, while most cabbage accessions (79.2%) harbored the OX-heart allele at this locus (Fig. 4b), suggesting its potential role in curd formation.

Several other FMI genes including *BoAP1.2*, *BoFUL1*, *BoFUL3*, and *BoSEP3* were also affected by selected SVs (Additional file 11), and all had relatively low expression at the vegetative, transition and curd formation stages, but significantly higher expression at the curd enlargement stage (Additional file 1: Figure S14). Studies in *Arabidopsis* suggest that an antagonistic interaction between the inflorescence meristem identity (IMI) gene *TFL1* and FMI genes regulates the developmental fate transitions [47–49]. The nearly opposite expression pattern of *BoTFL1.2* compared to that of FMI genes indicated its repression role (Additional file 1: Figure S14). While no selected SVs were identified in *BoTFL1.2*, a 13-bp deletion (SV_w_84836) in the intron of its positive regulator *BoAGL14* [50] was found under strong selection (Fig. 4b), and *BoAGL14* showed the same expression pattern as *BoTFL1.2* (Additional file 1: Figure S14), suggesting potentially important roles of both *BoTFL1.2* and *BoAGL14* in floral identity arrest and inflorescence proliferation for curd formation and maintenance. SVP is a key negative regulator of floral transition [51, 52]. A 420-bp (SV_w_74120) insertion in the promoter of *BoSVP1* was found in Korso and 98.1% of other cauliflower and all broccoli accessions, while only in 25.2% of the cabbage accessions (Fig. 4b). *BoSVP1* was significantly up-regulated from vegetative to transition stage and kept high expression levels throughout the curd formation (Fig. 4c), indicating its repressor role in flower bud development, as reported in *Arabidopsis* [53]. A cauliflower curd-specific gene, *BoCCE1*, was reported to have a potential role in the control of meristem development/arrest [29, 54]. Here, we identified a 1505-bp insertion (SV_b_67089a) covering the entire *BoCCE1* gene body in Korso. Genotyping of this insertion revealed that the *BoCCE1* gene was present in most cauliflower accessions (97.1%), but absent in most cabbage (86.5%) and broccoli (78.1%) accessions, suggesting a possible role of *BoCCE1* in floral arrest, as broccoli buds are arrested at later developmental stages compared to cauliflower buds (Fig. 4b, c).

Curd enlargement and spiral growth

Genes involved in organ size regulation, cell division and expansion, cell cycle, etc. can regulate the curd weight. *CYP78A5* (*KLU*) has been identified in *Arabidopsis* to prevent proliferation arrest and promote organ growth [55, 56]. The high expression of cauliflower *BoCYP78A5* was exclusively detected in curds (Additional file 11), especially at the curd formation and enlargement stages (Fig. 4c). A 2775-bp substitution (SV_b_76292) in the promoter of *BoCYP78A5*, present in 98.1% of cauliflower accessions while only in 8.2% of cabbage accessions (Fig. 4b), might contribute to the curd-specific expression of *BoCYP78A5*. *BoARL2* (or *CDAG1*) has been proved to play a role in the promotion of cauliflower curd size [57]. *BoARL2* was highly expressed in curd, with the highest expression during the curd enlargement phase of Korso (Fig. 4c). A 269-bp deletion (SV_w_38468) was detected in the promoter of *BoARL2*, and was present in all cauliflower and most broccoli (96.9%) accessions, while in only 41.7% of cabbage accessions (Fig. 4b).

The spiral arrangement of inflorescences is typical for cauliflower curds [27]. Transcription of the *DRNL* gene marks lateral organ founder cells in the peripheral zone of the inflorescence meristem [58]. We identified a 258-bp deletion (SV_w_30645) in the promoter of *BoDRNL1*, which was present in all cauliflower and broccoli accessions while only in 40.6% of cabbage accessions (Fig. 4b). *BoDRNL1* was specifically expressed in the curd with the highest expression at curd formation and enlargement stages (Fig. 4c), implying its potential role in determining curd architecture. Selected SVs were also identified in the alpha-tubulin gene *BoTUA2* and four *BoTUA3* genes (Fig. 4b and Additional file 11), whose homolog in *Arabidopsis* causes helical growth [59].

Discussion

The species *B. oleracea* includes a number of important vegetable crops displaying exceptionally high morphological diversity, with cauliflowers and cabbages representing two extreme morphotypes. In this study, we assembled high-quality chromosome-scale genome sequences for inbred lines of cauliflower and cabbage by integrating PacBio long-read sequences and Hi-C chromatin contact maps, which add important resources for future research and improvement of *B. oleracea* crops and provide the foundation for comprehensively exploring the phenotypic diversity of *B. oleracea*.

SVs play vital roles in the genetic regulation of plant phenotypic changes and are often the causative genetic variants for many important traits that are targets of crop domestication and breeding. However, population analysis of SVs in crops lags far behind that of SNPs, mainly due to the technological difficulties in accurate SV identification. The currently widely used SV calling approaches depend on the mapping of short sequencing reads to a reference genome, which are prone to both high false positive and high false negative rates [60]. The recent advances in long read sequencing technologies such as PacBio and Nanopore have helped read mapping in the detection of SVs. However, due to the restricted read length, some large SVs (e.g., insertions) cannot be detected [18]. In the present study, through direct comparison of the high-quality, reference-grade genome assemblies of cauliflower and cabbage, combined with long read mapping, we were able to identify ~ 120 K high-confidence SVs, with a number of them larger than 100 kb. Genotyping of this reference set of SVs in a population comprising 271 accessions representing different *B. oleracea* morphotypes and investigation of allele frequency difference of these SVs in different morphotype populations, mainly cauliflower, cabbage, and broccoli, revealed numerous SVs that are under selection in certain morphotypes, with many affecting genes associated with the corresponding unique phenotypes.

The curd of cauliflower is composed of thousands of inflorescence meristems that are spirally arranged on short enlarged inflorescence branches. This makes cauliflower an ideal model to analyze the genetic mechanism of inflorescence development and extreme organ genesis. SVs selected in cauliflower affected many genes. Combined with the analysis of expression profiles during curd development, we identified dozens of key SVs and associated genes that had potential associations with the unique curd phenotype of cauliflower. These included genes with roles in the different developmental stages of curd development. The first stage is curd initiation, involving the transition from the vegetative stage to the generative stage, with genes involved in flowering-time regulation affected (e.g., *FLC* and *FRI*). An essential step in curd formation is inflorescence proliferation, with genes like *WUS* and *MP* having cauliflower and broccoli-specific SVs. Cauliflower curds are further characterized by the floral meristem arrests, matching several floral identity genes (e.g., *CAL*, *AP1*, and *SEP3*) as well as their potential negative regulatory genes (e.g., *AGL14*, *SVP*, and *CCE1*) affected by selected SVs. Several genes with roles in cauliflower curd architecture were also affected by selected SVs. These include genes that play likely roles in organ size control (e.g., *CYP78A5* and *ARL*) and the curd spiral organization (e.g., *DRNL* and *TUA*) (Fig. 4d). Our analyses demonstrated the important contributions of SVs to the unique curd phenotype of cauliflower and shed light on the regulatory network of cauliflower curd formation.

Methods

Genome library construction and sequencing

Cauliflower (*B. oleracea* var. *botrytis*) accession Korso_1401 is a highly inbred line derived from Korso that was obtained from the Genebank of IPK Gatersleben (http://gbis.ipk-gatersleben.de; accession No. BRA2058) and has a white compact curd and long maturing time (> 95 d). Pointed cabbage (*B. oleracea* var. *capitate*) accession OX-heart_923, an inbred line with green pointed head and late bolting, was obtained from Vegetable Research Institute, Jiangsu Academy of Agricultural Science, Nanjing, China.

Young fresh leaves were collected from a single individual of each of the two accessions after a 24-h dark treatment and used for high molecular weight (HMW) DNA extraction using the cetyltrimethylammonium bromide method [61]. PacBio SMRTbell libraries were constructed from the HMW DNA using the SMRTbell Express Template Prep Kit 2.0 following the manufacturer's protocols (PacBio). A total of 24 Single-Molecule Real-Time (SMRT) cells (9 from PacBio RSII and 15 from PacBio Sequel) for Korso and 15 SMRT cells (all from PacBio Sequel) for OX-heart were sequenced by NextOmics Biosciences Co., Ltd. (Wuhan, China). For Illumina sequencing, paired-end libraries with insert sizes of ~ 400 bp were prepared using the NEBNext Ultra DNA Library Prep Kit and sequenced on a HiSeq 2500 system with 2× 150 bp mode. Hi-C libraries were constructed using the Proximo Hi-C plant kit following the manufacturer's instructions (Phase Genomics) and sequenced on an Illumina HiSeq X Ten system with 2 × 150 bp mode at Nextomics Biosciences.

For Korso, an optical map was generated using the Saphyr system (BioNano Genomics). Briefly, the HMW DNA labeling and staining were performed according to the manufacturer's protocols, and then loaded onto chips and imaged on the Saphyr System according to the user guide. Data processing, construction of the Direct Label and Stain (DLS) optical maps and the hybrid map assembly were performed using the BioNano Genomics Access software suite.

For genome resequencing, young leaves from 163 different *B. oleracea* accessions (89 cauliflower, 65 cabbage, and 9 broccoli accessions) were collected and used to extract DNA using the DNeasy Plant Mini Kit (Qiagen). Pair-end libraries with insert sizes of ~ 400 bp were constructed using the NEBNext Ultra DNA Library Prep Kit according to the manufacturer's instructions and sequenced on an Illumina HiSeq 2500 platform with the paired-end 2 × 150 bp (117 accessions) or 2 × 100 bp (46 accessions) mode.

Transcriptome sequencing and data processing

Seven different tissues of Korso (root, stem, leaf, curd, bud, flower, and silique) and OX-heart (root, stem, leaf, leafy head, bud, flower, and silique) were collected for transcriptome sequencing, which was mainly used to facilitate gene prediction. One to three biological replicates were conducted for each sample. Roots, stems, and leaves were sampled at the rosette stage of plants with 8–10 leaves (4–6 weeks after planting). The buds were about 2 mm in length, the flowers were blooming, and the siliques were at the developing stage including seeds. The curd of Korso and leafy head of OX-heart were collected when they were ready to be harvested. In addition, shoot apical meristem (SAM) samples from Korso were collected at the following developmental stages: vegetative, transition (curd initiation), curd formation (curd diameter of ~ 1 cm), pre-mature (curd diameter of 10 cm), and branch elongation (mature). For each SAM sample, two or three independent biological replicates were performed. RNA was extracted from each tissue using the TIANGEN RNAprep Pure Kit (Cat: No. DP441). RNA quality was assessed using an Agilent 2100 BioAnalyzer.

RNA-Seq libraries were prepared using the Illumina TruSeq RNA Sample Prep Kit and sequenced on an Illumina HiSeq X Ten system at WuXi NextCODE (Shanghai, China). Raw RNA-Seq data were preprocessed using the NGS QC Toolkit (v2.3.3) [62] to remove adaptors, low-quality bases, and reads containing more than 10% unknown bases ("N"). The cleaned reads were mapped to the reference genomes allowing up to two mismatches using HISAT2 [63]. Based on the alignments gene expression levels were estimated as fragments per kilobase of transcript per million mapped fragments (FPKM). Differential expression analysis between different developmental stages of SAM samples was performed using the DESeq package [64].

The PacBio Iso-Seq library was also constructed for Korso using equally mixed RNA samples from the seven tissues. The cDNA synthesis and amplification were performed using the NEBnext Single Cell/Low Input cDNA Sythesis & Amplification Module kit. SMRTbell libraries were constructed with the SMRTbell Express Template Prep kit 2.0. Three SMRT cells were sequenced on the PacBio Sequel system with the Sequel DNA Polymerase 2.0 and Sequel Sequencing Plate 2.0 by NextOmics Biosciences Co., Ltd. (Wuhan, China). Different subreads from the same polymerase read were used to generate circular consensus sequences (CCSs). The CCSs were then classified as full-length, non-chimeric, and non-full-length according to the presence or absence of 5′-primer, 3′-primer, and poly A/T tails. These sequences were then clustered using an Iterative Clustering and Error (ICE) correction algorithm incorporated in the IsoSeq_SA3nUP pipeline (https://github.com/PacificBiosciences/IsoSeq_SA3nUP). The resulting consensus isoforms were first polished using the non-full-length reads and then the RNA-Seq reads with LoRDEC (v0.3) [65].

Genome size estimation

The genome sizes of Korso and OX-heart were estimated using flow cytometry. For each sample, ~ 1 g leaves were finely chopped with a razor blade in 2000 μl LB01 isolation buffer [66]. The resulting suspension was filtered through 30-μm nylon, and then 2 μl 10 mg ml^{-1} RNase I was added at room temperature for 15 min. After centrifugation at 1000 r/min for 5 min, the supernatant was discarded, and the precipitated nuclei were collected. The nuclear DNA was fluorescently labeled with PI (Propidium iodide) staining solution and stained in the dark for 30 minutes. The DNA peak ratio was assessed by flow cytometry (BD FACSCalibur system, BD Biosciences) using *B. rapa* Chiifu-401-42 [20] as the internal reference. The ModFit software (Verity Software House) was used for data analyses.

De novo genome assembly

PacBio reads were de novo assembled using Falcon (v1.8.7) [67]. The assembled contigs were first polished with Arrow [68] using PacBio long reads, and then further polished with Illumina shotgun reads using Pilon [69]. The polished Korso contigs was further improved by BioNano optical genome maps. The PacBio contigs were anchored to optical maps to construct scaffolds and the resulting gaps between connected contigs were filled using PBJelly (https://github.com/esrice/PBJelly) with the following parameters: '-minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestn 1 -nCandidates 10 -maxScore -500 -noSplitSubreads'. The resulting contigs were polished again by Arrow and Pilon as described above.

To remove potential contaminations in the assemblies, the final contigs were divided into 50-kb bins and then searched against the GenBank nucleotide (nt) database using blastn [70]. Sequences with best hits not in the green plants were possible contaminations and discarded.

To construct pseudomolecules, raw Hi-C reads were processed to trim adapters and low-quality sequences using Trimmomatic [71] with parameters 'SLIDINGWINDOW:4:20 MINLEN:50'. The cleaned reads were aligned to the final contigs using bowtie2 end-to-end algorithm [72]. HiC-Pro pipeline [73] was then used to remove duplicated read pairs, detect valid ligation products, and perform quality controls. The assembled contigs were then clustered, ordered, and oriented into pseudomolecules using Lachesis [74].

Repetitive element and centromere prediction

A custom repeat library was constructed for each genome according to the pipeline described in http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced, using MITE-Hunter [75], RepeatModeler (http://www.repeatmasker.org/RepeatModeler/), and LTRharvest and LTRdigest from GenomeTools [76]. Repeat sequences were identified by scanning each genome assembly using the corresponding repeat library with RepeatMasker (http://www.repeatmasker.org/).

Full-length long terminal repeat retrotransposons (LTR-RTs) were identified using LTR_finder [77] and LTR_harvest [78], and redundancies in the full-length LTR-RTs were then removed using LTR_retriver [79]. The substitution rate between the two end sequences of each LTR-RT was calculated using PAML [80]. The LTR expansion time was estimated according to the formula $T=S/2\mu$, where S is the substitution rate and μ is the mutation rate (1.5×10^{-8} per site per year) [81].

The previously reported centromeric satellite repeats, including CentBr, CRB, TR238, and PCRBr [82], were used to scan the nine chromosomes of both Korso and OX-heart genome assemblies. The locations of centromeres were estimated based on the peak regions of these centromeric satellite repeats.

Prediction and annotation of protein-coding genes

Protein-coding genes were predicted from Korso and OX-heart assemblies using EVidenceModeler (EVM) [83] by integrating transcript evidence, *ab initio* prediction and protein homology searching. For transcript evidence, the Illumina RNA-Seq reads were mapped to the reference genomes using HISAT2 [63] and assembled into transcripts using StringTie [84]. PASA [83] was then used to determine potential intron-exon boundaries. The polished PacBio Iso-Seq reads of Korso were mapped to the Korso genome using Minimap2 [85] with parameters '--secondary=no -ax splice -uf -C5 -t 8' and also used as the transcript evidence. For *ab initio* prediction, AUGUSTUS [86], SNAP [87], and GENSCAN [88] were employed. AUGUSTUS and SNAP were trained for each genome using the high-confident gene models obtained with the PASA analysis, while the *Arabidopsis* gene models were used to train GENSCAN. For protein homology searching, protein sequences from relative and model species (*Arabidopsis thaliana*, *Aethionema arabicum*, *Brassica napus*, *Brassica nigra*, *Brassica rapa*, *Capsella rubella*, *Thellungiella halophilla*, *Brassica oleracea* TO1000, and HDEM) were aligned to the genome assemblies using GenBlastA [89] and based on the alignments GeneWise [90] was then used to predict gene structures. Finally, EVM was used to generate a consensus gene set for each genome by integrating evidence from transcript mapping, protein homology, and ab initio predictions. To annotate the predicted protein-coding genes, their protein sequences were searched against GenBank nr, SwissProt, KEGG, and TrEMBL protein databases. GO term annotation and enrichment analysis were carried out using the Blast2GO suite [91].

SV identification between Korso and OX-heart genomes

To identity SVs between genomes of OX-heart cabbage and Korso cauliflower, we first aligned the two genomes using the Minimap2 [85] with parameters '-ax asm5'. The resulting alignments were analyzed using Assemblytics [92] to call SVs. The resulting SVs spanning or close (distance < 50 bp) to gap regions in either of the two genomes were removed.

We further identified SVs by aligning PacBio reads from OX-heart to the Korso genome and Korso PacBio reads to the OX-heart genome, using Minimap2 with parameters: '-eqx -L -O 5,56 -e 4,1 -B 5 --secondary=no -z 400,50 -r 2000 -Y --MD -ax map-pb'. Based on the alignments, SVs were called using pbsv (https://github.com/PacificBiosciences/pbsv). The identified SVs spanning gap regions in the genomes were discarded. To further evaluate the SVs, sequences of the two 5-kb flanking regions of each SV were extracted from the genome and then blasted against to the other genome. The unique alignments between the two genomes identified by Assemblytics were used to filter SVs identified by pbsv. SVs were kept if the reliable blast hits of the two flank sequences (alignment length > 50 bp, identity > 90%, and *e*-value <1e−10) were found in the expected region on the query genome, and the gap size between the two hits was consistent with that estimated by pbsv. Specifically, for insertions, we required that the deviation between pbsv estimated size and the distance observed between the two blast hits was smaller than 20%. For deletions, the allowed gap or overlap between the two blast hits of flanking regions should be smaller than 3 bp.

It is noteworthy that, besides the simple indels with defined breakpoints, Assemblytics also reported four types of complex SVs without defined breakpoints, including repeat expansion, repeat contraction, tandem expansion and tandem contraction. If SVs identified by pbsv were in the regions of these complex SVs, their precise breakpoints could be defined, i.e., these complex SVs could be converted into one or more simple indels. SVs identified by Assemblytics and pbsv were merged if they overlapped with each other by at least 50% of their lengths.

Genotyping of SVs in *B. oleracea* accessions

Raw genome sequencing reads of the 163 *B. oleracea* accessions generated in this study and 108 accessions reported previously [14] were first processed to consolidate duplicated read pairs into unique read pairs. Duplicated read pairs were defined as those having identical bases in the first 90 bp for 100-bp reads or 100 bp for 150-bp reads of both left and right reads. The resulting reads were then processed to trim adapters and low-quality sequences using Trimmomatic [71] with parameters 'SLIDINGWINDOW:4:20 MINLEN:50'.

To genotype SVs in these accessions, the cleaned reads were aligned to the OX-heart and Korso genomes, respectively, using BWA-MEM [93], allowing no more than 3% mismatches. For each SV in each accession, we checked reads aligned to the regions spanning the breakpoints of the SV in both OX-heart and Korso genomes. For each breakpoint, we first required at least 3 split reads to support the SV call. If there were not enough split reads supporting the SV, we then checked the read coverage in the SV region. For a deletion, we required that < 50% of the deleted region was covered by reads with 2× depth, while > 50% of at least one flanking region with the same length of the deleted region was covered. Based on the split read and read depth information, SVs in a particular accession could be classified as Ox-heart genotype (same genotype as Ox-heart), Korso genotype (same genotype as Korso), heterozygous (containing both Ox-heart and Korso genotypes), and undetermined (genotype that were not able to be determined due to insufficient read mapping information).

For population analyses, we divided the genome into 25-kb non-overlapping windows and randomly selected one SV per window. The genotype data of the chosen SVs of the entire *B. oleracea* population were used to construct a maximum-likelihood phylogenetic tree using IQ-TREE [94] with 1000 bootstraps. The same SVs were also used to perform the principal component analysis (PCA) with TASSEL5 [95] and to investigate population structure of *B. oleracea* accessions with STRUCTURE [96]. Allele frequencies of 84,571 SVs with the genotype determined in at least 50% of accessions in both cauliflower and cabbage populations were calculated. Significance of the difference of the SV allele frequencies between cauliflower and cabbage groups was determined using Fisher's exact test, and the resulting raw *P* values were corrected using the Bonferroni method. SVs with adjusted *P* values < 0.001 and fold change ≥ 2 were defined as those highly differentiated between cauliflower and cabbage.

Availability of data and materials

The genome assemblies and raw genome and transcriptome sequencing reads of Korso and OX-heart have been deposited into the NCBI BioProject database under accession numbers PRJNA546441 [97] and PRJNA548819 [98], respectively. Raw genome sequencing reads of the 163 *B. oleracea* accessions have been deposited into the NCBI BioProject database under the accession number PRJNA700684 [99]. Genome assemblies and annotations of Korso and OX-heart are also available at Figshare [100].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-021-01031-2.

Additional file 1: Figure S1. Morphology of cauliflower Koroso_1401 and pointed cabbage OX-heart_923. **Figure S2**. Genome size estimation using flow cytometry analysis for cauliflower Korso and cabbage OX-heart. **Figure S3**. Chromosome conformation capture of cauliflower Korso and pointed cabbage OX-heart. **Figure S4.** Pairwise comparisons of the pseudo-chromosomes among cauliflower Korso, cabbage OX-heart and broccoli HDEM. **Figure S5**. Distribution of estimated insertion times of intact LTR retrotransposons in Korso, OX-heart and *B. rapa* (v3.0) genomes. **Figure S6**. Distribution of centromeric satellite repeats CentBr 1 and CentBr 2 on the nine chromosomes of Korso and OX-heart. **Figure**

S7. Segmental collinearity between the *A. thaliana* genome and genomes of Korso, OX-heart and *B. rapa*. **Figure S8**. Subgenomes in Korso and OX-heart. **Figure S9**. Detected SVs between the genomes of Korso and the OX-heart. **Figure S10**. Significantly enriched GO terms in genes overlapping with detected SVs in their gene body and/or promoter regions. **Figure S11**. Distribution of SV genotyping rates in 271 *B. oleracea* accessions. **Figure S12**. Population analysis of the 271 *B. oleracea* accessions using SNPs. **Figure S13**. Manhattan plot displaying the -log(q-value) of allele variations between cabbage and cauliflower for SVs across the Korso genome. **Figure S14**. Expression profiles of genes potentially involved in curd maintenance and floral arrest.

Additional file 2: Summary statistics of *Brassica oleracea* genome assemblies.

Additional file 3:. Base accuracy in genome assemblies estimated by consensus error calls using Illumina short reads.

Additional file 4:. Summary of RNA-seq data of 'Korso' and 'OX-heart'.

Additional file 5:. Repeat sequences in Korso, OX-heart and HDEM genome assemblies.

Additional file 6: Statistics of intact LTR-retrotransposons in Korso, OX-heart and *Brassica rapa* (V3.0) genome assemblies.

Additional file 7:. Insertions/deletions identified between the Korso and OX-heart genomes.

Additional file 8:. Distribution of SVs in different genomic regions.

Additional file 9: Summary of the SVs genotyping in 271 *Brassica oleracea* accessions.

Additional file 10:. Summary statistics of SV genotyping using Illumina short reads in the reference accessions.

Additional file 11:. List of candidate curd formation genes associated with selected SVs.

Acknowledgements

We thank Dr. Evelyn Klocke for providing the original seeds of Korso. This research was supported by grants from the National Key R&D Program of China (2016YFD0100204-14 to F.L. and 2016YFD0101702 to Jianbin Li), President Foundation for Science of Beijing Academy of Agricultural and Forestry Sciences (YZJJ201701 to F.L.), International Cooperation Foundation of Beijing Academy of Agricultural and Forestry Sciences (GJHZ2017-2 to F.L. and N.G.), National Modern Agriculture Industry Technology System (CARS-25 to Jianbin Li), Innovation Program of Beijing Academy of Agricultural and Forestry Sciences (KJCX20200205 to F.L. and KJCX20200108 to N.G.), National Natural Science Foundation of China (32072576 to N.G.). Overseas Training Program of State Administration of Foreign Experts Affairs of China (P172001038 to N.G.), and the US National Science Foundation (IOS-1855585 to Z.F.).

Authors' contributions

F.L., Z.F., Jianbin Li, G.B., and N.G. designed and managed the project. Jianbin Li., S.W., C.Z., H.J., G.B., S.H., M.Z., D.L., F.Y., and W.Z. collected samples. Jingjing Li, M.Y., and Y.L. contributed to the Korso and OX-heart genome assemblies and annotations. M.D., G.W., L.M., and N.G. generated RNA-Seq data. L.G., Y.L., H.S., and N.G. performed SV calling and genotyping. L.G., N.G., Y.L., Y.P., T.B., H.S., E.L., and X.W. performed data analysis. N.G. and L.G. wrote the manuscript. Z.F., F.L., G.B., and Y.X. revised the manuscript. The authors read and approved the final manuscript.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]National Engineering Research Center for Vegetables, Beijing Academy of Agriculture and Forestry Sciences, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (North China), Beijing Key Laboratory of Vegetable Germplasm Improvement, Beijing 100097, China. [2]Jiangsu Key Laboratory for Horticultural Crop Genetic Improvement, Institute of Vegetable Crops, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China. [3]Boyce Thompson Institute, Ithaca, NY 14853, USA. [4]CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Innovative Academy of Seed Design, Chinese Academy of Sciences, Wuhan 430074, China. [5]Nextomics Biosciences Institute, Wuhan 430074, China. [6]University of Chinese Academy of Sciences, Beijing 100049, China. [7]Plant Breeding, Wageningen University and Research, 6708 PB Wageningen, The Netherlands. [8]Tianjin GengYun Seed Company, Tianjin 300400, China. [9]Fuzhou Institute of Vegetable Science, Fuzhou 350012, China. [10]US Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA.

Received: 14 December 2020 Accepted: 19 April 2021
Published online: 05 May 2021

References

1. Dixon G. Origins and diversity of *Brassica* and its relatives. In: Vegetable Brassicas and Related Crucifers. 1st ed. Wallingford: CABI; 2007. p. 1–33.
2. Cheng F, Wu J, Wang X. Genome triplication drove the diversification of *Brassica* plants. Hortic Res. 2014;1(1):14024. https://doi.org/10.1038/hortres.2014.24.
3. Parkin IA, et al. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biol. 2014;15(6):R77. https://doi.org/10.1186/gb-2014-15-6-r77.
4. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, et al. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun. 2014;5(1):3930. https://doi.org/10.1038/ncomms4930.
5. Lv H, Wang Y, Han F, Ji J, Fang Z, Zhuang M, et al. A high-quality reference genome for cabbage obtained with SMRT reveals novel genomic features and evolutionary characteristics. Sci Rep. 2020;10(1):12394. https://doi.org/10.1038/s41598-020-69389-x.
6. Cai X, Wu J, Liang J, Lin R, Zhang K, Cheng F, et al. Improved *Brassica oleracea* JZS assembly reveals significant changing of LTR-RT dynamics in different morphotypes. Theor Appl Genet. 2020;133(11):3187–99. https://doi.org/10.1007/s00122-020-03664-3.
7. Sun D, et al. Draft genome sequence of cauliflower (*Brassica oleracea* L. var. *botrytis*) provides new insights into the C genome in *Brassica* species. Hortic Res. 2019;6:82.
8. Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants. 2018;4(11):879–87. https://doi.org/10.1038/s41477-018-0289-4.
9. Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, et al. The population genetics of structural variants in grapevine domestication. Nat Plants. 2019;5(9):965–79. https://doi.org/10.1038/s41477-019-0507-8.
10. Song JM, Guan Z, Hu J, Guo C, Guo L. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. Nat Plants. 2020;6(1):34–45. https://doi.org/10.1038/s41477-019-0577-7.
11. Lu S, van Eck J, Zhou X, Lopez AB, O'Halloran DM, Cosman KM, et al. The cauliflower *Or* gene encodes a DnaJ cysteine-rich domain-containing protein that mediates high levels of beta-carotene accumulation. Plant Cell. 2006;18(12):3594–605. https://doi.org/10.1105/tpc.106.046417.
12. 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. 2019;98:524–39.
13. Yin S, et al. Transposon insertions within alleles of *BnaFLC.A10* and *BnaFLC.A2* are associated with seasonal crop type in rapeseed. J Exp Bot. 2020;71:4729–41.
14. Cheng F, Sun R, Hou X, Zheng H, Zhang F, Zhang Y, et al. Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. Nat Genet. 2016;48(10):1218–24. https://doi.org/10.1038/ng.3634.
15. Stansell Z, Hyma K, Fresnedo-Ramírez J, Sun Q, Mitchell S, Björkman T, et al. Genotyping-by-sequencing of *Brassica oleracea* vegetables reveals unique phylogenetic patterns, population structure and domestication footprints. Hortic Res. 2018;5(1):38. https://doi.org/10.1038/s41438-018-0040-3.
16. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-

molecule sequencing. Nat Methods. 2018;15(6):461–8. https://doi.org/10.1038/s41592-018-0001-7.

17. Liu Y, du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-genome of wild and cultivated soybeans. Cell. 2020;182(1):162–76. https://doi.org/10.1016/j.cell.2020.05.023.
18. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell. 2020;182(1):145–61. https://doi.org/10.1016/j.cell.2020.05.021.
19. Waterhouse RM, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017;35:543–8.
20. Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, et al. Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. Hortic Res. 2018;5(1):50. https://doi.org/10.1038/s41438-018-0071-9.
21. Xiong Z, Pires JC. Karyotype and identification of all homoeologous chromosomes of allopolyploid *Brassica napus* and its diploid progenitors. Genetics. 2011;187(1):37–49. https://doi.org/10.1534/genetics.110.122473.
22. Wang X, et al. The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet. 2011;43(10):1035–9. https://doi.org/10.1038/ng.919.
23. Xie T, Zhang FG, Zhang HY, Wang XT, Hu JH, Wu XM. Biased gene retention during diploidization in *Brassica* linked to three-dimensional genome organization. Nat Plants. 2019;5(8):822–32. https://doi.org/10.1038/s41477-019-0479-8.
24. Lin Y, et al. Subtropical adaptation of a temperate plant (*Brassica oleracea* var. *italica*) utilizes non-vernalization-responsive QTLs. Sci Rep. 2018;8:13609.
25. Irwin JA, Lister C, Soumpourou E, Zhang Y, Howell EC, Teakle G, et al. Functional alleles of the flowering time regulator *FRIGIDA* in the *Brassica oleracea* genome. BMC Plant Biol. 2012;12(1):21. https://doi.org/10.1186/1471-2229-12-21.
26. Sadik S. Morphology of the curd of cauliflower. Am J Bot. 1962;49(3):290–7. https://doi.org/10.1002/j.1537-2197.1962.tb14940.x.
27. Kieffer M, Fuller MP, Jellings AJ. Explaining curd and spear geometry in broccoli, cauliflower and 'romanesco': quantitative variation in activity of primary meristems. Planta. 1998;206(1):34–43. https://doi.org/10.1007/s004250050371.
28. Kempin SA, Savidge B, Yanofsky MF. Molecular basis of the cauliflower phenotype in *Arabidopsis*. Science. 1995;267(5197):522–5. https://doi.org/10.1126/science.7824951.
29. Duclos DV, Björkman T. Meristem identity gene expression during curd proliferation and flower initiation in *Brassica oleracea*. J Exp Bot. 2008;59(2):421–33. https://doi.org/10.1093/jxb/erm327.
30. Smith LB, King GJW. The distribution of *BoCAL-a* alleles in *Brassica oleracea* is consistent with a genetic model for curd development and domestication of the cauliflower. Mol. Breed. 2000;6(6):603–13. https://doi.org/10.1023/A:1011370525688.
31. Labate JA, Robertson LD, Baldo AM, Björkman T. Inflorescence identity gene alleles are poor predictors of inflorescence type in broccoli and cauliflower. J Am Soc Hortic Sci. 2006;131(5):667–73. https://doi.org/10.21273/JASHS.131.5.667.
32. Michaels SD, Amasino RM. *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell. 1999;11:949–56.
33. Sheldon CC, Burn JE, Perez PP, Metzger J, Edwards JA, Peacock WJ, et al. The *FLF* MADS box gene: a repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. Plant Cell. 1999;11(3):445–58. https://doi.org/10.1105/tpc.11.3.445.
34. Okazaki K, Sakamoto K, Kikuchi R, Saito A, Togashi E, Kuginuki Y, et al. Mapping and characterization of *FLC* homologs and QTL analysis of flowering time in *Brassica oleracea*. Theor Appl Genet. 2007;114(4):595–608. https://doi.org/10.1007/s00122-006-0460-6.
35. Razi H, Howell E, Newbury H, Kearsey M. Does sequence polymorphism of *FLC* paralogues underlie flowering time QTL in *Brassica oleracea*? Theor Appl Genet. 2008;116(2):179–92. https://doi.org/10.1007/s00122-007-0657-3.
36. Irwin JA, Soumpourou E, Lister C, Ligthart JD, Kennedy S, Dean C. Nucleotide polymorphism affecting *FLC* expression underpins heading date variation in horticultural *Brassicas*. Plant J. 2016;87(6):597–605. https://doi.org/10.1111/tpj.13221.
37. Ridge S, Brown PH, Hecht V, Driessen RG, Weller JL. The role of *BoFLC2* in cauliflower (*Brassica oleracea* var. *botrytis* L.) reproductive development. J Exp Bot. 2015;66(1):125–35. https://doi.org/10.1093/jxb/eru408.
38. Wang Q, Zhang Y, Zhang L. A naturally occurring insertion in the *RsFLC2* gene associated with late-bolting trait in radish (*Raphanus sativus* L.). Mol Breed. 2018;38:137.
39. Sheldon CC, Conn AB, Dennis ES, Peacock WJ. Different regulatory regions are required for the vernalization-induced repression of FLOWERING LOCUS C and for the epigenetic maintenance of repression. Plant Cell. 2002;14(10):2527–37. https://doi.org/10.1105/tpc.004564.
40. Geraldo N, Bäurle I, Kidou S, Hu X, Dean C. FRIGIDA delays flowering in *Arabidopsis* via a cotranscriptional mechanism involving direct interaction with the nuclear cap-binding complex. Plant Physiol. 2009;150(3):1611–8. https://doi.org/10.1104/pp.109.137448.
41. Hasan Y, Briggs W, Matschegewski C, Ordon F, Stützel H, Zetzsche H, et al. Quantitative trait loci controlling leaf appearance and curd initiation of cauliflower in relation to temperature. Theor Appl Genet. 2016;129(7):1273–88. https://doi.org/10.1007/s00122-016-2702-6.
42. Choi K, Kim J, Hwang HJ, Kim S, Park C, Kim SY, et al. The FRIGIDA complex activates transcription of *FLC*, a strong flowering repressor in *Arabidopsis*, by recruiting chromatin modification factors. Plant Cell. 2011;23(1):289–303. https://doi.org/10.1105/tpc.110.075911.
43. Michaels SD, Bezerra IC, Amasino RM. *FRIGIDA*-related genes are required for the winter-annual habit in *Arabidopsis*. Proc Nat Acad Sci U S A. 2004;101(9):3281–5. https://doi.org/10.1073/pnas.0306778101.
44. Ma Y, et al. WUSCHEL acts as an auxin response rheostat to maintain apical stem cells in *Arabidopsis*. Nat Commun. 2019;10:1053.
45. Yamaguchi N, Wu MF, Winter CM, Berns MC, Nole-Wilson S, Yamaguchi A, et al. A molecular framework for auxin-mediated initiation of flower primordia. Dev Cell. 2013;24(3):271–82. https://doi.org/10.1016/j.devcel.2012.12.017.
46. Bhatia N, Heisler MG. Self-organizing periodicity in development: organ positioning in plants. Development. 2018;145:dev149336.
47. Wagner D. Key developmental transitions during flower morphogenesis and their regulation. Curr Opin Genet. Dev. 2017;45:44–50. https://doi.org/10.1016/j.gde.2017.01.018.
48. Teo ZWN, Song S, Wang Y, Liu J, Yu H. New insights into the regulation of inflorescence architecture. Trends Plant Sci. 2014;19(3):158–65. https://doi.org/10.1016/j.tplants.2013.11.001.
49. Liljegren SJ, Gustafson-Brown C, Pinyopich A, Ditta GS, Yanofsky MF. Interactions among *APETALA1*, *LEAFY*, and *TERMINAL FLOWER1* specify meristem fate. Plant Cell. 1999;11(6):1007–18. https://doi.org/10.1105/tpc.11.6.1007.
50. Pérez-Ruiz RV, García-Ponce B, Marsch-Martínez N, Ugartechea-Chirino Y, Villajuana-Bonequi M, de Folter S, et al. *XAANTAL2* (*AGL14*) is an important component of the complex gene regulatory network that underlies *Arabidopsis* shoot apical meristem transitions. Mol Plant. 2015;8(5):796–813. https://doi.org/10.1016/j.molp.2015.01.017.
51. Hartmann U, Hohmann S, Nettesheim K, Wisman E, Saedler H, Huijser P. Molecular cloning of *SVP*: a negative regulator of the floral transition in *Arabidopsis*. Plant J. 2000;21(4):351–60. https://doi.org/10.1046/j.1365-313x.2000.00682.x.
52. Gregis V, et al. Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in *Arabidopsis*. Genome Biol. 2013;14:R56.
53. Liu C, Xi W, Shen L, Tan C, Yu H. Regulation of floral patterning by flowering time genes. Dev Cell. 2009;16(5):711–22. https://doi.org/10.1016/j.devcel.2009.03.011.
54. Palmer JE, et al. A *Brassica oleracea* gene expressed in a variety-specific manner may encode a novel plant transmembrane receptor. Plant Cell Physiol. 2011;42:404–13.
55. Anastasiou E, Kenz S, Gerstung M, MacLean D, Timmer J, Fleck C, et al. Control of plant organ size by *KLUH/CYP78A5*-dependent intercellular signaling. Dev Cell. 2007;13(6):843–56. https://doi.org/10.1016/j.devcel.2007.10.001.
56. Stransfeld L, Eriksson S, Adamski NM, Breuninger H, Lenhard M. *KLUH/CYP78A5* promotes organ growth without affecting the size of the early primordium. Plant Signal Behav. 2010;5(8):982–4. https://doi.org/10.4161/psb.5.8.12221.
57. Li H, Liu Q, Zhang Q, Qin E, Jin C, Wang Y, et al. Curd development associated gene (*CDAG1*) in cauliflower (*Brassica oleracea* L. var. *botrytis*) could result in enlarged organ size and increased biomass. Plant Sci. 2017;254:82–94. https://doi.org/10.1016/j.plantsci.2016.10.009.
58. Comelli P, Glowa D, Frerichs A, Engelhorn J, Chandler JW, Werr W. Functional dissection of the *DORNRÖSCHEN-LIKE* enhancer 2 during

embryonic and phyllotactic patterning. Planta. 2020;251(4):90. https://doi.org/10.1007/s00425-020-03381-7.

59. Abe T, Hashimoto T. Altered microtubule dynamics by expression of modified-tubulin protein causes right-handed helical growth in transgenic *Arabidopsis* plants. Plant J. 2005;43(2):191–204. https://doi.org/10.1111/j.1365-313X.2005.02442.x.
60. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019; 20(1):246. https://doi.org/10.1186/s13059-019-1828-7.
61. Murray MG, Thompson WF. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 1980;8(19):4321–5. https://doi.org/10.1093/nar/8.19.4321.
62. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. Plos One. 2012;7(2):e30619. https://doi.org/10.1371/journal.pone.0030619.
63. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15. https://doi.org/10.1038/s41587-019-0201-4.
64. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106.
65. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;30(24):3506–14. https://doi.org/10.1093/bioinformatics/btu538.
66. Doležel J, Binarová P, Lcretti S. Analysis of nuclear DNA content in plant cells by flow cytometry. Biologia Plantarum. 1989;31(2):113–20. https://doi.org/10.1007/BF02907241.
67. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13(12):1050–4. https://doi.org/10.1038/nmeth.4035.
68. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563–9. https://doi.org/10.1038/nmeth.2474.
69. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. Plos One. 2014;9(11):e112963. https://doi.org/10.1371/journal.pone.0112963.
70. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421. https://doi.org/10.1186/1471-2105-10-421.
71. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
72. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
73. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16(1):259. https://doi.org/10.1186/s13059-015-0831-x.
74. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31(12):1119–25. https://doi.org/10.1038/nbt.2727.
75. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;38:e199.
76. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans. Comput Biol Bioinform. 2013;10(3):645–56. https://doi.org/10.1109/TCBB.2013.68.
77. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35(Web Server):W265–8. https://doi.org/10.1093/nar/gkm286.
78. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics. 2008;9(1):18. https://doi.org/10.1186/1471-2105-9-18.
79. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018; 176(2):1410–22. https://doi.org/10.1104/pp.17.01310.
80. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91. https://doi.org/10.1093/molbev/msm088.
81. Koch MA, Haubold B, Mitchell-Olds T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol Biol Evol. 2000;17(10):1483–98. https://doi.org/10.1093/oxfordjournals.molbev.a026248.
82. Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY, Kwon SJ, et al. Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related *Brassica* species. Plant J. 2007;49(2):173–83. https://doi.org/10.1111/j.1365-313X.2006.02952.x.
83. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9(1):R7. https://doi.org/10.1186/gb-2008-9-1-r7.
84. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.
85. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
86. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005; 33(Web Server):W465–7. https://doi.org/10.1093/nar/gki458.
87. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5(1):59. https://doi.org/10.1186/1471-2105-5-59.
88. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78–94. https://doi.org/10.1006/jmbi.1997.0951.
89. She R, Chu JS, Wang K, Pei J, Chen N. GenBlastA: enabling BLAST to identify homologous gene sequences. Genome Res. 2009;19(1):143–9. https://doi.org/10.1101/gr.082081.108.
90. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004;14(5):988–95. https://doi.org/10.1101/gr.1865504.
91. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36(10):3420–35. https://doi.org/10.1093/nar/gkn176.
92. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016;32(19):3021–3. https://doi.org/10.1093/bioinformatics/btw369.
93. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013. p. 1303.3997. https://arxiv.org/abs/1303.3997.
94. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74. https://doi.org/10.1093/molbev/msu300.
95. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633–5. https://doi.org/10.1093/bioinformatics/btm308.
96. Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour. 2009;9(5):1322–32. https://doi.org/10.1111/j.1755-0998.2009.02591.x.
97. Guo et al. Genome assembly and raw genome and transcriptome sequences of cauliflower (*Brassica oleracea* var. *botrytis* cv. Korso). 2019. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA546441
98. Guo et al. Genome assembly and raw genome and transcriptome sequences of cabbage (*Brassica oleracea* var. *capitata* cv. OX-heart). 2019. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA548819
99. Guo et al. Resequencing of *Brassica oleracea*. 2021. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA700684
100. Guo et al. Genome assemblies and annotations of cauliflower (*Brassica oleracea* var. *botrytis* cv. Korso) and cabbage (*Brassica oleracea* var. *capitata* cv. OX-heart). 2021. doi: https://doi.org/10.6084/m9.figshare.c.5392466

Publisher's Note