Title: Gapless pangenome analyses reveal fast *Brassica rapa* subspeciation

Authors: Wei Ma[1]†*, Yuanming Liu[1]†, Xiaochun Wei[2]†, Xiaomeng Zhang[1]†, Xiaonan Li[3]†, Zhaokun Liu[4]†, Lingyun Yuan[5]†, Guangguang Li[6]†, Shu Zhang[1], Qihang Yang[1], Xiaocong Chang[1], Zizhuo Han[1], Hao Liang[1], Zhaoshui Luan[7], Qianyun Wang[1], Yujie Gu[1], Xinlong Wang[1], Xianlei Zhao[1], Qing Liu[1], Xiaoxue Sun[1], Mengyang Liu[1], Daling Feng[1], Yin Lu[1], Shuangxia Luo[1], Lei Yang[1], Mengyuan Li[8], Robin Allaby[9], Kai Wang[10], Tianzhen Zhang[11], Shuxing Shen[1], Yves Van de Peer[12,13,14,15]*, Yiguo Hong[1,9]*, Yuxiang Yuan[2]*, Jianjun Zhao[1]*

Affiliations:

[1]State Key Laboratory of North China Crop Improvement and Regulation, Key Laboratory of Vegetable Germplasm Innovation and Utilization of Hebei, College of Horticulture, Hebei Agricultural University; Baoding, Hebei 071001, China.

[2]Institute of Vegetables, Henan Academy of Agricultural Sciences; Zhengzhou, Henan 450002, China.

[3]College of Horticulture, Shenyang Agricultural University; Shenyang, Liaoning 110866, China.

[4]Vegetable Research Institute, Suzhou Academy of Agricultural Sciences; Suzhou, Jiangsu 215000, China.

[5]College of Horticulture, Vegetable Genetics and Breeding Laboratory, Anhui Agricultural University; Hefei, Anhui 230036, China.

[6]Guangzhou Academy of Agricultural and Rural Sciences; Guangzhou, Guangdong 510640, China.

[7]Shandong Degao Seed Co., Ltd.; Dezhou, Shandong 253011, China.

[8]Grandomics Biosciences Co., Ltd.; Wuhan, Hubei 430070, China.

[9]Schoole of Life Sciences, University of Warwick; Coventry CV4 7AL, UK.

[10]School of Life Sciences, Nantong University; Nantong, Jiangsu 226019, China.

[11]The Advanced Seed Institute, Plant Precision Breeding Academy, Zhejiang Provincial Key Laboratory of Crop Genetic Resources, College of Agriculture and Biotechnology, Zhejiang University; Hangzhou, Zhejiang 310058, China.

[12]Department of Plant Biotechnology and Bioinformatics, Ghent University; 9052 Gent, Belgium.

[13]VIB Center for Plant Systems Biology, VIB; 9052 Ghent, Belgium.

[14]Department of Biochemistry, Genetics and Microbiology, Centre for Microbial Ecology and Genomics, University of Pretoria; Pretoria 0028, South Africa.

[15]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University; Nanjing, Jiangsu 210095, China.

†These authors contributed equally to this work

*Corresponding author. Email: yyzjj@hebau.edu.cn (J.Z.); yuanyuxiang@hnagri.org.cn (Y.Y.); yg.hong@hebau.edu.cn (Y.H.); yvpee@psb.vib-ugent.be (Y.V.P.); mawei0720@163.com (W.M.)

Abstract: *Brassica rapa* (*Br*) encompasses many morphotypes and subspecies, representing a unique model to investigate plant diversification and subspeciation. Here, we re-sequence genomes of 1,720 *Br* accessions and de novo assembled 11 representative T2T gapless genomes for 7 elite subspecies that underwent intensive morphotypification and developed distinct agronomic traits valued to agriculture. We identify 6,992 unknown genes, 110 complete (peri)centromeres, and 5 new satellites associated with *Br* morphotype/subspecies and *Brassica* species evolution. The pangenome built on 11 gapless and 20 published genomes reveals structural variations and gene diversities among *Br* subspecies. Pangenome-wide association studies uncover *BrLH1* controls leaf-head formation. We show structural changes have occurred in satellites, (peri)centromeres, and genes contributing to fast subspeciation/morphotypification during the short-history of *Br* cultivation, providing invaluable resources for *Brassica* breeding.

Main Text:

The drivers leading to the rapid rise and diversification of angiosperms since the middle Cretaceous period remain to be elucidated (*1, 2*). However, recent advances in functional (pan)genomics have opened new avenues to study fast evolution and development in flowering plants. For instance, a telomere-to-telomere (T2T) genome assembly reveals that dynamic changes of centromeres may have engendered speciation and rapid post-speciation divergence in cotton (*3*). Genomic rearrangements and structural variations (SVs) uncovered from pangenome graphs may have also underlain phenotypic variation and contributed to species diversification in forest trees and important crops (*4-9*). However, the lack of high-quality (pan)genomic resources has restricted our ability to fully appreciate the role of centromere and SV complexities in genomic diversity (*10, 11*). This hinders a more complete understanding of the pangenomic links with phenotypic variations in plants. We require data from (sub)species that have evolved rich morphotypes within a certain period to uncover (pan)genomic basis of fast plant diversification.

Brassica rapa (*Br*, AA, 2n=20) is one of the ancestral diploid species in the "triangle of U" model of Brassica genomes and provides unique insights into the evolution of polyploid crops (*B. napus* AACC, and *B. juncea*, AABB) (*12, 13*). *Br* has also diversified into numerous subspecies with diverse morphologies and physiologies, heavily shaped by modern agricultural demands during a relatively short history of domestication. Under human selection since 3,000-3,500 BCE, *Br* has diversified into leafy, root and oilseed crops (*12, 13*). For instance, Chinese cabbage (*B. rapa* ssp. *pekinensis*) develops leafy heads, Pak choi (*B. rapa* ssp. *chinensis*) forms no head but has smooth, darker green leaves with a prominent white midrib, and dark-leafy Wutacai (*B. rapa* ssp. *narinosa*) exhibits an overall flat appearance. Caixin (*B. rapa* ssp. *parachinensis*) requires no vernalization for flowering, and possesses a short vegetative

to reproductive growth cycle. Mizuna (*B. rapa* var. *nipposinica*) possesses serrated leaves and a piquant flavor. Unlike these leafy *Br* vegetables, oilseed (*B. rapa* ssp. *oleifera*) develops extensive lateral branches and exhibits enhanced photosynthetic efficiency, seed yield and quality. Turnip (*B. rapa* ssp. *rapa*) is characterized by nutrition-rich swollen roots. Both Turnip and Mizuna are more ancient than Chinese cabbage, Taicai (*B. rapa* ssp. *chinensis* var. *tai-tsai*) and other *Br* subspecies/morphotypes (*10, 11*). Such phenotypic diversity makes *Br* and its subspecies an ideal model for exploring the pangenomic basis of fast evolution of morphotypes and subspecies in flowering plants.

In previous studies, a pangenome built on 18 chromosome-level (*11*) and two telomere to telomere (T2T) assemblies (*14*) and a recent near-complete genome assembly (*15*) have been established as critical references for studying *Br* evolution and domestication. However, given the huge diversity found in *Br*, these resources are still limited and likely do not adequately represent the genomic diversity among many different morphotypes and subspecies within *Br*. Here we re-sequenced the genomes of 1,720 different *Br* accessions and de novo assembled T2T gapless genomes for 11 representative accessions, encompassing various morphotypes and subspecies that serve as elite breeding parents. We constructed a high-quality centromere and SV map for *Br*, enabling the analysis of centromeric elements, SV diversity and their rapid evolution. We established the most comprehensive gapless pangenome to date, and elucidated the pangenomic basis for morphotype diversification and subspeciation in *Br*. Using this unprecedented pangenomic resource, we also identified and functionally validated *BrLH1* as a key player controlling leafy head development in Chinese cabbage and *B. rapa* ssp. Taken together, our findings shed light on the intricate patterns of genetic diversity versus rapid morphotypification and subspeciation in *Br*, and enhance our understanding of pangenomic and genetic contribution to fast diversification and evolution of flowering plants.

T2T genome assemblies for 11 representative *B. rapa* accessions

To construct a pangenome that can cover *Br* genetic diversities and population structures, we re-sequenced the genomes of 1,720 accessions at an average 48× depth. These accessions belong to different subspecies and morphotypes, including 978 Chinese cabbage, 4 Taicai, 202 Caixin, 292 Wutacai, 134 Pak choi, 28 Oilseed, 5 Mizuna, and 77 Turnip (Fig. 1A and table S1). Based on their phylogenetic relatedness, population structures, morphotypical diversities and agricultural importance as elite breeding parental germplasm (Fig. 1A and fig S1), we selected 11 accessions for de novo gapless genome assembly (Fig. 1B and table S2).

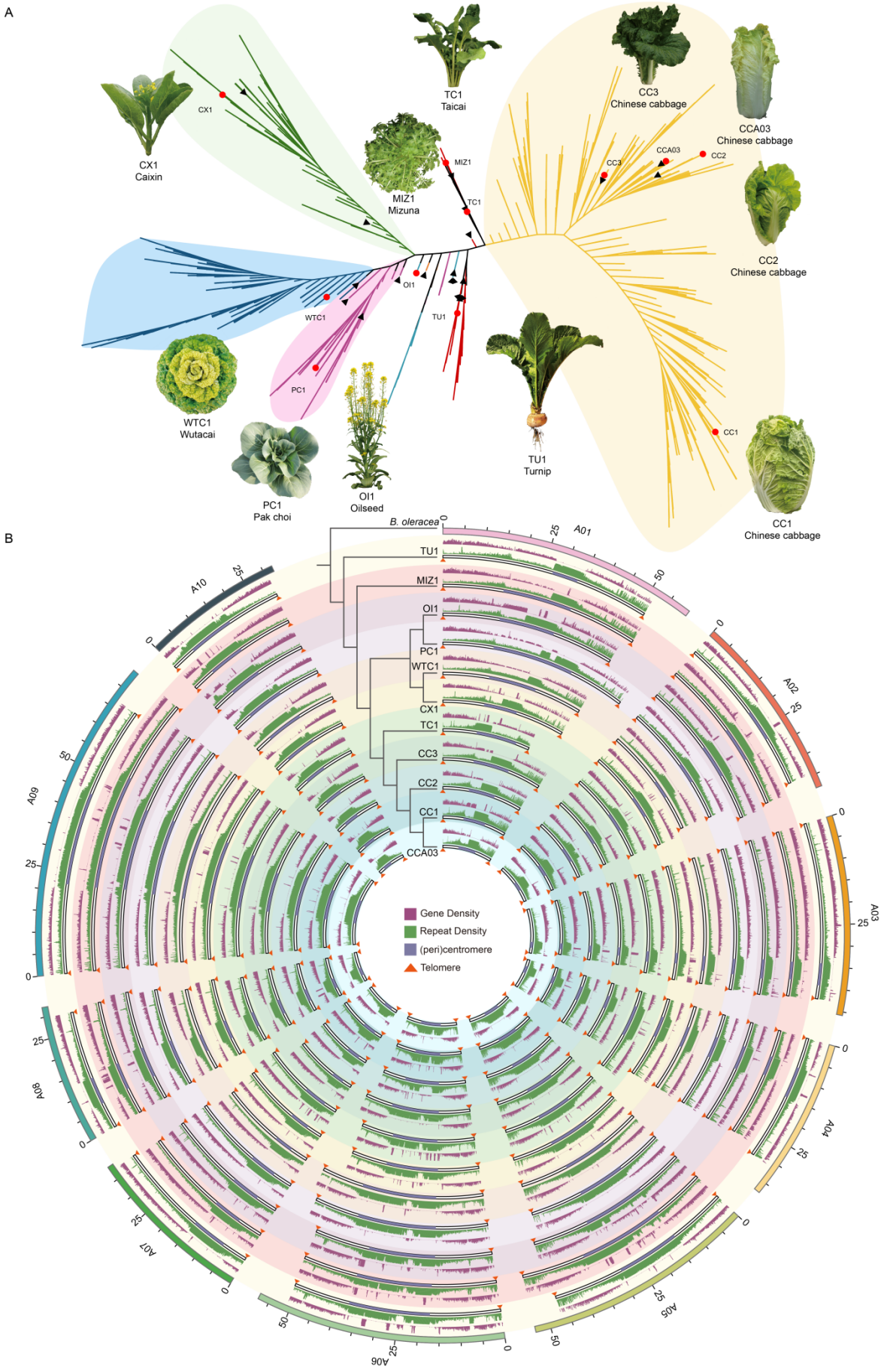

A
CX1
Caixin
TC1
Taicai
MIZ1
Mizuna
CC3
Chinese cabbage
CCA03
Chinese cabbage
CC2
Chinese cabbage
WTC1
Wutacai
PC1
Pak choi
OI1
Oilseed
TU1
Turnip
CC1
Chinese cabbage
B
B. oleracea
Gene Density
Repeat Density
(peri)centromere
Telomere

Fig. 1. Phylogenetics and gapless genome assemblies of *B. rapa* ssp.
(**A**) Phylogenetic tree of 1,720 accessions inferred from whole-genome SNPs. Branches are color-coded to indicate different accessions from different *B. rapa* subspecies and/or morphotype groups. Eleven accessions used for de novo gapless genome assembly are marked red-spot in the phylogenetic tree. Their common names, Chinese cabbage (CC1, CC2, CC3, and CCA03), Taicai (TC1), Caixin (CX1), Wutacai (WTC1), Pak choi (PC1), Oilseed (OI1), Mizuna (MIZ1) and Turnip (TU1), are indicated. Individual plants for each of the 11 accessions with distinctive phenotypes are shown near to each red spot. Black triangles denote 20 accessions whose genomes were previously sequenced. (**B**) Gapless genome assemblies of 11 representative accessions and their phylogenetic relatedness. Each arched panel stands for one of the 10 T2T-assembled chromosomes A01-A10 for each of the 11 *B. rapa* accession as indicated. Gene density (purple), repeat element density (green), (peri)centromere region (lavender), and telomeres (orange triangles) are indicated. *B. oleracea* was used as an outgroup to establish the root of the tree.

We constructed gapless T2T genomes for the 11 accessions using a comprehensive sequencing and assembly strategy. First, we generated high-coverage and quality sequencing data, including PacBio HiFi reads (average 62.6×), ONT ultralong reads (average 229.0×), Hi-C reads (average 264.3×), and NGS MGI-T7 paired-end reads (average 60.1×) (fig. S2 and table S3), and did preliminary genome assembly. We then performed de novo genome assembly (fig. S3), T2T genome assessment for each accession, and genomic collinearity analysis by aligning each assembly with its corresponding morphotype and subspecies (fig. S4) (*11, 14, 15*). We filled gaps (table S4 and S5), rectified assembly errors, and evaluated the copy number of satellite DNAs, especially the 45S rDNA (5.8S, 18S and 28S) to ensure our T2T gapless genomes being correctly assembled (fig. S5). To identify large SVs, we randomly selected and manually checked four inversions in each assembly. We also performed integrative genomics viewer (IGV) validation on the HiFi and ONT reads and confirmed the accuracy of our genome assemblies.

This approach enabled to assemble gapless genomes covering all telomeres and 110 complete centromeres for 11 *Br* accessions (Fig. 1B and table S6). Genome sizes of these assemblies ranged from 426.54 to 446.60 Mb, with an average contig N50 of 46.36Mb (Table 1 and table S7). Benchmarking Universal Single-Copy Orthologs (BUSCO) yielded an average score of 99.28%. High-quality values (QV) averaged 52.57, and exceeded the Vertebrate Genomes Project standard of QV40 (*16*). Long terminal repeat (LTR) assembly index (LAI) (*17*) approached or reached the ‘gold standard’ level (LAI > 20). Furthermore, MGI-T7 short reads (94.91%-98.61%), HiFi long reads (99.26%-99.97%), ultra-long reads (96.53%-99.39%) and Hi-C reads (99.89-100%) were all aligned to our gapless T2T genomes.

On average, 58.26% of the sequences were predicted as repeats, with LTR retrotransposons (LTR-RTs) being the most abundant, accounting for 23.16% to 36.15% of the whole genomes (table S8). Protein-coding genes were annotated using an integrated genome-guided transcript assembly, homologous-protein-based, and ab initio prediction pipeline (tables S7 and S9). After consolidation by Evidence Modeler (EVM), a total of 46,603 to 48,562 genes were annotated across each of the 11 genomes with an average gene length of 1,841 to 2,493-bp and 1,085 to 1,155-bases for the protein coding sequences (CDS) (Table 1 and table S7). Each assembly exhibited an average gene BUSCO score of 98.80%. We identified an average of 636 new genes in each genome and 6,992 in total, all of which were supported by databases (Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, Non-Redundant Protein Sequence Database, Swiss-Prot Protein Sequence Database) (tables S10 and S11) although they were not annotated in previously published genomes. We identified genome-wide non-coding RNAs including an average of 9,886 rRNAs, 1,574 small RNAs and 1,304 tRNAs for each of the 11 accessions (table S7). Taken together, these analyses demonstrate that we have produced the most extensive and comprehensive

high-quality *Br* genome sequences to date, and that the predicted gene models are of high quality and suitable for downstream analyses.

Table 1. Quality assessments and annotation of 11 assemblies

Accession	Total length (Mb)	Gaps	No. of telomeres	Assembly BUSCOs (%)	QV	LAI	Repeat elements (%)	No. of genes
TU1	427.71	0	20	99.60	51.82	19.21	56.88	48,562
MIZ1	444.51	0	20	99.32	53.25	17.32	59.17	46,603
OI1	437.06	0	20	99.44	51.92	23.78	58.16	46,770
PC1	446.60	0	20	99.44	51.72	23.03	58.51	47,840
WTC1	435.86	0	20	99.38	52.50	18.57	59.31	47,143
CX1	434.04	0	20	99.38	53.32	22.89	57.01	47,599
TC1	427.22	0	20	97.77	51.80	22.46	59.02	47,706
CC3	426.54	0	20	99.44	54.36	21.93	57.97	47,555
CC2	430.93	0	20	99.38	52.49	22.89	58.25	47,292
CC1	426.89	0	20	99.38	51.81	22.45	58.50	46,809
CCA03	434.60	0	20	99.50	53.30	21.17	58.09	47,102

Centromere evolution and diversification in *B. rapa*

We defined chromosomal sections that bind to characteristic centromeric histone H3 (CENH3) as centromeres in the genomes of *Br* subspecies and morphotypes (*18, 19*). Structural and sequence-based evidence such as high satellite density, high repeat density and/or low gene density was also used to reaffirm our complete assembly of these centromeric regions. To experimentally identify centromeres and to investigate the genomic landscape of centromeres, we performed chromatin immunoprecipitation with sequencing (ChIP-seq) using the centromere-specific antibody raised against the conserved BrCENH3 dodecapeptide (KHFASRARDRNP) (fig. S6 to S8 and table S12 to S14) (*19*). This analysis led to identifying a single distinct CENH3-rich region in each of the 10 chromosomes across each of 11 *Br* accessions (Fig. 2A and fig. S9). These centromeres varied in size, ranging from 0.58 to 2.20 Mb, with an average of 1.16 Mb (table S15). *Br* centromeres are characterized by long arrays of 176-bp satellite repeats (CentBr1 and CentBr2) (*20*). CentBr were found to span over the entire centromeric regions (Fig. 2A), validating the completeness and positional accuracy of the assembled centromeres. Consistent with previous findings (*15*), CentBr2 was specifically distributed in chromosomes A03 and A05, accounting for 25.0 and 38.6% of the centromeric regions, respectively, while CentBr1 was ubiquitous across the centromeres of the other eight chromosomes, contributing 38.2-69.4% of the centromeric regions (fig. S9). We noted that CentBr were not exclusively localized to centromeres but were also present extensively in the pericentromeric regions, suggesting that only partial CentBr arrays are located within the functional domains of *Br* centromeres. In addition, Ale/Copia and CRM/Gypsy retrotransposons accounted for 1.64-38.8% of each centromere region. The centromeric regions contained very few genes. 262 (106 non-TE and 156 TE) genes were detected among the 110 centromeres (table S16). Using RNA-seq data on leaf, stem, root, bud and silique at seedling (leaf, stem and root), mature plant (leaf and stem), flowering (flowering bud), and seeding (silique) stages of the 11 *Br* subspecies/morphotypes, 107 (51 non-TE and 56 TE) genes were expressed in at least one tissue. 43/51 expressed non-TE genes and 51/56 expressed TE genes bound to CENH3. However, the mRNA level of 69/94 expressed

TE and non-TE genes that bound to CENH3 was low (0<FPKM<0.1), suggesting that transcriptional activity is incompatible with CENH3 binding (table S16). Expressed vs non-expressed genes of the total centromeric genes in *Br* centromeres were similar to those in rice (*21*) and cotton (*22*), but some different from maize (*23*).

To further characterize highly repeated sequences, we performed de novo satellite identification across (peri)centromeres. Five novel satellites, namely pBrSTR497, pBrSTR92, pBrSTR144, pBrSTR197, and pBrSTR228, along with a previously reported PCR630 satellite (*15*) were identified. The tandem organization of these satellites was confirmed by HiFi long-read sequencing data (fig. S9, tables S17 and S18). However, these satellites localized exclusively on (peri)centromeres with a sharp decrease in gene density but a dramatic increase in TE density. A pericentromeric region refers to the chromosomal sections that surround the unique CENH3-binding centromeres. Unlike chromosomal arms, pericentromeric regions are characteristic of high satellite arrays (*24*), dense repeat sequences and low gene density (*8, 25, 26*). We define regions that meet at least two of the three criteria as pericentromere. Compared to the rest of chromosomal arms, those regions displayed distinct DNA methylation profiles, TADs, and AB compartments (figs. S10 and S11), highlighting the unique chromatin identity of the pericentromeric regions, spanning 0.60-25.56 Mb adjacent to defined centromeres (fig. S9 and table S15). We also detected 23,042 (1,088 TE and 21,954 non-TE) genes in the pericentromeres of the 11 *Br* genomes. Of these genes, 7,961 (454 TE and 7,507 non-TE) were expressed in at least one tissue. In general, transcript levels of genes in (peri)centromeric regions were significantly reduced compared to genes in chromosome arms (table S16 and table S19). Pericentromeres vs centromeres also had 10.6% vs 20.6% LTR content but only 3.3% LTR was found in chromosome arms (fig. S12A). The most abundant LTRs in pericentromeres were Ale/Copia, Bianca/Copia, CMR/Gypsy and Tekay/Gypsy, while Ale/Copia and CRM/Gypsy were predominant in centromeres (fig. S12, B and C). The LTR insertions were significantly younger in centromeres than chromosomal arms (fig. S13). These distinctive features of pericentromeric vs centromeric regions suggest their potential functional significance in chromosomal organization and genome stability maintenance (*18*) in *Br*.

Sequence analyses indicate that pBrSTR92 duplicated to form pBrSTR144 and pBrSTR197, but triplicated to generate pBrSTR228 (Fig. 2B). Five basic units A to E were identified, and an arrangement of these units ABCD, ABDAECD, ABCDABCD, and ABCDAECDABCD gave rise to pBrSTR92, pBrSTR144, pBrSTR197, and pBrSTR228, respectively (Fig. 2B and fig. S14). Continuous multiple arrays for each satellite repeat were identified by HiFi long reads (table S18). Copy number variations of pBrSTR497, pBrSTR92, pBrSTR144, pBrSTR197, and pBrSTR228 were significantly more pronounced than those of CentBr and PCR630 across different *Br* morphotypes and subspecies (fig. S15 and table S20). pBrSTR144, pBrSTR197, and pBrSTR228 were almost absent in Chinese cabbage (Fig. 2C). We interpret these findings to mean that dynamic contraction and expansion of rapidly evolving satellites may have driven centromere evolution, which is associated with the differentiation of *Br* morphotypes and subspeciation since morphotype diversification is mainly caused by functional gene variation and fixation (in a subpopulation), due to the fact that very few genes were identified in the centromeric regions (table S16).

To trace the evolutionary footprints of satellites during speciation within the *Brassica* genus, we selected 3 diploids (AA, BB, CC) (*27-31*) and 3 allotetraploids (AABB, AACC, BBCC) (*31-38*) to investigate satellite formation and accumulation

patterns (Fig. 2D and table S20). *Arabidopsis thaliana* (*39*) was included as a control due to its close relationship with Brassica. None of the seven centromeric/pericentromeric satellites was detected in the ancient *B. nigra* (BB) and *Arabidopsis*; and only CentBr and PCR630 were present in *B. oleracea* (CC; Fig. 2D and fig. S16). However, *Br* (AA) has evolved to encode all seven satellites, with CentBr and PCR630 showing approximately 3.31 and 5.81-fold copy number expansion, respectively, compared to their CC counterparts (Fig. 2D and fig S17). These satellites were preferentially retained in allotetraploids when present in at least one parental lineage (Fig. 2D and table S20), highlighting their CC to AA evolutionary route in the 6 *Brassica* species.

Using the collinearity analyses, we found that most centromeres have been highly dynamic and evolved rapidly across the 11 assembled accessions, with the centromere on chromosome A01 showing particularly marked changes (fig. S18) and the centromere on A10 remaining very conserved among accessions (fig. S19). On chromosome A10, CENH3 binding regions are conserved with approximate 80.0% sequence identities among 11 assemblies. We observed significant overlap between (peri)centromere repositioning regions and different satellite boundary regions (fig. S20). Within the 8.91Mb-12.96Mb (peri)centromeric region of A10, 25 non-TE genes on average were identified in each accession, some of which are involved in essential processes such as plant survival, growth, and development (table S21) (*40*-*52*). To further validate our findings, we conducted virus-induced gene silencing (VIGS) to knock down two genes *BrNfu2* and *BrQS* in Caixin CX1. VIGS of *BrNfu2* reduced CX1 growth (fig. S21, A to D). This is supported by that *BrNfu2* is an orthologous gene of *AtNfu2* in *Arabidopsis*, and *AtNfu2* mutants exhibit dwarf phenotypes due to impaired photosynthetic efficiency and metabolic processes (*43*). The *Arabidopsis* homolog of *BrQS* encodes quinolinate synthase and its dysfunction leads to early flowering (*44*). As expected, suppression of *BrQS* by VIGS caused CX1 plants to flower much earlier than control plants (fig. S21, A and E). These findings suggest that the evolutionary conservation of the A10 centromere structure is likely driven by its critical genomic functions in regulating essential developmental processes in *Br*.

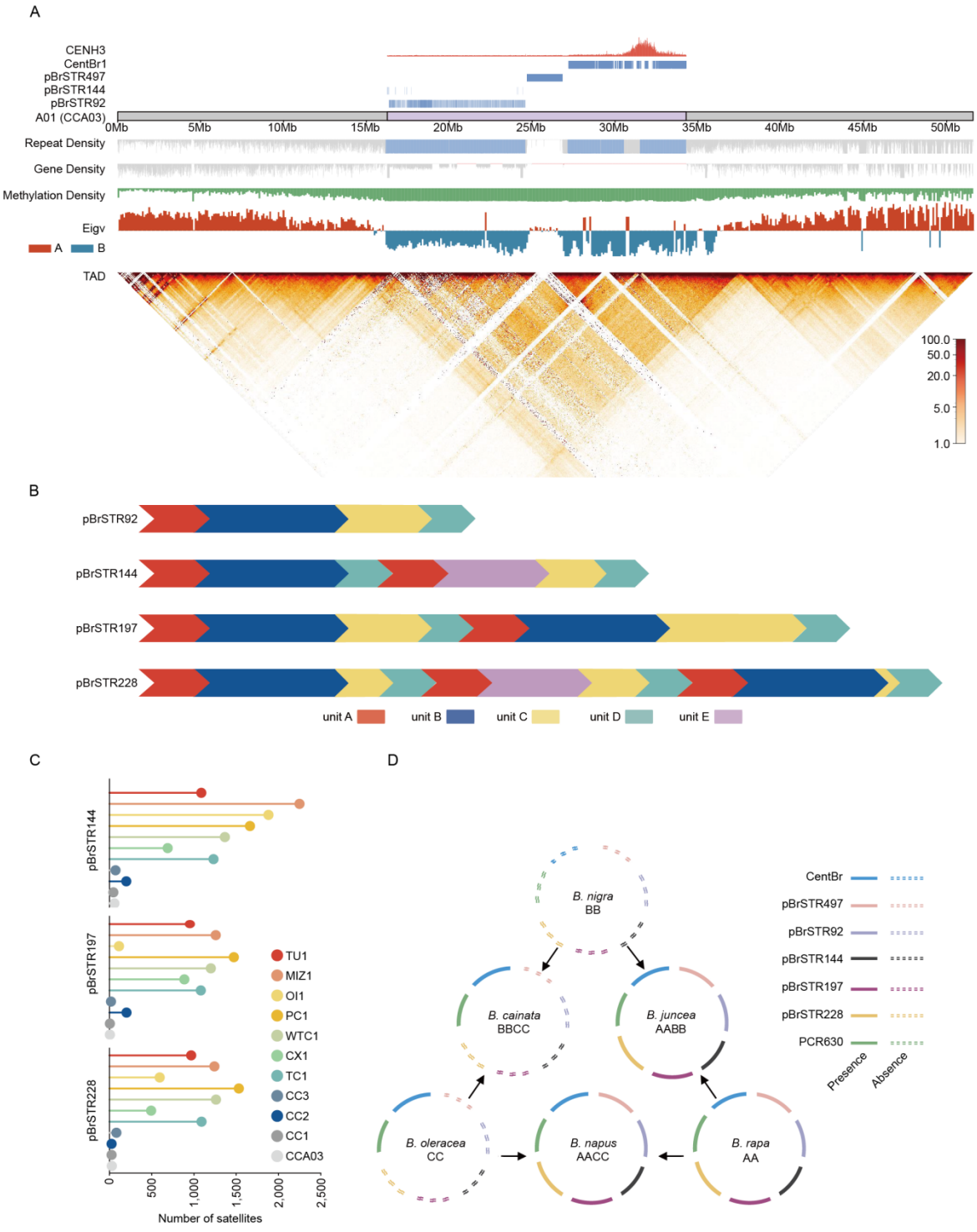


Fig. 2. Centromeres and satellites in *B. rapa* ssp. and *Brassica* sp.
(**A**) Diagram of centromere in CCA03 chromosome A01. Panels from top to bottom show CENH3 enrichment, satellite density, chromosome, repeat density, gene density, DNA methylation density, A/B compartment and topologically associating domain (TAD) as indicated. CENH3 (known as CENP-A in mammals) is a characteristic of eukaryotic centromeres which include CENH3-enriched core regions and interspersed nucleosome subdomains (*16, 17*). The CENH3-enriched region with strong signals (red peaks along the CENH3 panel) is defined as centromere. Proximal regions, in which satellites are also enriched, upstream and downstream of the centromere are regarded as pericentromere (purple box on A01). Eigenvector values of correlation matrix (Eigv) are indicated. The color scale shows the Pearson's correlation co-efficiency of normalized interaction matrix. (**B**) Satellites with specific units A-E shown by different colors. Units A, B, D and E are of a unique length of 17, 38, 14 and 28 nucleotides, respectively (fig. S14). Sequences of these four units are conserved among pBrSTR92, pBrSTR144, pBrSTR197 and pBrSTR228. The nucleotide length of Unit C ranges from 4 to 23 and varies even within same

individual satellites (e.g. pBrSTR197 and pBrSTR228) (fig. S14). (**C**) Satellite Numbers in different *B. rapa* accessions (see also fig. S15). (**D**) Evolutionary footprints of satellites in *Brassica* species. The 7 Satellites are shown by different colors as indicated. Solid and dotted lines show presence vs absence of these satellites in the given *Brassica* species.

A pangenomic insight into *B. rapa* diversification

We constructed a protein-coding gene-based *Br* pangenome using our assembled 11 gapless T2T and 20 previously published assemblies (*11, 14, 15, 53*). Through OrthoFinder, 1,439,526 genes were clustered into 47,946 non-redundant pan-gene families (Fig. 3). Simulation analyses of the 31 randomized *Br* accessions indicated that gene family numbers in this newly constructed pangenome reached saturation (Fig. 3A). Based on gene distributions across the 31 assemblies, we identified 15,558 core gene families (32.4%) present in all 31 accessions, and 32,388 (67.6%) as variable gene families absent in at least one accession. These variable gene families were further categorized into 8,536 (17.8%) softcore gene families present in 28 to 30 accessions, 1,512 (3.2%) specific gene families present only in one accession, and 22,340 (46.6%) dispensable gene families in 2 to 27 accessions (Fig. 3B-D). On average, each individual assembly comprised of 49.7%, 25.9%, 24.3%, and 0.2% core, softcore, dispensable and specific gene families, respectively. The number and length of CDSs of core and softcore gene were significantly greater than those of dispensable and specific genes (Fig. 3E and fig. S22). Furthermore, the non-synonymous to synonymous ratio (*Ka*/*Ks*) increased from core to specific genes (Fig. 3F), suggesting that these genes may be less constrained against functional change. Collectively, these results suggest that dispensable and specific genes have endured faster evolution rates and may possess functions associated with agronomic traits in morphotype diversification and subspeciation.

We then focused on subspecies-specific genes including 98 in Turnip (characterized by swollen roots), 62 in Caixin (not requiring vernalization), and 697 in Chinese cabbage (with leafy heading) (Fig. 3G), that may be associated with *Br* morphotypification and subspeciation. Among these genes, 35, 14 and 152 were not found in other *Brassica* species and *Arabidopsis*. GO-enrichment suggests some genes may have played a key role in *Br* diversification (Fig. 3G and fig. S23). For example, *PP2A-3* (*54*) and *CYP735A2* (*55*) in Turnip have been reported to be closely related to root development. In Caixin, *NRPD1* (*56*) is essential for flowering. *AXR1* and *DD46*, known to regulate leaf axial growth (*57*) and leaf development (*58*) in *Arabidopsis*, may influence leafy heading in Chinese cabbage.

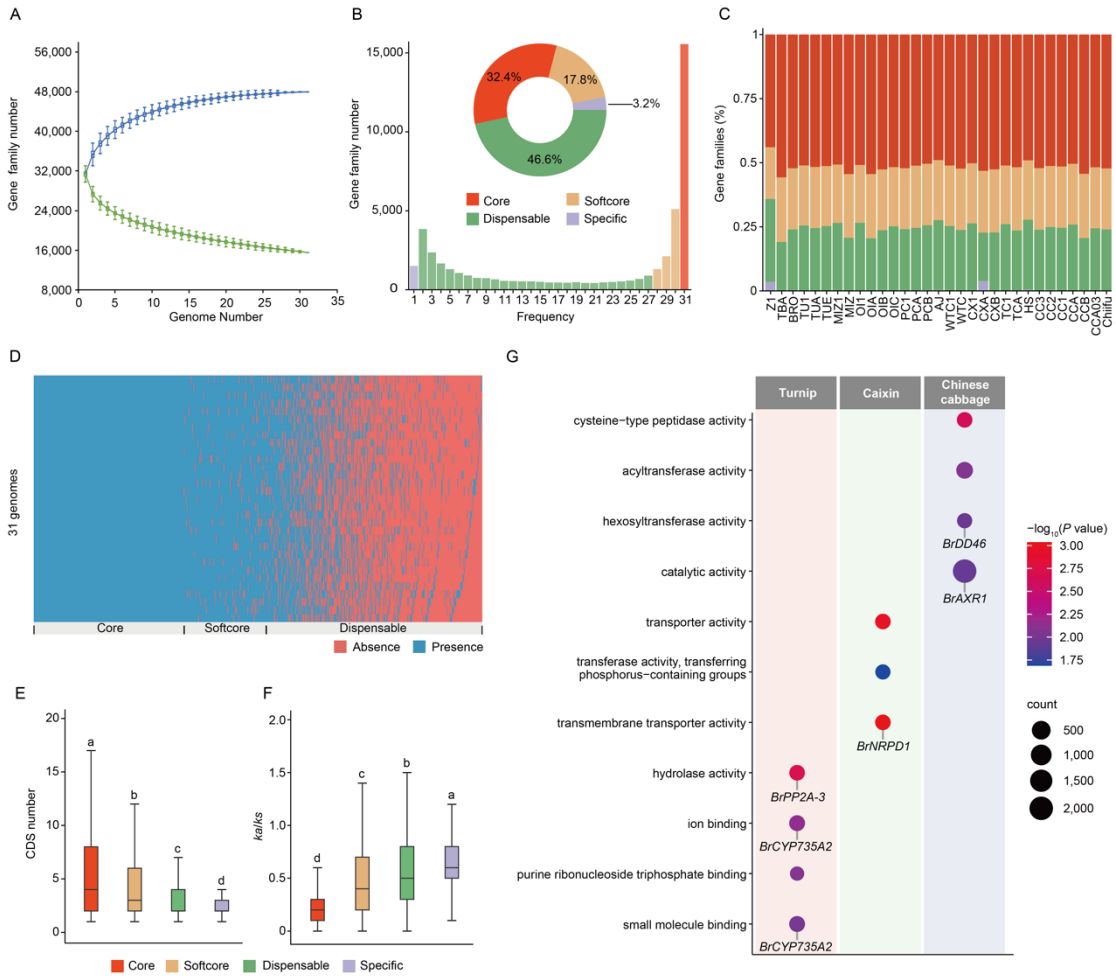


Fig. 3. The pan-genome constructed from 31 *B. rapa* assemblies.
(**A**) Number of pan- and core gene families in the 31 *B. rapa* genomes. (**B**) Compositions of the *B. rapa* pangenome. The histogram shows the number of gene families in the 31 genomes with different frequencies. The pie chart shows the proportion of the gene families marked by each composition. (**C**) Percentage of core, softcore, dispensable and specific gene families per genome assembly. Color codes are the same as indicated in (B). (**D**) Mapping the landscape of presence-absence variations (PAVs) across non-redundant gene families for the 31 *B. rapa* assemblies. (**E**) Boxplots of CDS number. Y-axis: CDS number of genes in core, softcore, dispensable and specific gene families. Different lowercase letters above the box-plots represent significant differences ($P < 0.05$). (**F**) Boxplots of *Ka/Ks* value. Y-axis: *Ka/Ks* value of genes in core, softcore, dispensable and specific gene families. Different lowercase letters above the box plots represent significant differences ($P < 0.05$). (**G**) Functional analysis (gene ontology) of specific genes on three morphotypes (subspecies), Turnip, Caixin and Chinese cabbage. Five representative known genes associated with different morphotype-specific phenotypes belonging to different terms are shown.

Linking SVs with subspeciation and diversification in *B. rapa*

To uncover SVs, we constructed a graph-based *Br* pangenome by aligning the genome reads of 30 accessions to the CCA03 reference genome. In total, 276,700 SVs (>50 bp) were identified, including 209,969 Presence-Absence Variations (PAVs), 1,458 inversions, 1,448 duplications and 63,825 translocations (Fig. 4A and fig. S24). We verified 10 randomly selected in silico SVs on 24 accessions from Chinese cabbage, Caixin, and Turnip (fig. S25), and validated the reliability of the SVs identified using computational methods (Fig. 4A and fig. S24). Among all the SVs, 95 were present in 30 accessions, 1,611 in 26-29 accessions, 149,816 in 2-25 accessions, and 123,987 unique in single accession; revealing most SVs exist only in one or less than 4 accessions (Fig. 4B). SVs, identified via read-based calling pipelines, were concentrated in chromosome arms (fig. S24) and enriched in intergenic regions, but low in exons (Fig. 4C), consistent with previous reports (*7, 9, 30, 59-61*). Population

genetics analyses show SVs had a strong contribution to morphotype diversification (fig. S26), and gene expression levels decreased as the distance from SV(s) increased (fig. S27). Additionally, 3,133 PAVs were identified among 110 centromeres, and 17 PAVs of centromere-associated genes were found, which might be linked with *Br* subspeciation/morphotypification. We also found that 41.07% PAVs overlapped with TEs, and LTR-TE PAVs located the longest distance away from protein-coding genes, followed by DNA-TE PAVs and those without TEs (Fig. 4D). Such close TE-SV correlations indicate that TEs may have been critical to drive SVs in *Br*. Moreover, we used the ancient subspecies Turnip as an outgroup to deduce SV origins, and SVs state that is the same with Turnip is defined as ancient, whilst different as derived. Across 31 accessions, we identified 33,812 to 59,038 derived SVs, and 6,225 to 17,358 derived subspecies-specific SVs, further supporting that SVs have contributed to *Br* morphotypification/subspeciation (Fig. 4E).

The pan-SV graph enables detection of SVs using NGS short reads at the population level, and advances population genomics studies. We constructed a graph-based *Br* pangenome by integrating all identified SVs into the CCA03 reference genome using the vg toolkit based on genomic coordinates (Fig. 4F). We then mapped the clean short reads of 1,720 accessions to the pan-SV graph, and investigated if these SVs could be linked to phenotypic variations. We found 200, 203 and 503 PAVs are potentially relevant (80% conformity) to diversification of Chinese cabbage, Caixin and Turnip, respectively (fig. S28, A to H and table S22). For instance, an 8,653bp PAV on chromosome A04 resulted in the presence-and-absence of an RZ-1B encoding gene which regulates *Arabidopsis* leaf axial growth (*62*). We found that accessions with or without such PAV had higher tendency to heading vs non-heading, respectively (Fig. 4G), indicating that this PAV might be one of the genetic variations responsible for leafy head formation in Chinese cabbage. We also found a PAV encompassing *BrWIP4* (fig. S28A), a root development gene homologue (*63*), in Turnip, and PAVs with *BrTFL1* (fig. S28H) or *BrFCA* (fig. S28F) in Caixin, which known to control vernalization and flowering time (*64, 65*). Presence of the 5,168-bp PAV containing *BrWIP4* plus up/downstream regulatory elements (fig. S28A) is likely to bring new function to control turnip root development. Regarding *BrTFL1* (fig. S28H) or *BrFCA* (fig. S28F) in Caixin, how presence of 191-bp or 1,298-bp PAV affects *BrTFL1* or *BrFCA* to control vernalization/flowering remains unclear since both PAVs were mapped to the downstream of the two genes. Various mechanisms could be envisaged, for instance, those PAVs might act as distal regulatory elements to affect *BrTFL1*/*BrFCA* expression to impact their biological functions. Deep mining pan-SVs further revealed several other SVs that may also influence root and vernalization in *B. rapa* ssp. (fig. S28, B to E, and G).

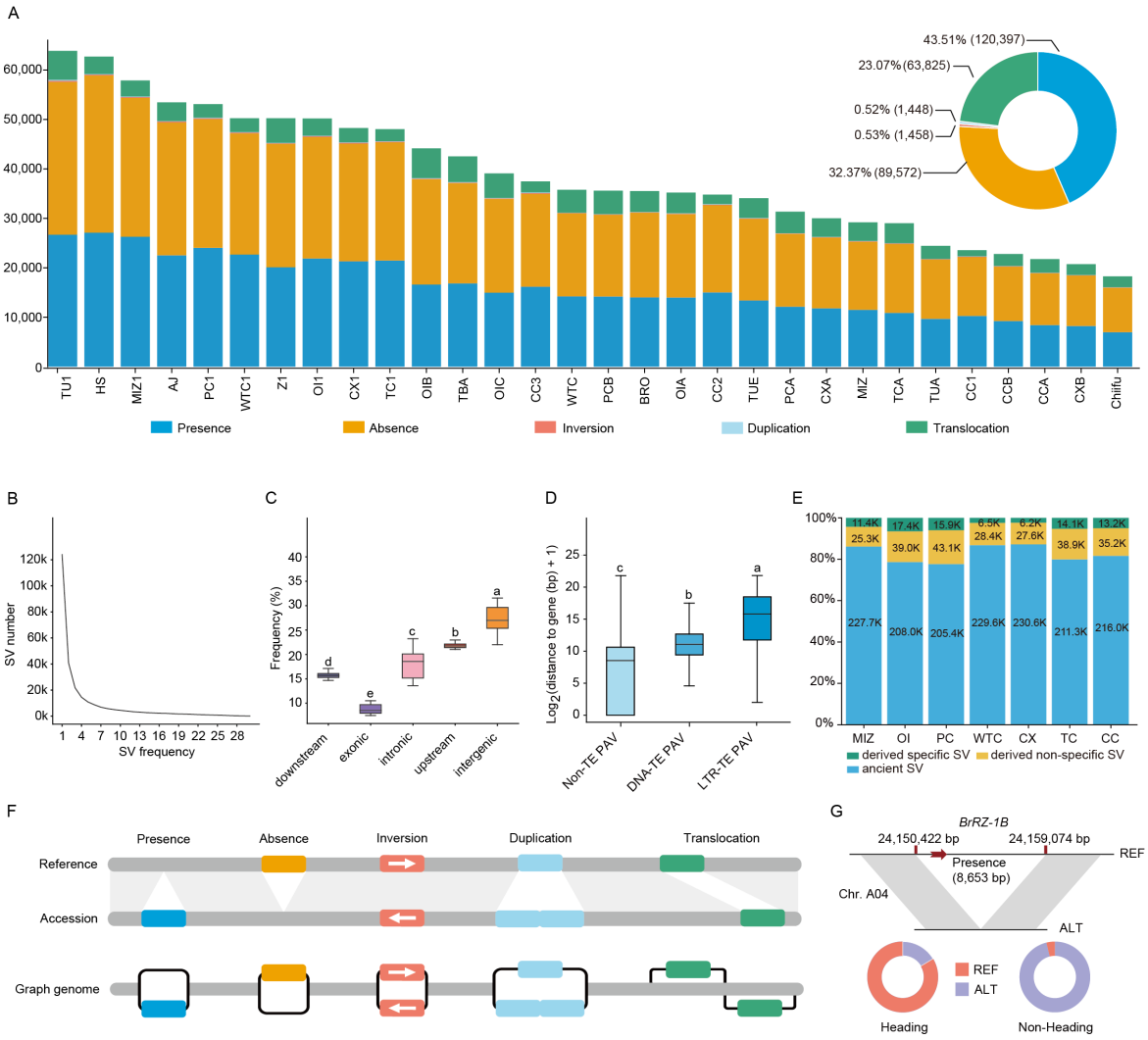


Fig. 4. The landscape of the genetic SVs among 31 *B. rapa* genome assemblies.

(**A**) The number of different types of SVs in each *B. rapa* genome. (**B**) The frequency of SVs in *B. rapa* genome assemblies. (**C**) The frequency distribution of SVs in different gene regions, including downstream, exon, intron, upstream, and intergenic regions. Different lowercase letters above the box plots represent significant differences between pairwise comparisons ($P < 0.05$). (**D**) Distance of PAVs to their closest protein-coding genes. A two-sided Wilcoxon test was used to determine the significant levels. Different lowercase letters above the box plots represent significant differences between pairwise comparisons ($P < 0.05$). (**E**) Characterization of ancient and modern SVs across *B. rapa* ssp. The ancient subspecies Turnip was utilized as an outgroup to deduce the origins of SVs across different *B. rapa* subspecies Chinese cabbage (CC), Taicai (TC), Caixin (CX), Wutacai (WTC), Pak choi (PC), Oilseed (OI), and Mizuna (MIZ). The numbers of SV count in thousands (k) on the bar chart columns and proportion of ancient SV, derived subspecies non-specific SV and derived subspecies-specific SVs in different *B. rapa* morphotypes are shown. (**F**) Schematic representation of SVs originating from the genomes of two accessions, alongside the linear reference genome sequences utilized for developing the graph-based pangenome. Types of SVs include PAVs (Presence or Absence), inversion, duplication, and translocation. (**G**) The PAV results in the presence-and-absence of *RZ-1B* (red arrow), and the comparison of leafy heading ability between two haplotypes of the PAV. REF represents the same sequence as the reference genome CCA03, and ALT indicates that the sequence is different from REF. Gray boxes represent the genomic collinearity region.

BrLH1 controls leafy heading in Chinese cabbage

Leafy head is one of the most important traits of Chinese cabbage (*66*). We performed pan-GWAS with SNPs, InDels (i.e. insertion-deletion < 50-bp), and PAVs on leafy heading ability in 1,720 accessions, and identified a genomic region associated with head formation with an extremely strong peak on chromosome A07. Within this region, a single copy intron-free candidate gene *BraA07g02453.V2*, namely *BrLH1* comprising

3,147-bp and encoding a 1,048-amino acid (AA) protein, was identified (Fig. 5A and fig. S29). Three non-synonymous SNPs at nucleotide 539, 556, and 658 in *BrLH1* (fig. S30A) were associated with heading formation. Among 1,720 accessions, plants carrying the AAC haplotype formed leafy heads, whereas those with alternative haplotypes (predominantly CTC and CTT) displayed flat leaves, demonstrating their mutation effects on *Br* heading and non-heading (Fig. 5B, fig. S29 and S30A).

We isolated a *BrLH1* loss-of-function mutant *fg7* (fig. S31) from an EMS-mutagenized CCA03 population (*67, 68*). *fg7* plants were dwarf and unable to form heads, while leafy heads developed properly in wild-type CCA03 (Fig. 5C). We then performed DNA bulked segregant analysis and confirmed loss-of-function *BrLH1* in *fg7* (Fig. 5D and fig. S30, B and C). After SNPs-index (≥0.9) filtration of MutMap and Kompetitive Allele Specific PCR genotyping, we identified a C-to-T SNP within the *BrLH1* exon on chromosome A07 (fig. S30B and tables S23 to S25). The mutated *BrLH1* carries a premature termination codon TAG at nucleotide 1,264 and encodes truncated BrLH1 of 421-AA (fig. S30, B and C). This SNP completely co-segregated with the non-heading phenotype in 275 F_2 individuals (tables S23 to S25). Collectively, these findings provide compelling pangenomic and genetic evidence for *BrLH1* as the causal morphogenetic regulator to control leafy heading in *Br*. BrLH1 is a Multiple C2 Domains and Transmembrane domains protein. Its *Arabidopsis* homolog QUIRKY interacts with the STRUBBELIG (SUB) receptor-like kinase and SCRAMBLED to controls the CAPRICE protein movement for regulating plant cell morphogenesis and patterning (*69-72*). We also found that BrLH1 interacts with BrSUB (fig. S32). Thus, like QUIRKY, BrLH1 may involve BrSUB to modulate heading in *Br*.

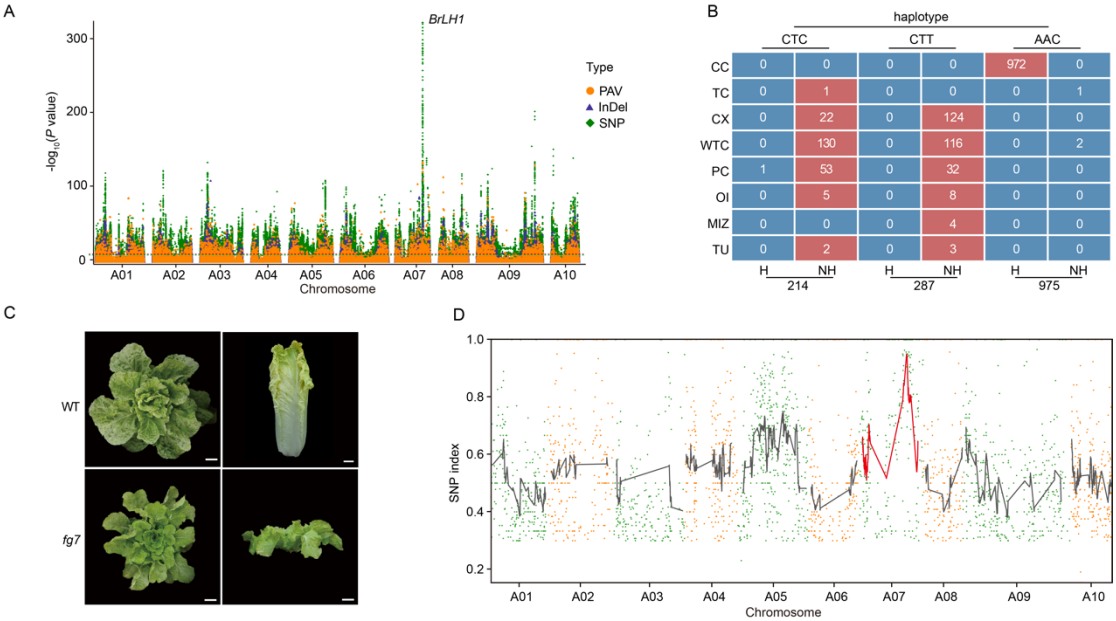


Fig. 5. *BrLH1* controls leafy heading in Chinese cabbage

(**A**) Manhattan plots and haplotype analysis of the leafy head pan-GWAS. Notable association locus with *BrLH1* is indicated. The dotted line represents the pan-GWAS significance threshold set at 0.05/total number of each type of variations. (**B**) Distributions of the *BrLH1* haplotypes in morphotype populations of *B. rapa* ssp. H: heading, NH: non-heading. (**C**) Phenotypes of WT vs mutant *fg7* Chinese cabbage at the heading stages. Plants were photographed at 80 days post seed sowing. Bar=5 cm. (**D**) Identification of genomic regions harboring causal mutations in chromosomes A01-A10 for *fg7* using MutMap. SNP index plots. Each symbol corresponds to an SNP, and the black regression line shows the average value of the SNP index based on a

sliding window analysis. Along with the red regression line, a peak appears on chromosome A07 and it indicates the location of causal mutation(s) for *fg7*.

DISCUSSION

In this study, we re-sequenced the genomes of 1,720 accessions, de novo assembled 11 T2T genomes, and constructed a gapless pangenome for *Br*. Bioinformatics mining of such unprecedented genomic and pangenomic datasets uncovers 110 complete (peri)centromeres, 6,992 novel protein-coding genes which are absent from all published genomes, over 1,500 accession-specific gene families, dozens to hundreds of subspecies-specific genes, 123,987 accession-unique SVs, and five previously unknown (peri)centromere-localized satellites in *Br*. The comprehensive (pan)genomic resources capture a significant degree of genetic diversity, particularly in (peri)centromeric regions. Our 'gold standard' reference gapless genomes and pangenomes to date facilitate mapping and linking mutations within and/or outside (peri)centromere regions to diverse phenotypic variations across *Br* morphotypes/subspecies. Indeed, functional (pan)genomic studies coupled with forward genetic analyses (*66-68*) have revealed *BrLH1* as the main causal morphogenetic factor to control leafy heading vs non-heading in Chinese cabbage and other *Br* subspecies.

Centromeres and centromere-specific satellites in genome dynamics, *Brassica* speciation and *B. rapa* subspeciation

Sequencing and characterizing the 110 complete centromeres led to identification of five novel centromere-specific satellites. We demonstrated that contraction and expansion of these satellites reflect the centromeric-to-genomic diversity of *B. rapa* ssp. and morphotypes, and of other species in the *Brassica* genus. The *B. rapa* ssp. are associated with rapid turnover of centromeric satellites, as has been observed in *Arabidopsis* speciation (*73*) and genomic structural diversification underlying morphotypes. Moreover, the unusual conservation of the A10 centromere juxtaposed with hyperdynamic conversation of A01-A09 centromeres suggests that "modular evolution" of functional centromeres may have occurred during *Br* subspeciation. In dynamic (peri)centromeres, satellite and basic satellite unit reorganizations facilitate chromosomal rearrangements and novel allele generation, while conserved (peri)centromeres (A10) are constrained by strong purifying selection due to their association with essential developmental genes (e.g., *BrNfu2* and *BrQS*). We also observed overlaps co-exist between different satellite boundaries and centromere repositioning regions (fig. S20). This implies that during *B. rapa* ssp. diversification, centromere formation and rapid turnover may occur through satellite rearrangements. Such "modular" rearrangements may have facilitated centromere evolution while avoiding overall genome instability that could be caused by random genome breakage, an intriguing phenomenon which has been reported in fruit fly, humans and maize (*74*), associated with *Br* morphotypification and subspeciation.

SVs in fast morphotype diversification and subspeciation during *B. rapa* subspeciation

Construction of the graph-based *Br* pangenome provides an unprecedented platform for systematic characterization and dissection of functional SVs and gene families. The *Br* morphotype/subspecies-specific SVs and gene families identified in our pangenome could have acted as crucial drivers during phenotypic divergence. They have been likely to play some essential roles in *Br* morphotype diversification and subspeciation during the process of their evolution. Indeed, the pan-SV analyses unveiled an 8,653-bp PAV

on chromosome A04, which encodes the key regulatory factor RZ-1B associated with leaf curling. This PAV is closely linked to differentiation of leafy heading ability in Chinese cabbage. Meanwhile, we also found some candidate SVs related to swollen root and vernalization flowering (fig. S28), such as a 1,130-bp PAV on chromosome A09, which encodes *LBD31* associated with flower development (fig. S28E) (*75*). This PAV is closely linked to non-vernalization flowering in Caixin. We propose a tripartite evolutionary model where SVs and morphotype/subspecies-specific genes function collectively to drive *Br* diversification, which usually is accompanied by modular centromere dynamics. This model proposes a possible mechanism in which SVs (fig. S28) (*76, 77*) and specific genes (Fig. 3G, fig. S23) would generate the primary genetic bases for diversification, while centromere modularity (fig. S20) balances genomic innovation with developmental robustness to ensure successful subspeciation as part of environmental and agricultural selection.

B. rapa subspeciation, a case of fast flowering plant diversification

Like other species in the *Brassica* genus (*27-38*), *Br* separated from their common ancestor with *Arabidopsis circa* 23 million years ago (*78*). This has diversified into numerous subspecies with varied morphologies and physiologies during post domestication evolution and artificial selection since approximately 3500 BCE (*12, 13*). This makes *Br* and its subspecies ideal to address whether and how genomic and dynamic genome and pangenome (*12-14*) could influence fast expansion and diversification of flowering plants, one of central aspects of the Darwin's abominable mystery (*1-3*). Darwin's abominable mystery refers to the rapid rise and diversification of flowering plants since the middle Cretaceous period (*1, 2, 79*). Many including Darwin have studied and tried to identify causal explanation for this puzzling phenomenon in flowering plant evolution. The central issue associated with the mystery stands on two aspects, i.e., the "sudden" evolution of flowering plants, and their subsequent fast expansions to ecological dominance on earth. The fossil record, systems classification and molecular phylogenetics have brought about insight into the origin and lineage of flowering plants. However, the drivers leading to fast diversification of these plant species within their evolutionary history remains deeply mystical (*80*). This study has demonstrated that vibrant changes at the (pan)genomic and genetic levels have facilitated fast morphotype diversification and subspeciation in *Br*, and deepened our understanding of the contributions of (pan)genomic and genetic diversity on Darwin's abominable mystery.

Pangenomic basis of trait development in _B. rapa_

Pan-GWAS of SNPs, InDels and PAVs can help identify genes associated with traits and phenotypical variations at the population level, the accession/morphotype/subspecies level in our case. The 11 *Br* accessions which were selected for our T2T gapless genome and pangenome assemblies have been often used as elite breeding germplasm with distinct agronomic traits and regional adaptability to cultivation. They have been successfully utilized to develop commercially important cultivars with specific characteristics such as early maturation in the CC1-derived "Yuxin" series, heat tolerance in the CC2-derived "Yuxia" series, bolting resistance in CC3-derived "Degao Qinglang 1", broad adaptability in CX1- "Youlv 703", cold tolerance in WTC1-derived "Huiwu 11", and clubroot resistance in TU1. These cultivars are grown in major production regions across China and of great agricultural significance and economic value. As parts of our pan-GWAS coupled with genetic mapping of Chinese cabbage mutant population, we established the pangenomic and

genetic basis of a single dominant gene *BrLH1* that acts as the main morphogenetic regulator for leafy head formation in the 11 elite *Br* accessions. Unlike *fg7*, accessions without AAC-haplotype are not dwarf. It is possible that, contrasting the premature stop-gain in *BrLH1* in *fg7*, the three SNPs in *BrLH1* among natural populations without the AAC-haplotype cause three-AA changes in BrLH1. Such AA substitution may negatively affect BrLH1 to modulate heading but not plant growth. Additionally, overall genetic background may also contribute to plant development in accessions with vs without AAC haplotype. Nevertheless, this work has showcased that the 1,720 re-sequenced genomes, 11 T2T genomes and the gapless pangenome for *Br* can not only be used to underpin fundamental questions over plant evolution but also provide invaluable resources for future *Brassica* crop breeding.

Materials and Methods

Sample selection and sequencing

1,720 accessions (table S1) from 7 elite subspecies (8 major morphotypes) of *B. rapa* were selected to conduct genome re-sequencing. We specifically chose 11 representatives from among the 1,720 accessions for T2T genome assembly. High molecular weight genomic DNA was extracted from fresh young leaves of two-week-old seedlings by the CTAB method and followed by purification with GrandOmics BAC-Long kit V1.0 (Wuhan Grandomics Biotechnology Co., Ltd.), according to the standard operating procedure provided by the manufacturer. For re-sequencing, whole-genome sequencing libraries for short reads were constructed by MGIEasy Universal DNA Library Prep Kit V1.0 (CAT#1000005250, MGI) following the standard protocol and sequenced on DNBSEQ-T7RS platform.

To de novo assemble gapless genomes for the 11 accessions, we performed multiple sequencing technologies. For Ultra-long Nanopore library, genomic DNA was selected (> 100 kb) with the SageHLS HMW library system (Sage Science), and then processed using the Ligation sequencing 1D Kit (SQK-LSK114, Oxford Nanopore Technologies, Oxford, UK) according to the manufacturer's instructions. ONT DNA libraries (approximately 400 ng) were constructed and sequenced on the PromethION48 (Oxford Nanopore Technologies). PacBio HiFi libraries were constructed based on the protocol of SMRT bell prep kit 3.0 kit manual and sequencing was performed using the PacBio Revio according to the operating manual provided by PacBio. Hi-C library of 11 accessions was constructed according to the manufacturers' instructions and were subsequently sequenced on the DNBSEQ-T7RS platform. We extracted total RNA using the TRIzol CTAB-LiCl method from root, bud and silique at seedling (leaf, stem and root), mature plant (leaf and stem), flowering (flowering bud), and seeding (silique) of 11 accessions to prepare RNA sequencing libraries and then to sequence on the DNBSEQ-T7RS platform. The extracted RNA was used to construct an Iso-seq library, which was then sequenced on the PacBio Sequel II platform.

Antibody production

To produce anti-BrCENH3 antibody, we first identified CENH3 protein(s) by performing BLAST searches using the *Arabidopsis thaliana AtCENH3* gene (also known as *HTR12*, *AT1G01370.1*) as the query against each of the assembled genomes

of 11 *B. rapa* subspecies/morphotypes. We uncovered each genome possesses a single *CENH3* gene with the following gene identifier (table S13), TU1: *BraA09g06535.TU1*; MIZ1: *BraA09g06418.MIZ1*; OI1: *BraA09g06449.OI1*; PC1: *BraA09g06475.PC1*; WTC1: *BraA09g06551.WTC1*; CX1: *BraA09g06692.CX1*; TC1: *BraA09G06864.TC1*; CC3: *BraA09g06585.CC3*; CC2: *BraA09G06630.CC2*; CC1: *BraA09G06587.CC1*; and CCA03: *BraA09g06511.V2*. The deduced BrCENH3 proteins from these genes are almost identical among the 11 *B. rapa* subspecies/morphotypes, but exhibit only 65.2 - 67.6% amino acid identities with AtCENH3 (table S12). Nevertheless, the C-termini of BrCENH3s, AtCENH3 and NbCENH3 are highly conserved and comprise the typical CATD domain which is crucial for centromere targeting (*81*). However, the N-terminal sequences vary significantly among those CENH3 proteins (fig. S6A and table S13). We further compared BrCENH3s, BraH3s, AtH3 and NbH3 and selected a specific and highly antigenic dodecapeptide (KHFASRARDRNP) at the N-terminal of BrCENH3s (fig. S6, A to C). This dodecapeptide plus an extra cysteine residue was chemically synthesized, conjugated to the Keyhole Limpet Hemocyanin, a carrier protein through the cysteine linker, and injected into two rabbits. Blood was collected to obtain antisera which were then purified through affinity chromatography to produce specific polyclonal antibodies against BrCENH3s.The antibody was produced via the commercial service from Nantong Teastog Biotechnology Co., Ltd., China.

Specificity of the BrCENH3 antibody

We tested and validated the specificity of the anti-BrCENH3 antibody in three different experimental settings. (i) Transient assay. A CaMV 35S promoter-controlled BraCENH3 expression cassette 35S::BrCENH3 was constructed and cloned into Agrobacterium tumefaciens GV3101. Young leaves of *Nicotiana benthamiana* were agroinfiltrated with *Agrobacterium. tumefaciens* GV3101 carrying 35S:BrCENH3. At 4 days post-agroinfiltration, total protein was extracted from these infiltrated leaf tissues (OE) and analyzed by western blot using the anti-BrCENH3 or anti-H3 antibody. The antibody specifically reacted to BrCENH3-CCA03 that was transiently expressed in *Nicotiana benthamiana.* This is evidenced by Western detection of a single band of the predicted BrCENH3 size using the anti-BrCENH3 antibody (fig. S7A). (ii) Western detection of BrCENH3s and BrH3s in *B. rapa*. The antibody was able to detect a single specific nuclear protein of the predicted BrCENH3 size in total proteins extracted from 11 *B. rapa* subspecies/morphotypes, but not from *N. benthamiana*, consistent with the fact that the BrCENH3-specific dodecapeptide is absent in the *N. benthamiana* homologue NbCENH3 (fig. S7B). (iii) Immunostaining of chromosomes with anti-BrCENH3 antibodies. Immunostaining was performed on root tips of four example *B. rapa* subspecies CCA03, PC1, WTC1, and CX1. The anti-BrCENH3 antibody produced a single intensive signal at the primary constriction of each chromosome (primary chromosome constriction refers to a narrowing or indentation on a chromosome mostly at the CENH3-binding centromeric region) of the four *B. rapa* subspecies. No such immunostaining was observed on chromosomes with negative control antisera (fig. S8). Collectively, these results demonstrate the anti-BrCENH3 antibody we generated is highly specific to BrCENH3.

ChIP and ChIP-seq

The ChIP experiments were carried out in accordance with a previously published protocol (*82*). Untreated DNA was used as the input control. The ChIP and input DNA samples were employed for library construction, following the guidelines outlined in the Illumina protocol (NEBNext-Ultra™ DNA Library Prep Kit for Illumina E7370; New England Biolabs, https://www.neb.com). These libraries were then sequenced using the HiSeq 2,500 platform (Illumina, https://www.illumina.com).

Western Blot

Histones were extracted from 300 mg of fresh leaf tissue harvested from two-week-old plants using the plant-specific histone extraction kit (Solarbio, Cat# EX1530, China) following the manufacturer's protocol with minor modifications. The extracts were dissolved in SDS sample buffer (60Mm Tris-HCl, pH6.8, 2% (w/v) SDS, 10% (v/v) glycerol, 5% (v/v) β-mercaptoethanol, 0.01% (w/v) bromphenol blue), and denaturated by heating at 95 °C for 5 minutes. The proteins were transferred onto PVDF membranes (Millipore). The membranes were blocked with 5% (w/v) non-fat dry milk in Tris-buffered saline (TBS)/Tween20. Subsequently the membranes were probed 2 hours at room temperature with a rabbit anti-BrCENH3 antibody/ anti-H3 antibody (Abcom, ab1791). After washing, the membranes were incubated with anti-rabbit secondary antibody conjugated to HRP (Bioss, bs-0295G-HRP). The histone H3 abundance is used as a control. To verify the specificity of the anti-BrCENH3 antibody, we overexpressed the *BrCENH3* gene from CCA03 in *N. benthamiana*.

Chromosomal immunofluorescence

Immunostaining was carried out based on established published protocols (*82*), with some modifications. Following fixation with paraformaldehyde, root tips were gently squashed onto glass slides and subsequently dehydrated in 75% ethanol for 5 minutes. Anti-BrCENH3 antibodies were then applied to the slides and incubated in a humid chamber at room temperature (RT, 20-25°C) for 3 hours. The slides underwent three washes in full-strength PBS before being incubated with Alexa Fluor 594 chicken anti-rabbit IgG (ThermoFisher Scientific, https://www.thermofisher.com) as the secondary antibody at 37°C for 1 hour. After another three rounds of washing in full-strength PBS, the slides were air-dried at RT, and the chromosomes were counterstained with 4', 6-diamidino-2-phenylindole (DAPI) before being examined under an epifluorescence microscope.

Genome assembly of the 11 *B. rapa* accessions

(i) Preliminary genome assembly

(a) Sequencing data quality control was performed on Nanopore ultra-long reads, PacBio Revio HiFi reads, and MGI paired-end reads using dorado (https://github.com/nanoporetech/dorado/), fastp (v0.23.4), and smrtlink (v13.0, https://www.pacb.com/smrt-link/), respectively, to obtain high quality reads.

(b) Initial genome assembly (v1) was conducted using Hifiasm (*83*) (with default parameters) based on HiFi reads. Subsequently, genome assemblies for versions v2 and v3 were independently generated using Hifiasm (*83*) and verkko (*84*), respectively, by integrating HiFi reads and nanopore ultra-long reads. These assemblies underwent iterative evaluation based on multiple criteria, including the longest assembly length,

the fewest contigs, minimal divergence from the reference genome (assessed via minimap2 (*85*) -x asm5 alignment), and the highest BUSCO completeness score (using the BUSCO database – embryophyte_odb10). If the assembly result is very poor, parameter adjustments were performed iteratively until satisfactory assemblies were achieved for both versions, which including some completed chromosomes. Ultimately, v2 was identified as the optimal assembly (among v1, v2, v3) and selected for downstream analysis due to its superior performance across these metrics, providing the most reliable basic contig sequences.

(c) Chromosome scaffolding was performed using HapHiC (*86*) based on Hi-C reads based on v2, and manual adjustments were made using Juicebox to generate the preliminary chromosome-level genome sequence (pre-v4). Synteny analysis and visualization were conducted against the reference genome using minimap2 (-x asm5) and SyRI (*87*). Meanwhile, HiFi reads (minimap2 -x map-hifi) and nanopore ultra-long reads (minimap2 -x map-ont) were mapped to pre-v4 genome to verify positions with discrepancies from the reference genome. These positions were inspected to determine whether they represented assembly errors or true biological differences. Errors were corrected through scaffolding revisions and positions confirmed as non-errors were retained, which basing on the IGV, to form the final chromosome-level genome (v4). Up to now, this analysis revealed 0-3 gaps per genome (table S4) and 0-2 unassembled telomeres per genome (table S5).

(ii) T2T genome assembly

(a) Based on v4, gaps were filled using HiFi and ONT ultra-long reads (>=50kb) to generate the gap-free genome. This genome was then validated through read mapping (nanopore ultra-long reads, HiFi reads, MGI paired-end reads) and synteny analysis against the reference genome. If correct, it was retained as the gap-free genome.

(b) Gap filling: Sequenced ultra-long and HiFi reads were mapped to the genome using minimap2 with parameters (-ax map-ont --secondary=no for ONT reads and -ax map-hifi --secondary=no for HiFi reads). Reads flanking each gap were selected as anchor reads, which must be uniquely mapped with mapping identity ≥ 99% and mapping coverage ≥ 99% and be the first reads adjacent to the gap on both sides. Unmapped reads (excluding those containing telomeric features, i.e., reads with more than 10 consecutive telomeric repeats) and reads mapping within the coordinate range defined by the paired anchor reads flanking the gap were collected and mutually aligned to identify overlaps between reads. Extension paths were built based on the overlap relationships with anchor reads and among other reads until the reads from both sides of the gap overlapped. Reads within telomeric gaps were extended from anchor reads until telomeric reads to build extension paths. The extended reads within the tiling path were concatenated based on their overlap relationships, and the resulting contig was used to replace the corresponding gap region. The assembled genome after telomere completion was subjected to base error correction using nextpolish2 (*88*). Reads were mapped to this corrected genome using minimap2 to check for regions with significantly abnormal coverage (regions with fewer than 5 "perfectly mapped" ONT reads, where "perfectly mapped" reads are defined as uniquely mapped with identity > 90% and coverage > 90%). If no issues were found, the genome was designated as the

gapfree genome (v5). Nextpolish2 were used to polish the genome.

(c) Telomere assembly: Sequenced HiFi reads were mapped to the gapfree genome using minimap2 with parameters (-x map-hifi --secondary=no for HiFi reads). Reads mapping to the chromosome terminal were selected as terminal reads, when uniquely mapping with mapping identity ≥ 99% and mapping coverage ≥ 99%, and being the innermost reads adjacent to the unassembled telomere. Additionally, reads containing telomeric repeat units were also used as terminal reads. Unmapped reads and terminal reads were assembled using Hifiasm (*83*). The assembled sequence with telomere was mapped to chromosome terminal with minimap2 (-cx asm10) and selected to fill the missing telomeric sequence at the chromosome terminus to form the T2T genome (v6).

(iii) T2T genome assessment

(a) For v6 genome, HiFi reads, ultra-long reads and NGS reads were mapped to the T2T genome using minimap2 with the options -x map-hifi, -x map-ont and BWA (*89*) with default parameters, respectively. Genome coverage was calculated by samtools (*90*). Quality value (QV) (*16*) was obtained by Yak and Merqury based on the NGS data. Genome completeness was evaluated by BUSCO using the embryophyta_odb10 database. The final Hi-C heatmap was generated using HiCExplorer v3.7.2 (*91*). The LTR assembly index (LAI) value (*17*), which uses repeat sequences, evaluated the assembly level of the genome. Such comprehensive assessments of the 11 pre-T2T genomes (v6) led to identify 110 centromeres across the 11 assemblies using ChIP-based techniques and 22 telomeres across the 11 assemblies using the telomeric repeat unit AAACCCT.

(b) Genomic collinearity was analyzed using minimap2 (-x asm5) and SyRI (*87*). To validate certain SVs, we conducted a detailed inspection of the reads coverage at the SV loci using integrative genomics viewer (IGV) (*92*). Through whole-genome collinearity analysis between the 11 newly assembled genomes and the published reference genomes in the *B. rapa* morphotypes, SVs visible in the map were identified. Four randomly selected inversions were manually checked to confirm the assembly accuracy of these SVs using IGV visualization (fig. S4). If an SV breakpoint was supported by at least five perfectly matching reads (with a mapping quality score ⩾ 40), the region was deemed to be correctly assembled; regions failing to meet this criterion were segmented and subjected to reassembly. For certain complex regions of the genome, such as rDNAs, where read alignment results could not conclusively determine assembly accuracy, digital PCR experiments were employed to verify the copy numbers of rDNA and centromeric repeat units. If the assembly results closely matched the experimentally determined values, the region was considered correctly assembled; otherwise, local reassembly was performed for correction. When all evaluation results met the required standards, the genome was finalized as the complete Telomere-to-Telomere (T2T) version (v7).

Genome annotation

EDTA (*93*) was utilized to detect the repeat elements among the genome sequences with --sensitive 1 --anno 1 --evaluate 1. As for protein-coding genes, ab initio prediction, homology search, and transcripts prediction were employed to predict genes based on a repeat-masked sequences. For the transcripts-based approach, RNA-seq data from 7

tissues were quality control and to assemble the transcripts by using fastp, STAR (*94*) (v2.7.9a), and Stringtie (*95*) (v1.3.4d). Iso-seq were aligned to T2T genome using minimap2 with the parameter "-x splice -uf", and nonredundant transcripts were obtained using the IsoSeq3 (v3.8.2, https://github.com/PacificBiosciences/IsoSeq). Non-redundant transcripts from RNA-seq and Iso-seq were used to predict gene models via PASA software (v2.5.2). GeMoMa (*96*) and AUGUSTUS were chosen to carry out the homology search and ab initio prediction, respectively. EVidenceModeler (EVM) (*97*) integrated the gene set, removing genes associated with transposable elements (TEs) using TransposonPSI (http://transposonpsi.sourceforge.net/) and filtering out misannotated genes. Untranslated regions (UTRs) and alternative splicing regions were identified using PASA, with the longest transcripts retained for each locus. SwissProt, NR and KOG were annotated by using BLASTp with E-value cutoff of 1E-05 based on the protein-coding sequences. KEGG, GO and motifs were marked by InterProScan. Noteworthily, we used genome-guided transcript assembly (STAR + StringTie) for RNA-seq. On the other hand, the default Isoseq3 was used to analyze Iso-seq reads for characterizing isoform transcript reads. We then combined and mapped transcripts from both RNA-seq and Iso-seq analyses to genome to predict protein-coding gene expression profiles using PASA (-f --ALIGNERS blat, gmap). This "combined" strategy guarantees to optimize annotation completeness and accuracy in our study.

Synteny analysis

Minimap2 was used for synteny analysis between the reference genome and the T2T genome to obtain the correspondences and directional relationships between each chromosome. Then, SyRI and plotsr (*98*) were used to generate whole-genome plots of the reference genome and the T2T genome.

Evaluation of the copy number of satellite DNAs and the 45S rDNA

To evaluate the copy number of satellite DNAs and the 45S rDNA, we first conducted de novo annotation using Satellite Repeat Finder (SRF). This led to identification of five novel satellites pBrSTR497, pBrSTR92, pBrSTR144, pBrSTR197, and pBrSTR228 along with previously reported two satellites CentBr and PCR630. A multi-species rRNA database was built for BLASTp genome alignment (similarity & coverage >95%) to identify candidate rDNAs. As a complementary means, RNAmmer was also used to predict rRNA loci with sequencing-based rDNA profiling. Combining all these bioinformatics analyses with rDNA structural features (18S-ITS1-5.8S-ITS2-28S clusters interspersed with 5S), regions fulfilling either of alignment criteria plus structural patterns were confirmed as genomic rDNA clusters. To ensure the accurate assembly and accurate evaluation of the copy number of satellite DNAs and the 45S rDNA, we used manual inspection and Integrative Genomics Viewer (IGV) to confirm that long reads fully cover the satellite and 45S rDNA regions and these regions exhibit uniform long-read coverage. We then performed digital PCR to experimentally measure and validate the copy numbers of satellite DNAs. Due to PCR630 exhibits greater sequence conservation, PCR630 was selected as a representative marker for digital PCR quantification. For 45S rDNA, we targeted conserved regions within the 5.8S, 18S, and 28S rDNAs. Among all accessions tested, the copy numbers of PCR630 and the 45S

rDNAs (5.8S, 18S, and 28S) measured by digital PCR vs in silico assays on genome alignments are similar (fig. S5 and fig. S17).

Centromere and pericentromere analysis

To identify/verify the centromeric regions, the filtered ChIP-seq data was mapped to the reference genome using BWA (*89*), after which the the aligned files were processed with samtools (*90*), after which MACS3 (*99*) was used to perform peak calling based on input and output data. This analysis revealed a distinct and consecutive region with strong peaks per chromosome in each accession with high read coverage, indicating the positions of centromeres.

To predict pericentromeric regions, we first assessed the complex and repetitive sequences of the genome. Using KMC (v3.2.1) (https://github.com/refresh-bio/KMC), we performed k-mer counting and statistical analysis on 11 *B. rapa* genome sequences. We used the kmc and kmc_dump commands with the following parameters: -fm (enable memory mode), -k151 (set k-mer length to 151), -ci200 (set the minimum occurrence count for k-mer), and -cs1000000 (set the k-mer count upper limit to 1,000,000). Next, we utilized the SRF tool (*100*) (https://github.com/lh3/srf) to identify and analyze satellite repetitive sequences within the genomes, using the default parameters. The identified satellite repetitive sequences were manually curated to determine their smallest repetitive units, ultimately identifying six satellite sequences. To count these satellite sequences in each genome, we used blastn for whole-genome alignment based on two criteria: sequence similarity greater than 95% and matched sequence length greater than 95%. Sequences meeting these criteria were considered as satellite sequences.

To comprehensively characterize the pericentromeric regions of *B. rapa*, we integrated the distribution patterns of satellites, repeat element density, and annotated gene density. We defined the boundary of the first candidate region as the location of five consecutive 100 kb windows containing more than three satellites; the boundary of the second candidate region as the location of 50 consecutive 10 kb sliding windows with a repeat element coverage greater than 0.95; and the boundary of the third candidate region as the location of 50 consecutive 10 kb sliding windows with a gene density less than 0.05. We defined regions that simultaneously met two criteria as the boundaries of the final outer pericentromeric regions. Furthermore, to verify the accuracy of these pericentromeric positions, we also assessed DNA methylation density based on HiFi reads and validated the positions using TADs and AB compartment data based on Hi-C reads as auxiliary tools. The sliding window for DNA methylation density was set to 100kb, with 100kb resolution for AB compartments and 40kb resolution for TADs.

Profiling DNA methylation, TADs and AB compartments

Genome-wide distributions of DNA methylation, TADs and AB compartments were experimentally examined and their distinct profiles in (peri)centromeres vs chromosomal arms were analyzed by bioinformatics tools. To achieve these, high molecular weight genomic DNA was extracted from fresh young leaves of two-week-old seedlings by the CTAB method. After library construction, PacBio HiFi and Hi-C sequencing were carried out to obtain the raw data of whole-genome DNA methylation,

TADs, and AB compartments. The DNA methylation profiles were derived from HiFi sequencing data generated via the PacBio Revio platform as described (29). We then plotted the genome-wide DNA methylation levels in the CG, CHG, and CHH contexts (fig. S9) and found that the DNA methylation levels were higher in (peri)centromeric regions than chromosome arms (fig. S10). However, we have not used DNA methylation data to define (peri)centromere. On the other hand, the TAD and AB compartment results were directly generated from Hi-C sequencing data. We observe that TADs in (peri)centromeric regions were distinct from those in the chromosome arms, and that B compartments (closed chromatin) were predominantly located at (peri)centromeric regions in contrast to the open chromatin A compartments which were mainly found in chromosome arms (fig. S11).

Gene-based pangenome construction

To construct the pangenome based on orthologous relationships, we employed gffread (*101*) to extract the longest protein sequences for each gene from both the sample genomes and the reference genome, utilizing the corresponding annotation files. Gene family identification was conducted using Orthofinder (*102*) (-S diamond), which generated orthologous groups. The pangenome was subsequently categorized according to the number of samples in each orthologous group: groups containing only one sample were classified as private, those with more than one but fewer than 28 samples were designated as dispensable, groups with 28 to 30 samples were termed soft-core, and those with exactly 31 samples were classified as core.

Structural variation (SV) identification and wet-experimental validations

For read-based SVs calling, third-generation sequencing reads from 30 samples were aligned to the reference genome CCA03 using minimap2 (-x map-hifi for hifi reads; -x map-ont for nanopore reads; -x map-pb for sequel reads). SV calling was subsequently performed for each sample using Sniffles (*103*) and CuteSV (*104*) (-l 50). The SV results for each sample were consolidated using SURVIVOR (parameters set to merge 1000,2,1,1,0,50) (https://github.com/fritzsedlazeck/SURVIVOR) to ensure completeness and accuracy. Finally, the SV results across all samples were integrated using SURVIVOR with parameters set to merge 1000,1,1,1,0,50, thereby providing an accurate and comprehensive summary of structural variations. Furthermore, digital PCR experiments were performed to support the reliability of the detected structural variations. We have randomly selected 10 SVs and performed PCR amplification to detect these SVs using eight accessions each from Chinese cabbage, Caixin, and Turnip (fig. S25). Our results demonstrate clear validations on the reliability of the structural variations revealed via the dry-computational methods.

GWAS for leafy head

We obtained 98,227 PAVs, 787,027 InDels and 4,015,383 bi-allelic SNPs were filtered by BCFtools with a missing rate less than (0.5) and a minor allele frequency over (0.05) for GWAS analysis. GWAS were performed using mixed linear model in rMVP package (*105*). We defined the whole genome significance cutoff at 0.05/total number of each type of variations. The results of GWAS were visualized using custom codes based on the R package ggplot2.

PCR and RT-qPCR

For PCR, 20 μL PCR containing 100 ng of genomic DNA, 2 × Rapid Taq Master Mix, 10 μM primers was used for various fragment amplifications. Primer sequences are shown in table S26. A touchdown PCR protocol was used as followings: initial denaturation at 95 ℃ for 15 seconds; annealing at 62 ℃ for 15 seconds, decreasing by 0.6 ℃ per cycle; and extension at 72 ℃ for 45 seconds, followed by repeating these steps for 20 cycles. After enrichment, the program continued for 20 cycles as follows: 95 ℃ for 15 seconds, 58 ℃ for 15 seconds, and 72 ℃ for 45 seconds.

For RT-qPCR, 1 μg aliquot of total RNA for each sample was used for reverse transcription and first-strand cDNA synthesis using the PrimeScript™ RT reagent Kit with gDNA Eraser (TAKARA). SYBR Green Master Mix (Vazyme) was used in RT-qPCR analyses. RT-qPCR analyses were performed with three technical replicates in the LightCycler ® 96 (Roche) under the following conditions: 10 minutes at 95 ℃, 40 cycles of 10 seconds at 95 ℃, 10 seconds at 57 ℃, and 10 seconds at 72 ℃. After PCR, a melting curve was generated by 10 seconds at 95 ℃, 60 seconds at 60 ℃, and 1 seconds at 97 ℃. The $2^{-\Delta\Delta Ct}$ method was used to calculate relative gene expression levels. Gene-specific primers for RT-qPCR are presented in table S26.

Digital PCR

Digital PCR was employed to quantify absolute copy numbers of satellite and rDNA. Genomic DNA was used as template in a 20 μL reaction mixture containing 2× T5 Fast qPCR Mix (probe), specific primers, fluorescent probe, and ROX reference dye. Reactions were partitioned into droplets and amplified on the SinofU DQ24 digital PCR system using a cycling program of denaturation at 95 ℃ and annealing/extension at 60 ℃ for 40 cycles. After amplification, fluorescence of each droplet was measured, and the ratio of positive to negative droplets was analyzed using a poisson distribution to determine the absolute copy number of satellite and rDNA.

Virus induced gene silencing (VIGS) and gene expression assays

For functional genomics, VIGS was performed to silence to *BrNfu2* and *BrQS* in Caixin CX1, and RT-qPCR was carried out to analyze silencing efficiency as previously described (*106*). Impacts of gene silencing on CX1 plant development (sizes) and flowering time were examined, measured and recorded during the courses of repeated VIGS experiments.

References and Notes

1. T. J. Davies *et al.*, Darwin's abominable mystery: Insights from a supertree of the angiosperms. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 1904-1909 (2004). doi: 10.1073/pnas.0308127100; pmid:14766971
2. R. J. A. Buggs, The deepening of Darwin's abominable mystery. *Nat. Ecol. Evol.* **1**, 0169 (2017). doi: 10.1038/s41559-017-0169; pmid:28812628
3. H. Yan *et al.*, Post-polyploidization centromere evolution in cotton. *Nat. Genet.* **57**, 1021-1030 (2025). doi: 10.1038/s41588-025-02115-3; pmid:40033059
4. S. Secomandi *et al.*, Pangenome graphs and their applications in biodiversity genomics. *Nat. Genet.* **57**, 13-26 (2025). doi: 10.1038/s41588-024-02029-6; pmid:39779953
5. Y. C. Liu *et al.*, Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162-176 (2020). doi: 10.1016/j.cell.2020.05.023; pmid:32553274

6. P. Qin *et al.*, Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542-3558 (2021). doi: 10.1016/j.cell.2021.04.046; pmid:34051138
7. T. Shi *et al.*, The super-pangenome of *populus* unveils genomic facets for its adaptation and diversification in widespread forest trees. *Mol. Plant* **17**, 725-746 (2024). doi: 10.1016/j.molp.2024.03.009; pmid:38486452
8. Y. L. Zhang *et al.*, Telomere-to-telomere super-pangenome provides direction for watermelon breeding. *Nat. Genet.* **56**, 1750-1761 (2024). doi: 10.1038/s41588-024-01823-6; pmid:38977857
9. Z. J. Liu *et al.*, Grapevine pangenome facilitates trait genetics and genomic breeding. *Nat. Genet.* **56**, 2804-2814 (2024). doi: 10.1038/s41588-024-01967-5; pmid:39496880
10. F. Cheng *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). doi: 10.1038/ng.3634; pmid:27526322
11. X. Cai *et al.*, Impacts of allopolyploidization and structural variation on intraspecific diversification in *Brassica rapa*. *Genome Biol.* **22**, 166-190 (2021). doi: 10.1186/s13059-021-02383-2; pmid:34059118
12. A. C. McAlvay *et al.*, *Brassica rapa* domestication: untangling wild and feral forms and convergence of crop morphotypes. *Mol. Biol. Evol.* **38**, 3358-3372 (2021). doi: 10.1093/molbev/msab108; pmid:33930151
13. X. Qi *et al.*, Genomic inferences of domestication events are corroborated by written records in *Brassica rapa*. *Mol Ecol.* **26**, 3373-3388 (2017). doi: 10.1111/mec.14131; pmid:28371014
14. Y. F. Zhou *et al.*, The complexity of structural variations in *Brassica rapa* revealed by assembly of two complete T2T genomes. *Sci. Bull.* **69**, 2346-2351 (2024). doi: 10.1016/j.scib.2024.03.030; pmid:38548570
15. L. Zhang *et al.*, A near-complete genome assembly of *Brassica rapa* provides new insights into the evolution of centromeres. *Plant Biotechnol. J.* **21**, 1022-1032 (2023). doi: 10.1111/pbi.14015; pmid:36688739
16. A. Rhie, B. P. Walenz, S. Koren, A. M. Phillippy, Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245-272 (2020). doi: 10.1186/s13059-020-02134-9; pmid:32928274
17. S. Ou, J. Chen, N. Jiang, Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018). doi: 10.1093/nar/gky730; pmid:30107434
18. K. L. McKinley, I. M. Cheeseman, The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16-29 (2016). doi: 10.1038/nrm.2015.5; pmid:26601620
19. M. Han, Y. Yang, M. Zhang, K. Wang, Considerations regarding centromere assembly in plant whole-genome sequencing. *Methods* **187**, 54-56 (2021). doi: 10.1016/j.ymeth.2020.09.006; pmid:32920129
20. K. B. Lim *et al.*, Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassica rapa*. *Mol. Cells* **19**, 436-444 (2005). doi: 10.1016/S1016-8478(23)13190-6; pmid:15995362
21. J. M. Song *et al.*, Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* **14**, 1757-1767 (2021). doi: 10.1016/j.molp.2021.06.018; pmid:34171480
22. X. Chang *et al.*, High-quality *Gossypium hirsutum* and *Gossypium barbadense* genome assemblies reveal the landscape and evolution of centromeres. *Plant Commun.* **5**, 100722-100735 (2024). doi: 10.1016/j.xplc.2023.100722; pmid:37742072
23. H. N. Zhao *et al.*, Gene expression and chromatin modifications associated with maize centromeres. *G3-Genes Genom Genet* **6**, 183-192 (2016). doi: 10.1534/g3.115.022764; pmid:26564952

24. J. Chen *et al.*, A complete telomere-to-telomere assembly of the maize genome. *Nat. Genet.* **55**, 1221-1231 (2023). doi: 10.1038/s41588-023-01419-6; pmid:37322109
25. G. J. Hu *et al.*, A telomere-to-telomere genome assembly of cotton provides insights into centromere evolution and short-season adaptation. *Nat. Genet.* **57**, 1031-1043 (2025). doi: 10.1038/s41588-025-02130-4; pmid:40097785
26. Y. B. Wang *et al.*, Four near-complete genome assemblies reveal the landscape and evolution of centromeres in Salicaceae. *Genome Biol.* **26**, 111-137 (2025). doi: 10.1186/s13059-025-03578-7; pmid:40317068
27. W. L. Wang *et al.*, Chromosome level comparative analysis of *Brassica* genomes. *Plant Mol. Biol.* **99**, 237-249 (2019). doi: 10.1007/s11103-018-0814-x; pmid:26601620
28. S. Perumal *et al.*, A high-contiguity *Brassica nigra* genome localizes active centromeres and defines the ancestral *Brassica* genome. *Nat. Plants* **6**, 929-941 (2020). doi: 10.1038/s41477-020-0735-y; pmid:32782408
29. K. Paritosh, A. K. Pradhan, D. Pental, A highly contiguous genome assembly of *Brassica nigra* (BB) and revised nomenclature for the pseudochromosomes. *BMC Genomics* **21**, 887-899 (2020). doi: 10.1186/s12864-020-07271-w; pmid:33308149
30. X. Li *et al.*, Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in *Brassica oleracea*. *Nat. Genet.* **56**, 517-529 (2024). doi: 10.1038/s41588-024-01655-4; pmid:38351383
31. J. H. Yang *et al.*, The genome sequence of allopolyploid and analysis of differential homoeolog gene expression influencing selection. *Nat. Genet.* **48**, 1225-1232 (2018). doi: 10.1038/ng.3657; pmid:27595476
32. J. H. Yang *et al.*, Genomic signatures of vegetable and oilseed allopolyploid and genetic loci controlling the accumulation of glucosinolates. *Plant Biotechnol. J.* **19**, 2619-2628 (2021). doi: 10.1111/pbi.13687; pmid:34448350
33. K. Paritosh *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). doi: 10.1111/pbi.13492; pmid:33073461
34. K. B. Lim *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 (2006). doi: 10.1111/j.1365-313X.2006.02952.x; pmid:17156411
35. J. M. Song *et al.*, Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* **6**, 34-45 (2020). doi: 10.1038/s41477-019-0577-7; pmid:31932676
36. X. M. Song *et al.*, *Brassica carinata* genome characterization clarifies U's triangle model of evolution and polyploidy in *Brassica Plant Physiol.* **186**, 388-406 (2021). doi: 10.1093/plphys/kiab048; pmid:33599732
37. W. C. Yim *et al.*, The final piece of the Triangle of U: Evolution of the tetraploid *Brassica carinata* genome. *Plant Cell* **34**, 4143-4172 (2022). doi: 10.1093/plcell/koac249; pmid:35961044
38. Y. Niu *et al.*, A *Brassica carinata* pan-genome platform for *Brassica* crop improvement. *Plant Commun.* **5**, 100725 (2024). doi: 10.1016/j.xplc.2023.100725; pmid:37803826
39. X. R. Hou, D. P. Wang, Z. K. Cheng, Y. Wang, Y. L. Jiao, A near-complete assembly of an *Arabidopsis thaliana* genome. *Mol. Plant* **15**, 1247-1250 (2022). doi: 10.1016/j.molp.2022.05.014; pmid:35655433

40. E. J. Drdová *et al.*, Developmental plasticity of Arabidopsis hypocotyl is dependent on exocyst complex function. *J. Exp. Bot.* **70**, 1255-1265 (2019). doi: 10.1093/jxb/erz005; pmid:30649396
41. D. Safavian *et al.*, RNA silencing of exocyst genes in the stigma impairs the acceptance of compatible pollen in Arabidopsis. *Plant Physiol.* **169**, 2526-2538 (2015). doi: 10.1104/pp.15.00635; pmid:26443677
42. M. Fendrych *et al.*, The exocyst complex is involved in cytokinesis and cell plate maturation. *Plant Cell* **22**, 3053-3065 (2010). doi: 10.1105/tpc.110.074351; pmid:20870962
43. B. Touraine *et al.*, Nfu2: a scaffold protein required for [4Fe-4S] and ferredoxin iron-sulphur cluster assembly in chloroplasts. *Plant J.* **40**, 101-111 (2004). doi: 10.1111/j.1365-313X.2004.02189.x; pmid:15361144
44. J. H. M. Schippers *et al.*, The mutation of quinolinate synthase affects nicotinamide adenine dinucleotide biosynthesis and causes early ageing. *Plant Cell* **20**, 2909-2925 (2008). doi: 10.1105/tpc.107.056341; pmid:18978034
45. K. Chapman *et al.*, CEP receptor signalling controls root system architecture in Arabidopsis and Medicago. *New Phytol.* **226**, 1809-1821 (2020). doi: 10.1111/nph.16483; pmid:32048296
46. A. C. Bryan, A. Obaidi, M. Wierzba, F. E. Tax, XYLEM INTERMIXED WITH PHLOEM1, a leucine-rich repeat receptor-like kinase required for stem growth and vascular development in *Arabidopsis thaliana*. *Planta* **235**, 111-122 (2012). doi: 10.1007/s00425-011-1489-6; pmid:21853254
47. D. R. Gallie, Z. Chen, Chloroplast-localized iron superoxide dismutases FSD2 and FSD3 are functionally distinct in *Arabidopsis*. *Plos One* **14**, e0220078 (2019). doi: 10.1371/journal.pone.0220078; pmid:31329637
48. F. Myouga *et al.*, A heterocomplex of iron superoxide dismutases defends chloroplast nucleoids against oxidative stress and is essential for chloroplast development in *Arabidopsis*. *Plant Cell* **20**, 3148-3162 (2008). doi: 10.1105/tpc.108.061341; pmid:18996978
49. M. Sustr, H. Konrádová, M. Martincová, A. Soukup, E. Tylová, Potassium transporter KUP9 regulates plant response to K plus deficiency and affects carbohydrate allocation in *A. thaliana*. *J. Plant Physiol.* **292**, 154147 (2024). doi: 10.1016/j.jplph.2023.154147; pmid:38096629
50. M. Okamoto *et al.*, High-affinity nitrate transport in roots of Arabidopsis depends on expression of the *NAR2*-like gene *AtNRT3.1*. *Plant Physiol.* **140**, 1036-1046 (2006). doi: 10.1104/pp.105.074385; pmid:16415212
51. M. H. Kang *et al.*, The pan-genome and local adaptation of *Arabidopsis thaliana*. *Nat. Commun.* **14**, 55-73 (2023). doi: 10.1038/s41467-023-42029-4; pmid:37802986
52. J. Du *et al.*, Nitric oxide induces cotyledon senescence involving co-operation of the *NES1*/*MAD1* and *EIN2*-associated *ORE1* signalling pathways in *Arabidopsis*. *J. Exp. Bot.* **65**, 4051-4063 (2014). doi: 10.1093/jxb/ert429; pmid:24336389
53. C. Belser *et al.*, Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Biotechnol.* **4**, 879-887 (2018). doi: 10.1038/s41477-018-0289-4; pmid:30390080
54. K. Yue *et al.*, PP2A-3 interacts with ACR4 and regulates formative cell division in the *Arabidopsis* root. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 1447-1452 (2016). doi: 10.1073/pnas.1525122113; pmid:26792519
55. E. Ramireddy, L. Chang, T. Schmülling, Cytokinin as a mediator for regulating root system architecture in response to environmental cues. *Plant Signaling Behav.* **9**, e27771 (2014). doi: 10.4161/psb.27771; pmid:24509549
56. F. Berger, J. R. Haag, O. Pontes, C. S. Pikaard, Metal A and metal B sites of nuclear RNA polymerases

Pol IV and Pol V are required for siRNA-dependent DNA methylation and gene silencing. *Plos One* **4**, e4110 (2009). doi: 10.1371/journal.pone.0004110; pmid:19119310

57. N. Dharmasiri *et al.*, *AXL* and *AXR1* have redundant functions in RUB conjugation and growth and development in Arabidopsis. *Plant J.* **52**, 114-123 (2007). doi: 10.1111/j.1365-313X.2007.03211.x; pmid:17655650
58. J. O. Narciso *et al.*, Biochemical and functional characterization of GALT8, an *Arabidopsis* GT31 β-(1,3)-galactosyltransferase that influences seedling development. *Front. in Plant Sci.* **12**, 678564 (2021). doi: 10.3389/fpls.2021.678564;
59. Q. H. Li *et al.*, Haplotype-resolved T2T genome assemblies and pangenome graph of pear reveal diverse patterns of allele-specific expression and the genomic basis of fruit quality traits. *Plant Commun.* **5**, 101000-101021 (2024). doi: 10.1016/j.xplc.2024.101000; pmid:38859586
60. X. F. Yu *et al.*, Super pan-genome reveals extensive genomic variations associated with phenotypic divergence in Actinidia. *Mol. Hortic.* **5**, 4-20 (2025). doi: 10.1186/s43897-024-00123-1; pmid:39849617
61. N. Li *et al.*, Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Na. Genet.* **55**, 852-860 (2023). doi: 10.1038/s41588-023-01340-y; pmid:37024581
62. Z. Wu *et al.*, RNA binding proteins RZ-1B and RZ-1C play critical roles in regulating pre-mRNA splicing and gene expression during development in *Arabidopsis*. *Plant Cell* **28**, 55-73 (2016). doi: 10.1105/tpc.15.00949; pmid:26721863
63. B. C. W. Crawford *et al.*, Genetic control of distal stem cell fate within root and embryonic meristems. *Science* **347**, 655-659 (2015). doi: 10.1126/science.aaa0196; pmid:25612610
64. M. Cerise *et al.*, Two modes of gene regulation by TFL1 mediate its dual function in flowering time and shoot determinacy of *Arabidopsis*. *Development* **150**, dev202089 (2023). doi: 10.1242/dev.202089; pmid:37971083
65. Y. H. Wang *et al.*, Molecular variation in a functionally divergent homolog of FCA regulates flowering time in *Arabidopsis thaliana*. *Nat. Commun.* **11**, 5830-5844 (2020). doi: 10.1038/s41467-020-19666-0; pmid:33203912
66. X. M. Zhang *et al.*, OCTOPUS regulates BIN2 to control leaf curvature in Chinese cabbage. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2208978119 (2022). doi: 10.1073/pnas.2208978119; pmid:35969746
67. X. Sun *et al.*, Construction of a high-density mutant population of Chinese cabbage facilitates the genetic dissection of agronomic traits. *Mol. Plant* **15**, 913-924 (2022). doi: 10.1016/j.molp.2022.02.006; pmid:35150930
68. W. Ma, P. Zhang, J. Zhao, Y. Hong, Chinese cabbage: an emerging model for functional genomics in leafy vegetable crops. *Trends Plant Sci.* **28**, 515-518 (2023). doi: 10.1016/j.tplants.2023.02.008; pmid:36914552
69. C. Trehin *et al.*, *QUIRKY* interacts with *STRUBBELIG* and *PAL OF QUIRKY* to regulate cell growth anisotropy during *Arabidopsis* gynoecium development. *Development* **140**, 4807-4817 (2013). doi: 10.1242/dev.091868; pmid:24173806
70. L. Fulton *et al.*, *DETORQUEO*, *QUIRKY*, and *ZERZAUST* represent novel components involved in organ development mediated by the receptor-like kinase STRUBBELIG in *Arabidopsis thaliana*. *PLoS Genet.* **5**, e1000355 (2009). doi: 10.1371/journal.pgen.1000355; pmid:19180193
71. J. H. Song, S. H. Kwak, K. H. Nam, J. Schiefelbein, M. M. Lee, QUIRKY regulates root epidermal cell patterning through stabilizing SCRAMBLED to control CAPRICE movement in *Arabidopsis*.

Nat. Commun. **10**, 1744-1756 (2019). doi: 10.1038/s41467-019-09715-8; pmid:30988311

72. P. Vaddepalli, L. Fulton, M. Batoux, R. K. Yadav, K. Schneitz, Structure-function analysis of STRUBBELIG, an Arabidopsis atypical receptor-like kinase involved in tissue morphogenesis. *Plos One* **6**, e19730 (2011). doi: 10.1371/journal.pone.0019730; pmid:21603601
73. P. Wlodzimierz *et al.*, Cycles of satellite and transposon evolution in *Arabidopsis* centromeres. *Nature* **618**, 557-565 (2023). doi: 10.1038/s41586-023-06062-z; pmid:37198485
74. B. G. Mellone, D. Fachinetti, Diverse mechanisms of centromere specification. *Curr. Biol.* **31**, 1491-1504 (2021). doi: 10.1016/j.cub.2021.09.083; pmid:34813757
75. Q. Hou *et al.*, ZmMS1/ZmLBD30-orchestrated transcriptional regulatory networks precisely control pollen exine development. *Mol. Plant* **16**, 1321-1338 (2023). doi: 10.1016/j.molp.2023.07.010; pmid:37501369
76. Z. Gompert *et al.*, Adaptation repeatedly uses complex structural genomic variation. *Science* **388**, eadp3745 (2025). doi: 10.1126/science.adp3745; pmid:40245138
77. Q. He *et al.*, The near-complete genome assembly of hexaploid wild oat reveals its genome evolution and divergence with cultivated oats. *Nat. Plants* **10**, 2062-2078 (2024). doi: 10.1038/s41477-024-01866-x; pmid:39627369
78. P. Lou *et al.*, Genetic and genomic resources to study natural variation in *Brassica rapa*. *Plant Direct* **4**, e00285 (2020). doi: 10.1002/pld3.285; pmid:33364543
79. D. Dimitrov *et al.*, Diversification of flowering plants in space and time. *Nat. Commun.* **14**, 7609-7625 (2023). doi: 10.1038/s41467-023-43396-8; pmid:37993449
80. A. Tanentzap, K. A. Simonin, A. B. Roddy, Genome downsizing, physiological novelty, and the global dominance of flowering plants. *PLoS Biol.* **16**, e2003706 (2018). doi: 10.1371/journal.pbio.2003706; pmid:29324757
81. Z. Zhou *et al.*, Structural basis for recognition of centromere histone variant CenH3 by the chaperone Scm3. *Nature* **472**, 234-237 (2011). doi: 10.1038/nature09854; pmid:21412236
82. J. L. Han *et al.*, Rapid proliferation and nucleolar organizer targeting centromeric retrotransposons in cotton. *Plant J.* **88**, 992-1005 (2016). doi: 10.1111/tpj.13309; pmid:27539015
83. H. Y. Cheng, G. T. Concepcion, X. W. Feng, H. W. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170-175 (2021). doi: 10.1038/s41592-020-01056-5; pmid:33526886
84. M. Rautiainen *et al.*, Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474-1482 (2023). doi: 10.1038/s41587-023-01662-6; pmid:36797493
85. H. Li, New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572-4574 (2021). doi: 10.1093/bioinformatics/btab705; pmid:34623391
86. X. F. Zeng *et al.*, Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10**, 1184-1200 (2024). doi: 10.1038/s41477-024-01755-3; pmid:39103456
87. M. Goel, H. Q. Sun, W. B. Jiao, K. Schneeberger, SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277-290 (2019). doi: 10.1186/s13059-019-1911-0; pmid:31842948
88. J. Hu *et al.*, Nextpolish2: a repeat-aware polishing tool for genomes assembled using hifi long reads. *GENOM PROTEOM BIOINF* **22**, qzad009 (2024). doi: 10.1093/gpbjnl/qzad009; pmid:38862426
89. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Bioinformatics* **00**, 1-2 (2013). doi: 10.48550/arXiv.1303.3997;

90. H. Li *et al.*, The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078-2079 (2009). doi: 10.1093/bioinformatics/btp352; pmid:19505943

91. J. Wolff *et al.*, Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177-W184 (2020). doi: 10.1093/nar/gkaa220; pmid:32301980

92. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinf.* **14**, 178-192 (2013). doi: 10.1093/bib/bbs017; pmid:22517427

93. S. J. Ou *et al.*, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275-293 (2019). doi: 10.1186/s13059-019-1905-y; pmid:31843001

94. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013). doi: 10.1093/bioinformatics/bts635; pmid:23104886

95. M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, S. L. Salzberg, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650-1667 (2016). doi: 10.1038/nprot.2016.095; pmid:27560171

96. J. Keilwagen, F. Hartung, J. Grau, GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* **1962**, 161-177 (2019). doi: 10.1007/978-1-4939-9173-0_9; pmid:31020559

97. B. J. Haas *et al.*, Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008). doi: 10.1186/gb-2008-9-1-r7; pmid:18190707

98. M. Goel, K. Schneeberger, plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922-2926 (2022). doi: 10.1093/bioinformatics/btac196; pmid:35561173

99. Y. Zhang *et al.*, Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137. (2008). doi: 10.1186/gb-2008-9-9-r137; pmid:18798982

100. Z. Y, C. J, C. H, L. H, De novo reconstruction of satellite repeat units from sequence data. *Genome Res.* **33**, 1994-2001 (2023). doi: 10.1101/gr.278005.123; pmid:37918962

101. G. Pertea, M. Pertea, GFF utilities: gffread and gffcompare. *F1000Research* **9**, 304-323 (2020). doi: 10.12688/f1000research.23297.1; pmid:32489650

102. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238-252 (2019). doi: 10.1186/s13059-019-1832-y; pmid:31727128

103. M. Smolka *et al.*, Detection of mosaic and population-level structural variants with sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024). doi: 10.1038/s41587-023-02024-y; pmid:38168980

104. T. Jiang *et al.*, Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189-213 (2020). doi: 10.1186/s13059-020-02107-y; pmid:32746918

105. L. L. Yin *et al.*, rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *GENOM PROTEOM BIOINF* **19**, 619-628 (2021). doi: 10.1016/j.gpb.2020.10.007; pmid:33662620

106. Z. Yu *et al.*, ETHYLENE RESPONSE FACTOR 070 inhibits flowering in Pak-choi by indirectly impairing *BcLEAFY* expression. *Plant Physiol.* **195**, 986-1004 (2024). doi: 10.1093/plphys/kiae021; pmid:38269601

107. W. Ma *et al.*, *Brassica rapa* pan-genome publication. prerelease. *Zenodo*, (2025). doi:

10.5281/zenodo.17000216;

Acknowledgments:

We thank Xiaoming Wu at Oil Crops Research Institute of Chinese Academy of Agricultural Sciences for supplying the critical oilseed materials used in this study. We thank Yangyong Zhang and Honghao Lv at Institute of Vegetables and Flowers of Chinese Academy of Agricultural Sciences for supplying the DNA from eight *B. oleracea* accessions used in this study. We thank Yuannian Jiao at Institute of Botany of the Chinese Academy of Sciences for helpful discussions about variation calling method. We thank Ray Ming at Fujian Agriculture and Forestry University and Xizhe Sun at Hebei Agriculture University for helpful discussions. We are also grateful to Richard Napier in the School of Life Sciences at University of Warwick for his insightful comments on the article and his efforts to improve the English language.

Funding:

National Natural Science Foundation of China 32222076 (WM)

National Natural Science Foundation of China 32330096 (JZ)

National Natural Science Foundation of China 32372736 (XZ)

National Natural Science Foundation of China 32402565 (YL)

Hebei Natural Science Foundation C2024204246 (JZ)

Hebei Natural Science Foundation C2023204308 (WM)

Science Research Project of Hebei Education Department YJZ2024001 (JZ)

Science Research Project of Hebei Education Department JCZX2025020 (WM)

Science Research Project of Hebei Education Department JZX2024001 (XZ)

China Agriculture Research System of MOF and MARA CARS-12 (XL)

Zhongyuan Sci-Tech Innovation Leading Talents 244200510041 (YY)

Guangdong Modern Vegetable Industry Technology System Project 2024CXTD08 (GL)

Author contributions:

Conceptualization: JZ, WM

Methodology: WM, YL

Investigation: XW, XZ, XL, ZL, LY, GL

Funding acquisition: JZ, WM, YL, XZ, XL, YY, GL

Project administration: YL, QY, XC, ZH, YG, ML, SZ, ZL, QW, XZ, QL, XS, ML, DF, YL, SL, LY

Visualization: HL, XW

Writing – original draft: YL, WM

Writing – review & editing: YH, JZ, YY, YVDP, SS, TZ, KW, AR

Competing interests:

Authors declare that they have no competing interests.

Data and materials availability:

All data supporting this study are available in the article and Supplementary Information. Genome assemblies and sequencing data of the newly assembled *B. rapa* genomes are deposited at NCBI under BioProject accession number PRJNA1297980. Scripts used to generate and analyze data are available as a Zenodo repository (*107*). Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Supplementary Materials:

Figs. S1 to S32

Tables S1 to S26

References (30, 69-72)

Science
AAAS

Supplementary Materials for

Gapless pangenome analyses reveal fast *Brassica rapa* subspeciation

Wei Ma[1]†*, Yuanming Liu[1]†, Xiaochun Wei[2]†, Xiaomeng Zhang[1]†, Xiaonan Li[3]†, Zhaokun Liu[4]†, Lingyun Yuan[5]†, Guangguang Li[6]†, Shu Zhang[1], Qihang Yang[1], Xiaocong Chang[1], Zizhuo Han[1], Hao Liang[1], Zhaoshui Luan[7], Qianyun Wang[1], Yujie Gu[1], Xinlong Wang[1], Xianlei Zhao[1], Qing Liu[1], Xiaoxue Sun[1], Mengyang Liu[1], Daling Feng[1], Yin Lu[1], Shuangxia Luo[1], Lei Yang[1], Mengyuan Li[8], Robin Allaby[9], Kai Wang[10], Tianzhen Zhang[11], Shuxing Shen[1], Yves Van de Peer[12,13,14,15]*, Yiguo Hong[1,9]*, Yuxiang Yuan[2]*, Jianjun Zhao[1]*

Corresponding author: yyzjj@hebau.edu.cn (J.Z.), yuanyuxiang@hnagri.org.cn (Y.Y.), yg.hong@hebau.edu.cn (Y.H.), yvpee@psb.vib-ugent.be (Y.V.P.), mawei0720@163.com (W.M.)

The PDF file includes:

figs. S1 to S32
References

Other Supplementary Materials for this manuscript include the following:

tables S1 to S26

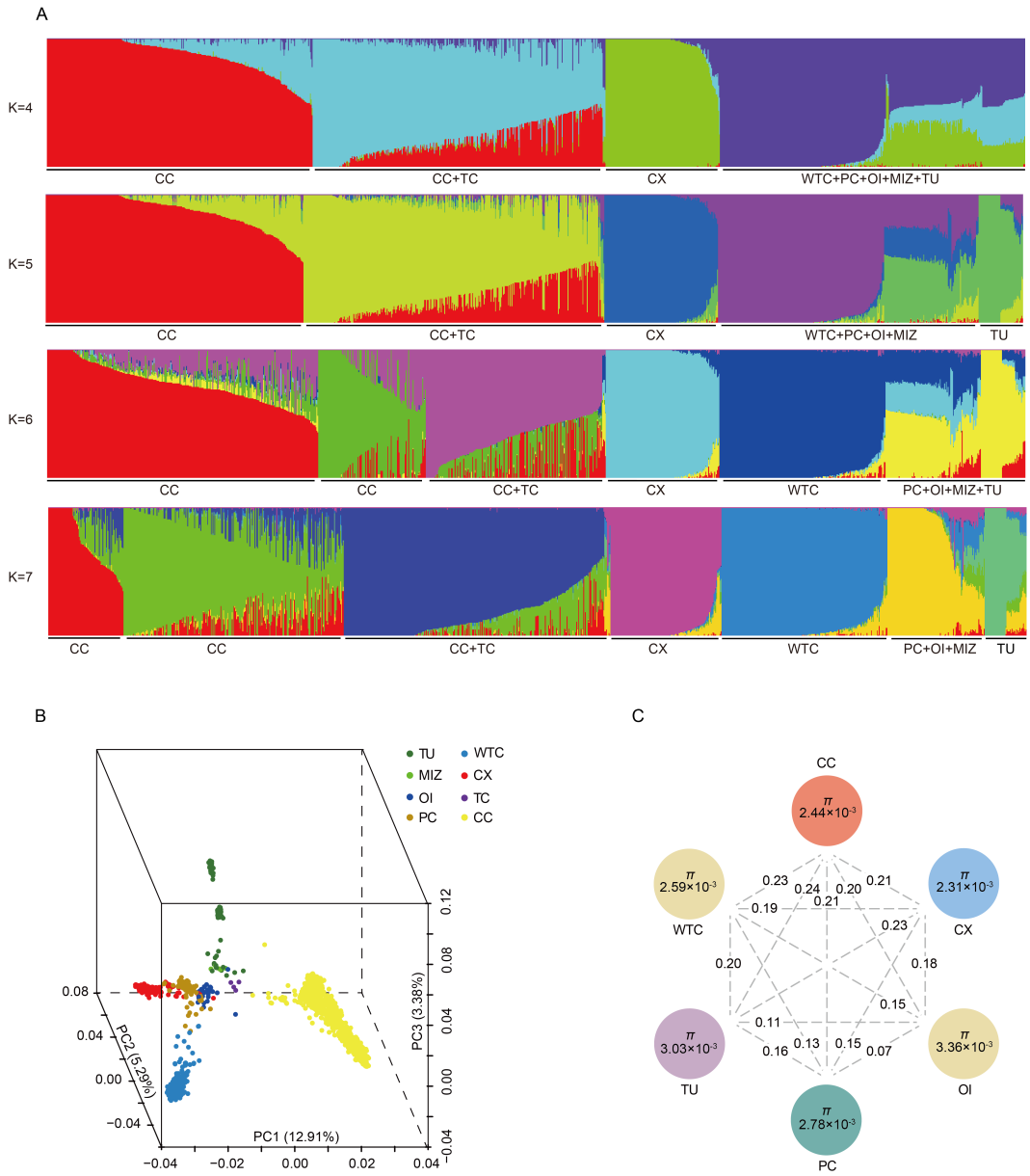


fig. S1. Population structure and genomic diversity of 1,720 *B. rapa* accessions based on SNPs.

(**A**) Structure of 1,720 *B. rapa* accessions based on SNPs. Bar-plots showing the inferred ancestral components at K = 4 to 7. Each vertical bar represents a group of accessions. Colored segments within each bar indicate the proportional contributions from different ancestral population clusters. (**B**) Three-dimensional PCA of *B. rapa* accessions based on SNPs. Each dot represents a group colored by morphotypes, including Turnip (TU), Mizuna (MIZ), Oilseed (OI), Pak choi (PC), Wutacai (WTC), Caixin (CX), Taicai (TC), and Chinese cabbage (CC). PC1, PC2, and PC3 show 12.91%, 5.29%, and 3.38% of the total genetic variation, respectively. (**C**) Nucleotide diversity (π) and population divergence (F_{ST}) across the six morphotypes (accession number >20). The value in each circle represents a measure of π for each morphotype and values on each line indicate F_{ST} between two morphotypes.

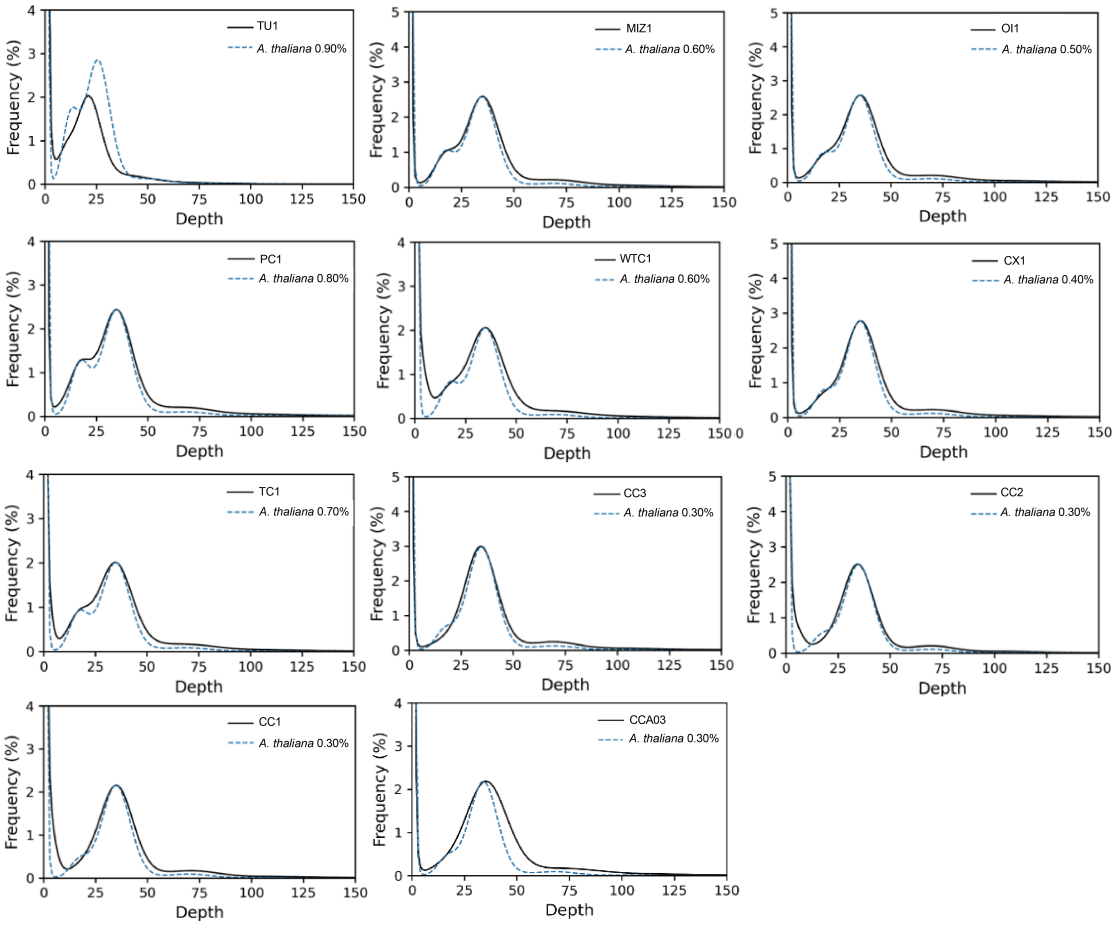


fig. S2. K-mer depth-frequence distribution in 11 *B. rapa* genomes.
Each graph represents one of 11 gapless genome assemblies, including Turnip (TU1), Mizuna (MIZ1), Oilseed (OI1), Pak choi (PC1), Wutacai (WTC1), Caixin (CX1), Taicai (TC1) and Chinese cabbage (CC1, CC2, CC3, CCA03). X- and Y-axis show k-mer (k=21) depth and k-mer frequency, respectively. Heterozygosity values below 0.3% were set to 0.3%.

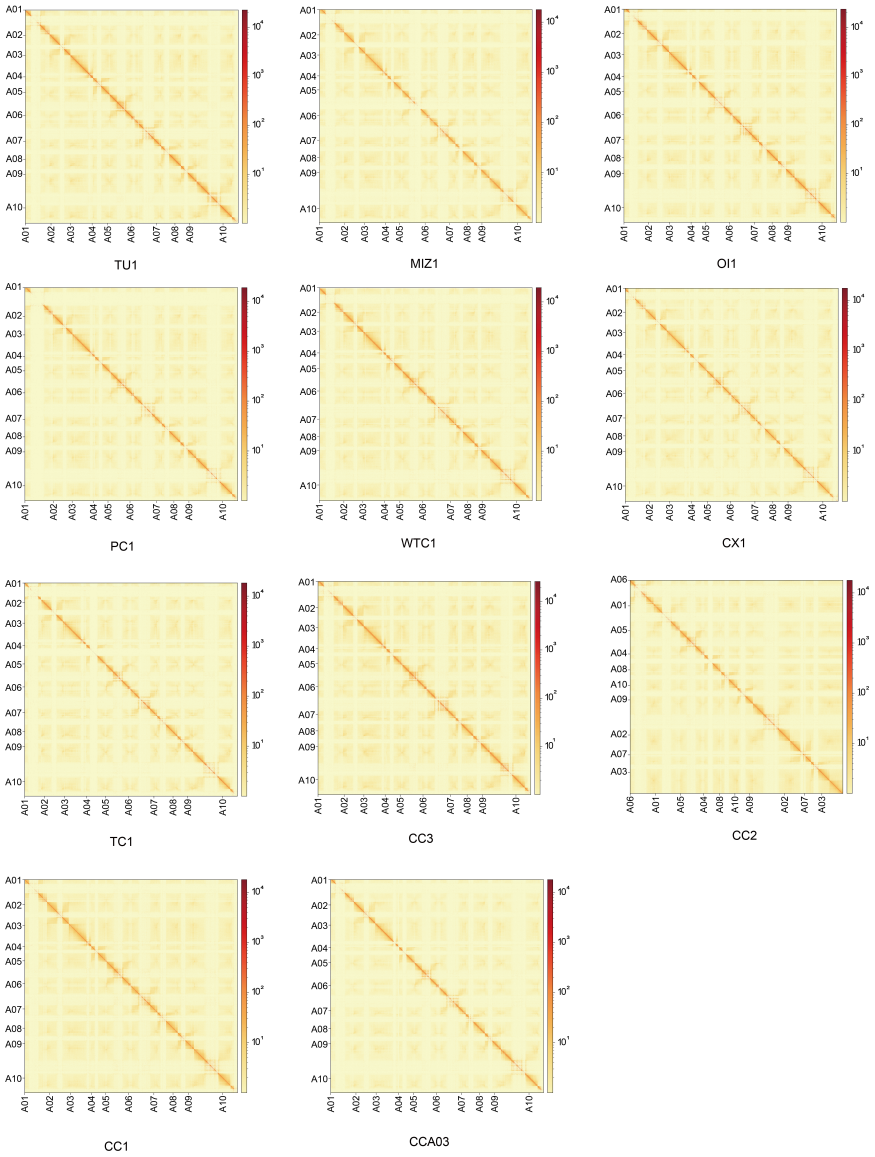
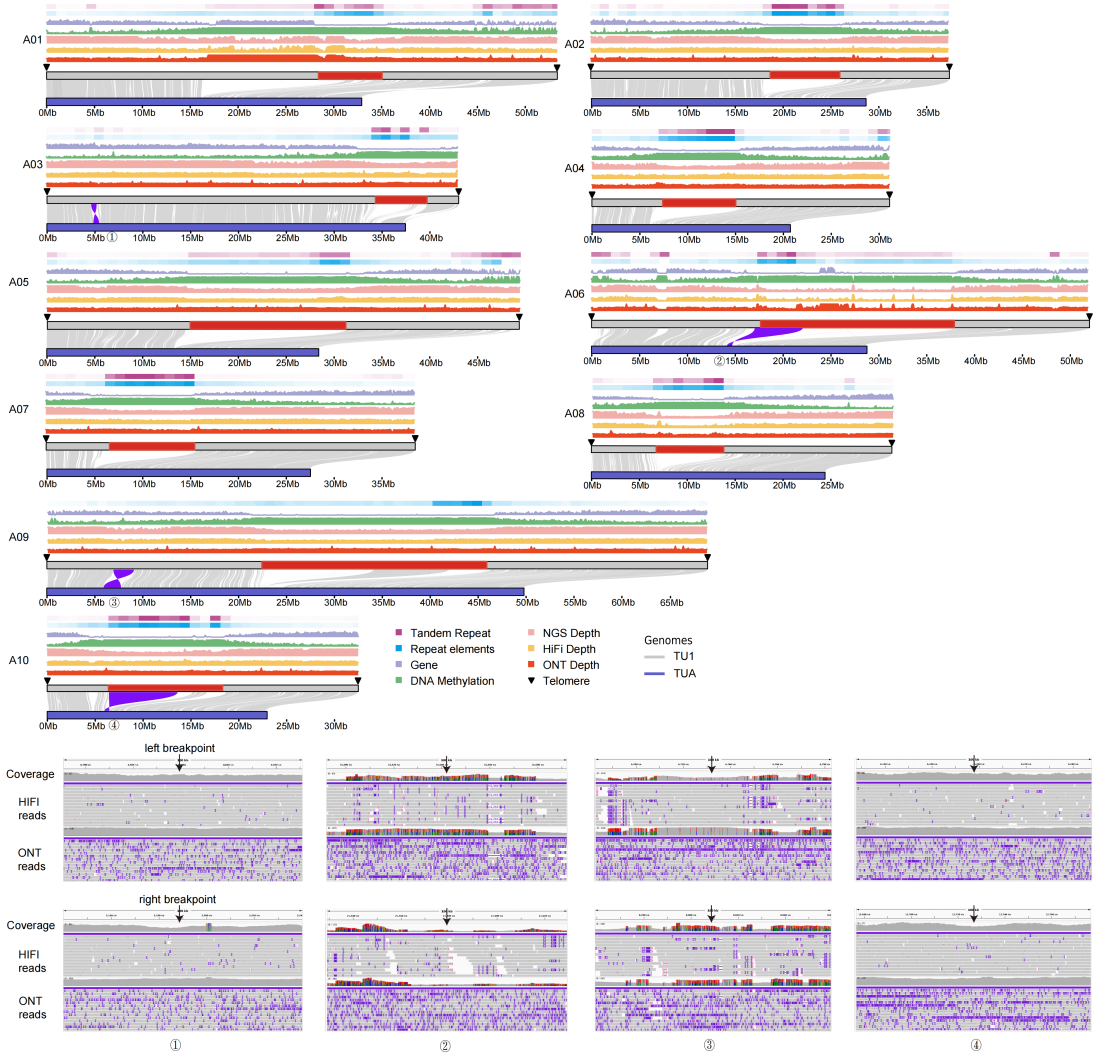


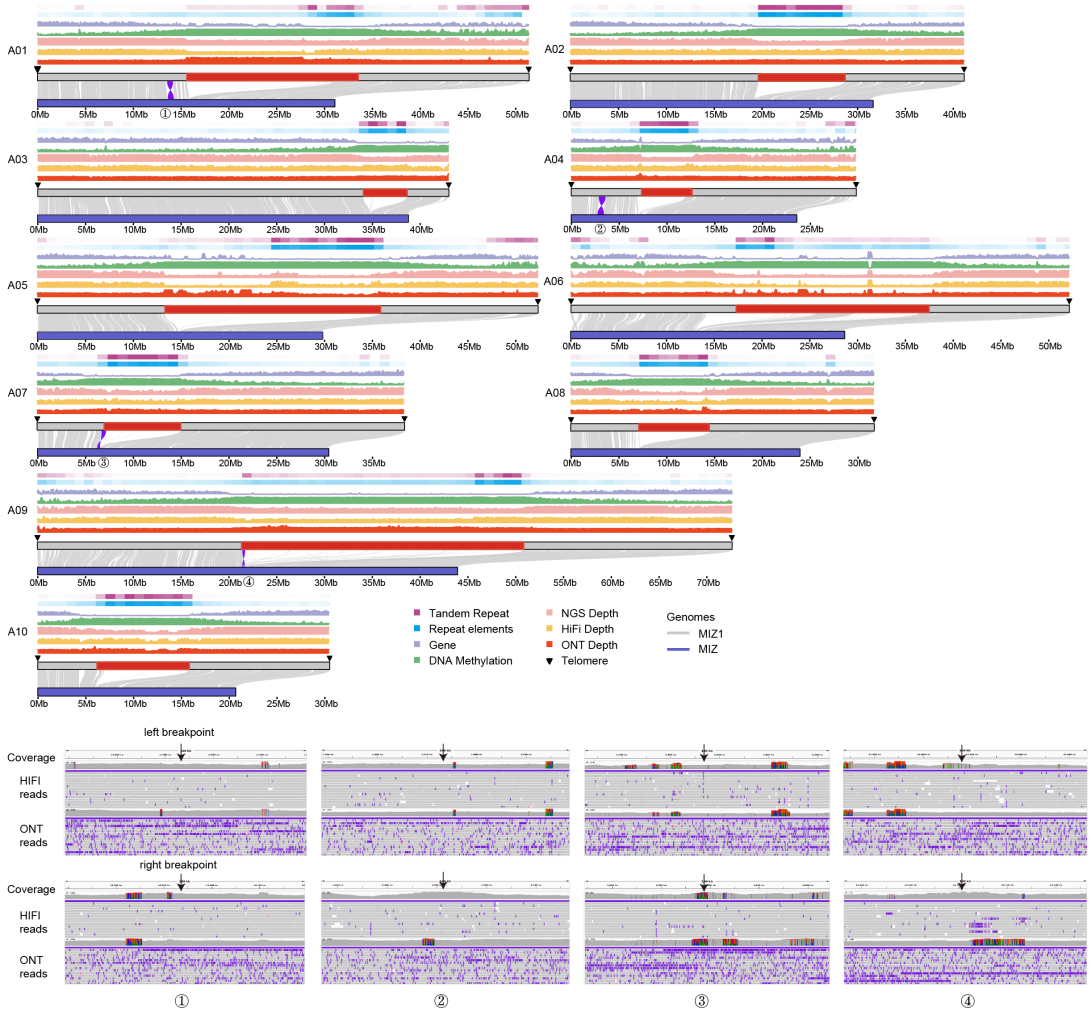
fig. S3. Interaction heatmap of 11 *B. rapa* genomes at 100-kb resolution based on Hi-C analysis.

The heatmaps visualize the Hi-C contact matrices at a resolution of 100-kb window, providing a comprehensive overview of the chromatin interactions across the genomes. Within each accession, interactions along the diagonal exhibit higher intensity compared to off-diagonal positions, indicating that in the Hi-C assembled chromosomes, proximal sequences (diagonal positions) have stronger interaction frequencies while distal sequences (off-diagonal positions) display weaker interaction signals. The level of Hi-C interaction strength is indicated by the color scale. This observation aligns with the principles of Hi-C-assisted genome assembly. The absence of significant noise (strong interaction intensities) outside the diagonal regions further substantiates the high quality of the genome assembly. Each graph represents one of 11 assemblies, including Turnip (TU1), Mizuna (MIZ1), Oilseed (OI1), Pak choi (PC1), Wutacai (WTC1), Caixin (CX1), Taicai (TC1) and Chinese cabbage (CC1, CC2, CC3, CCA03).

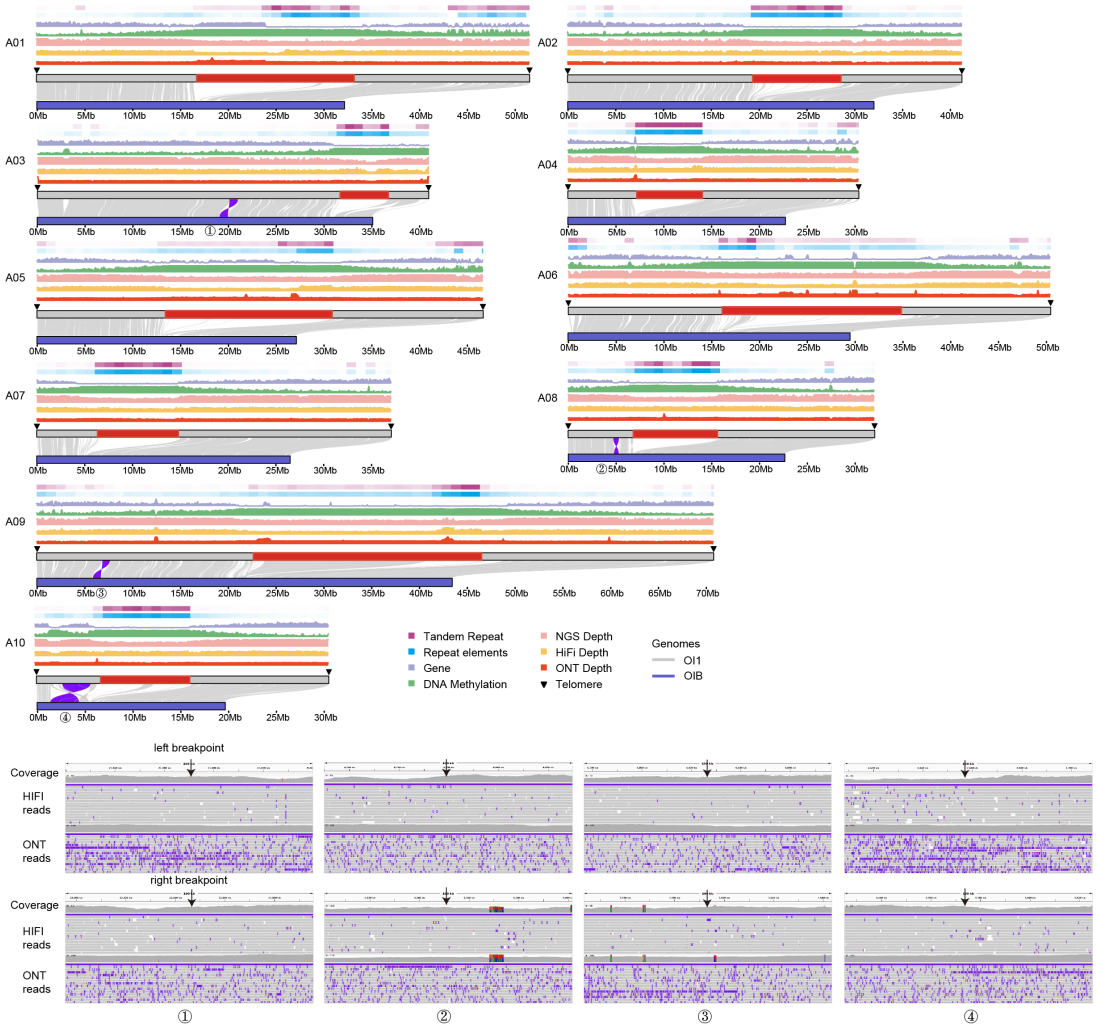
A Comprehensive map on chromosomes A01-A10 of TU1



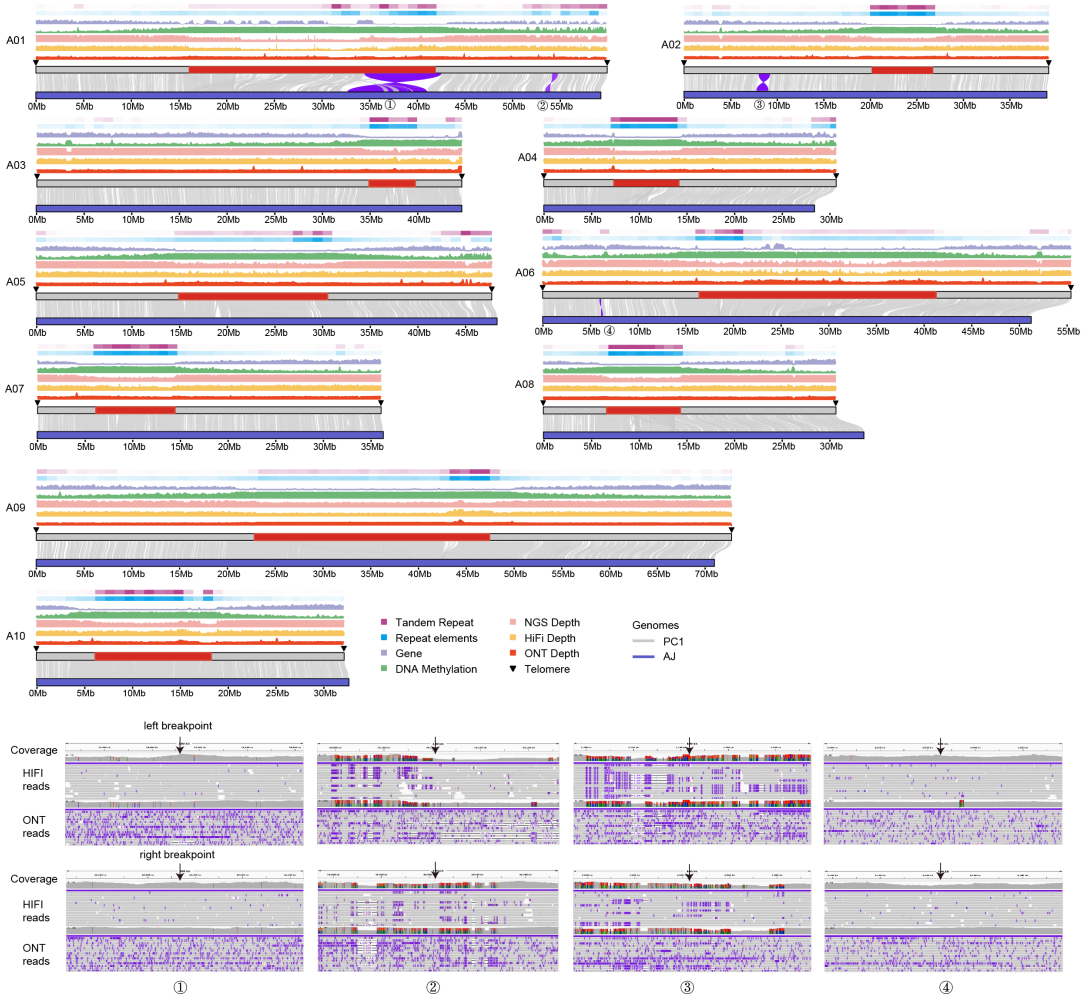
B Comprehensive map on chromosomes A01-A10 of MIZ1



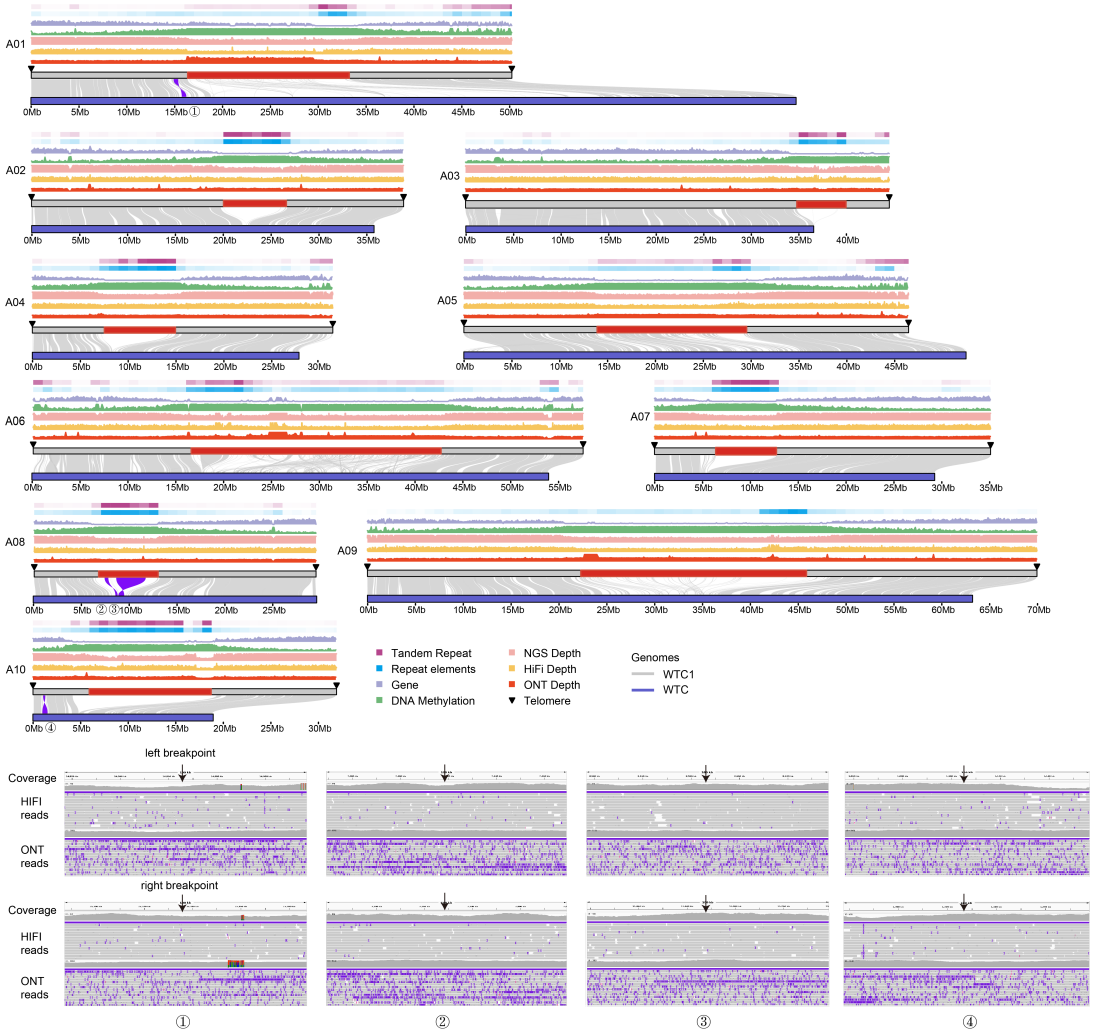
C Comprehensive map on chromosomes A01-A10 of OI1



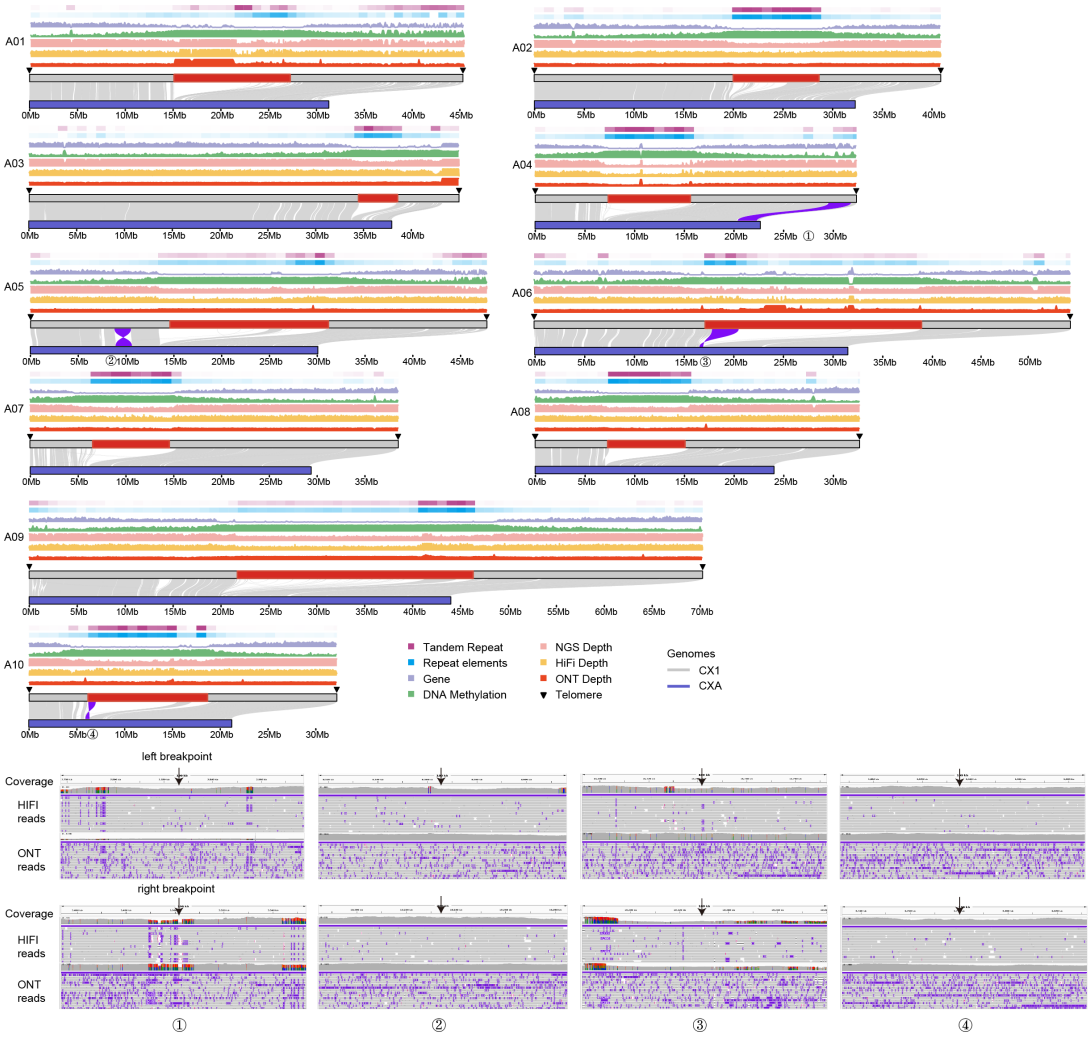
D Comprehensive map on chromosomes A01-A10 of PC1



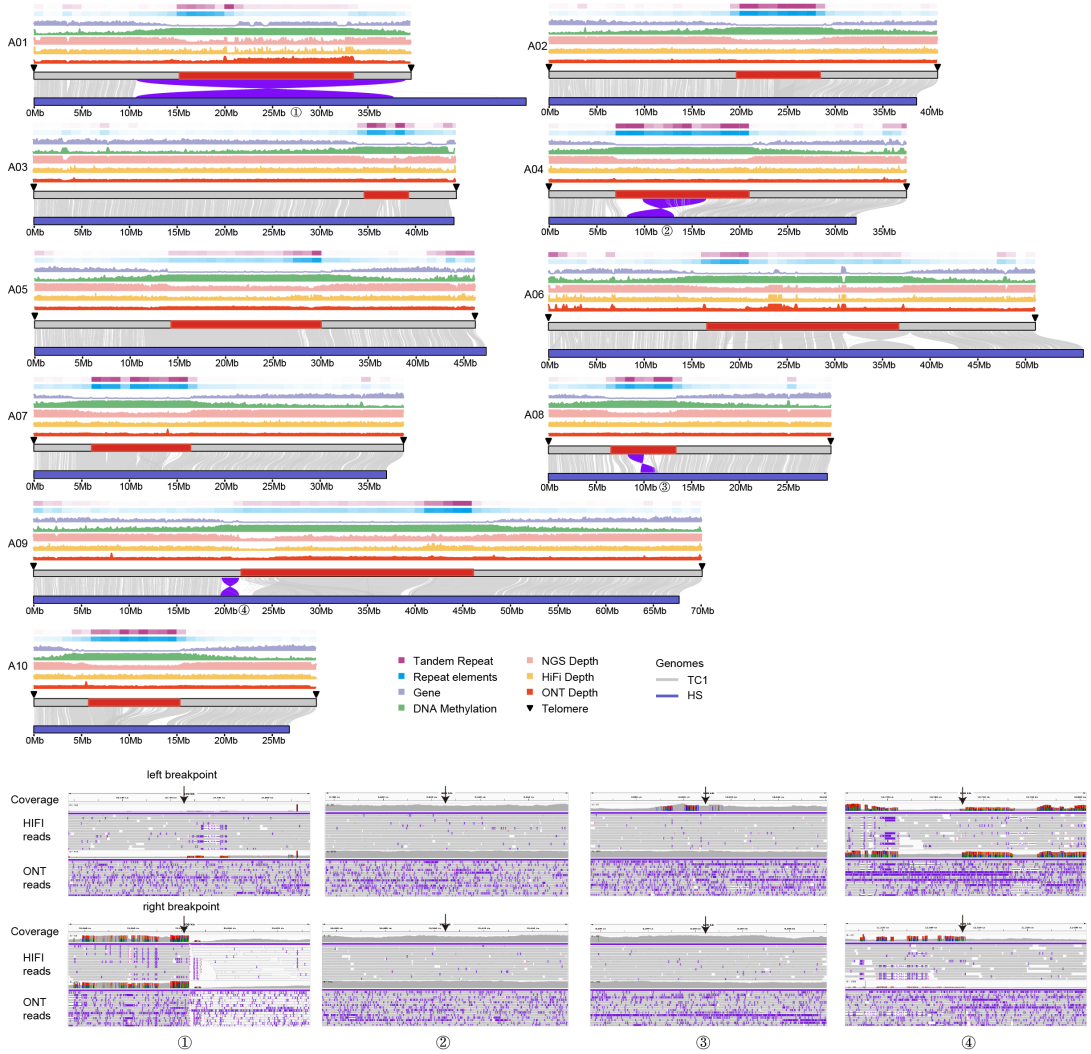
E Comprehensive map on chromosomes A01-A10 of WTC1



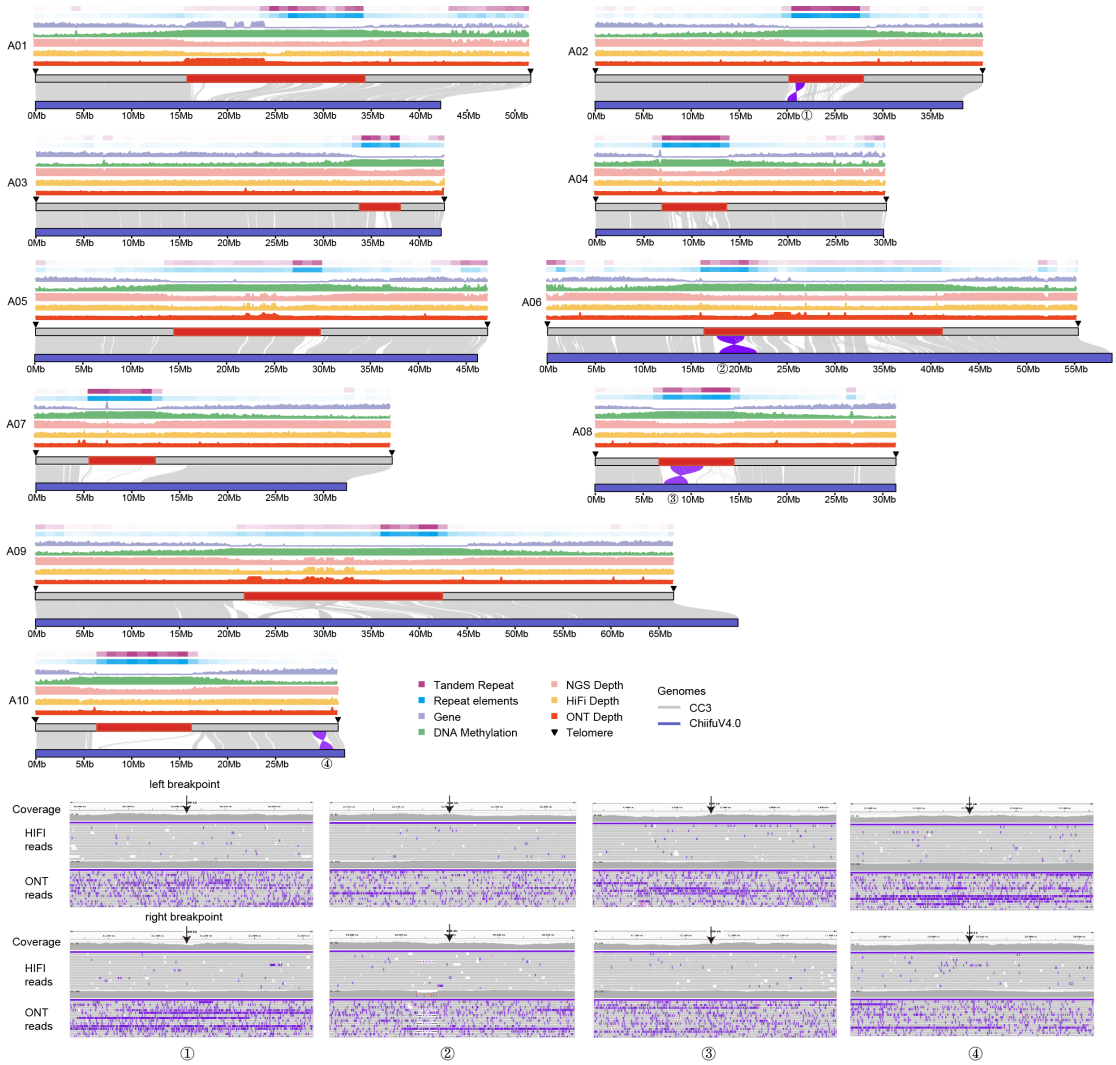
F Comprehensive map on chromosomes A01-A10 of CX1



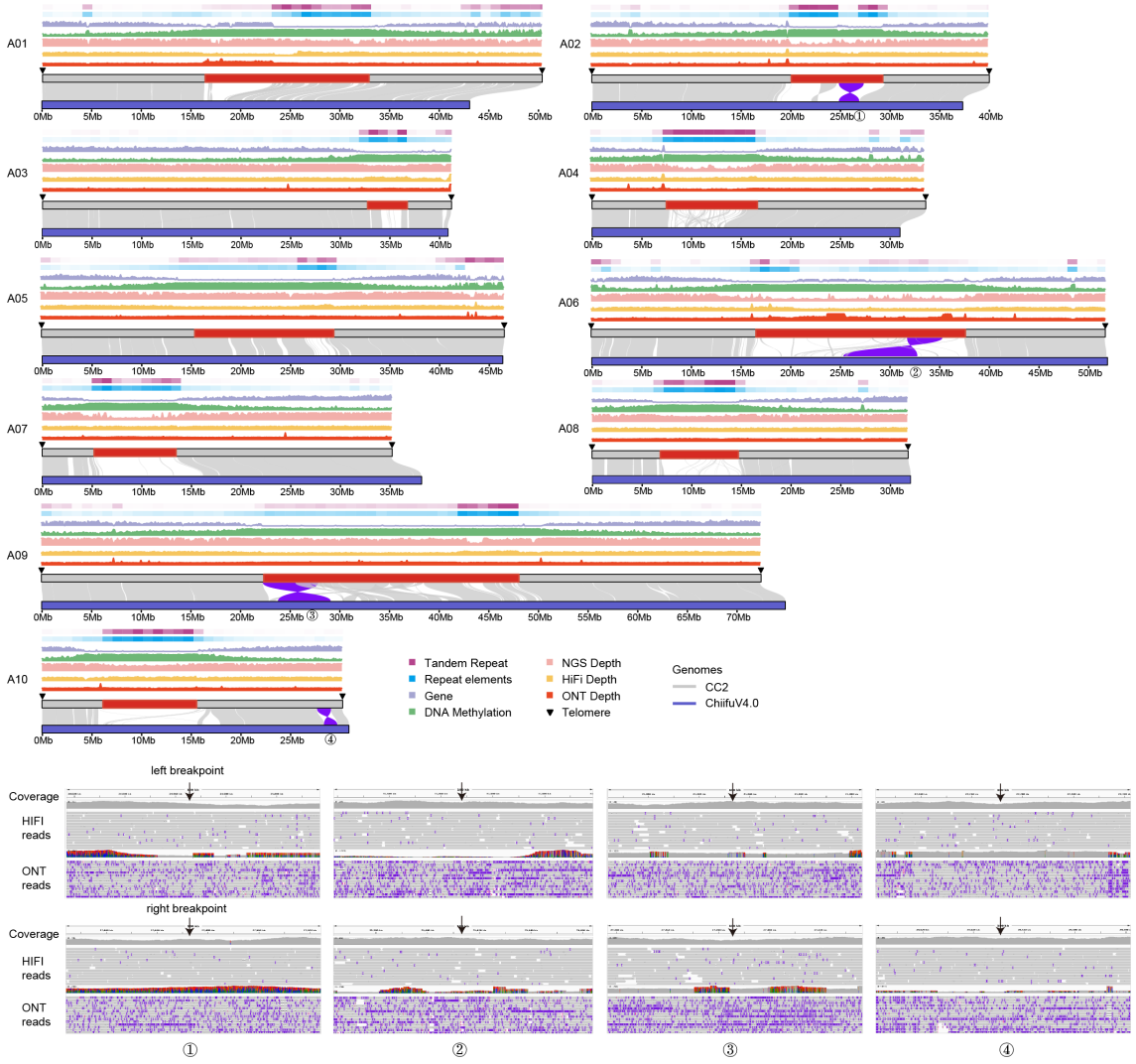
G Comprehensive map on chromosomes A01-A10 of TC1



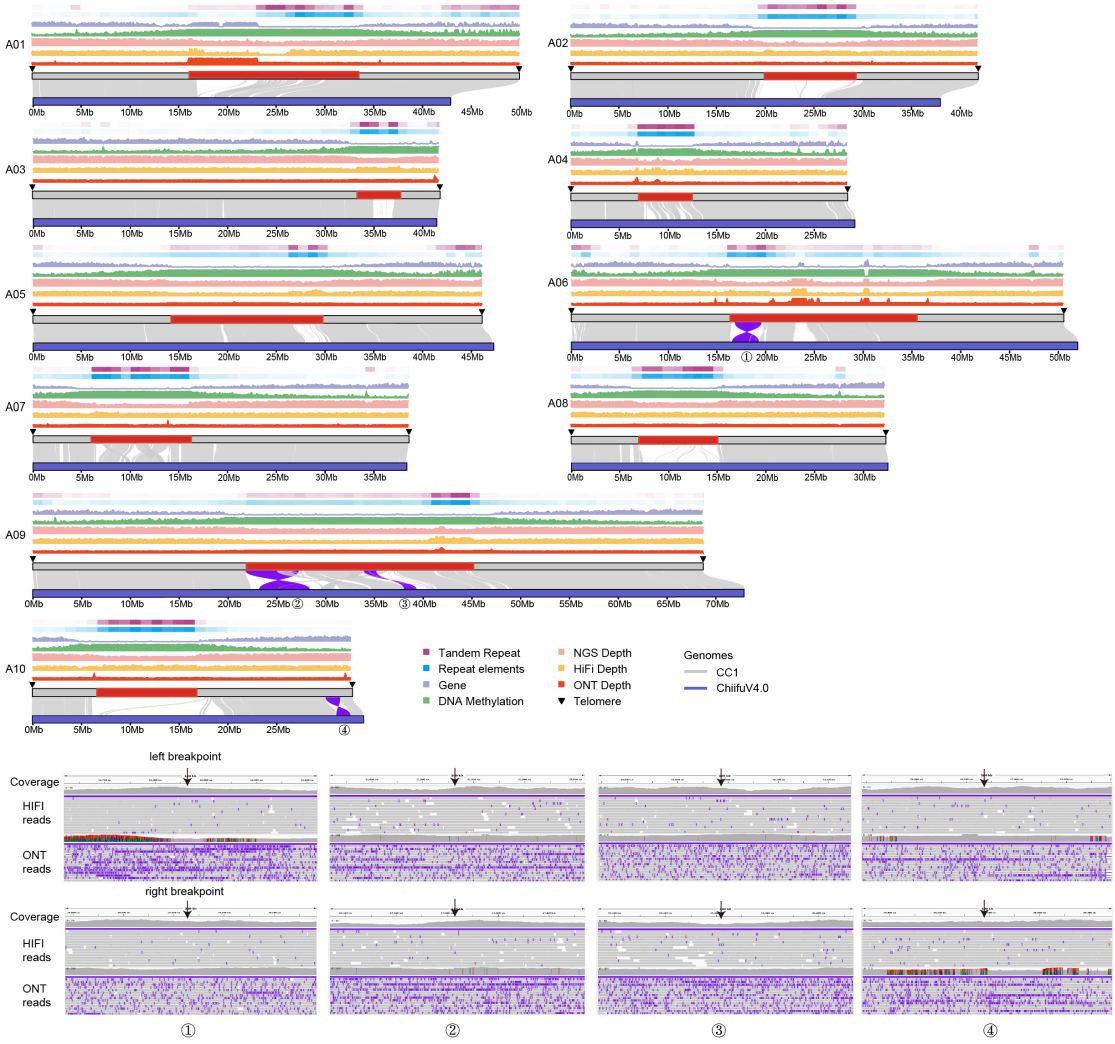
H Comprehensive map on chromosomes A01-A10 of CC3



I Comprehensive map on chromosomes A01-A10 of CC2



J Comprehensive map on chromosomes A01-A10 of CC1



K Comprehensive map on chromosomes A01-A10 of CCA03

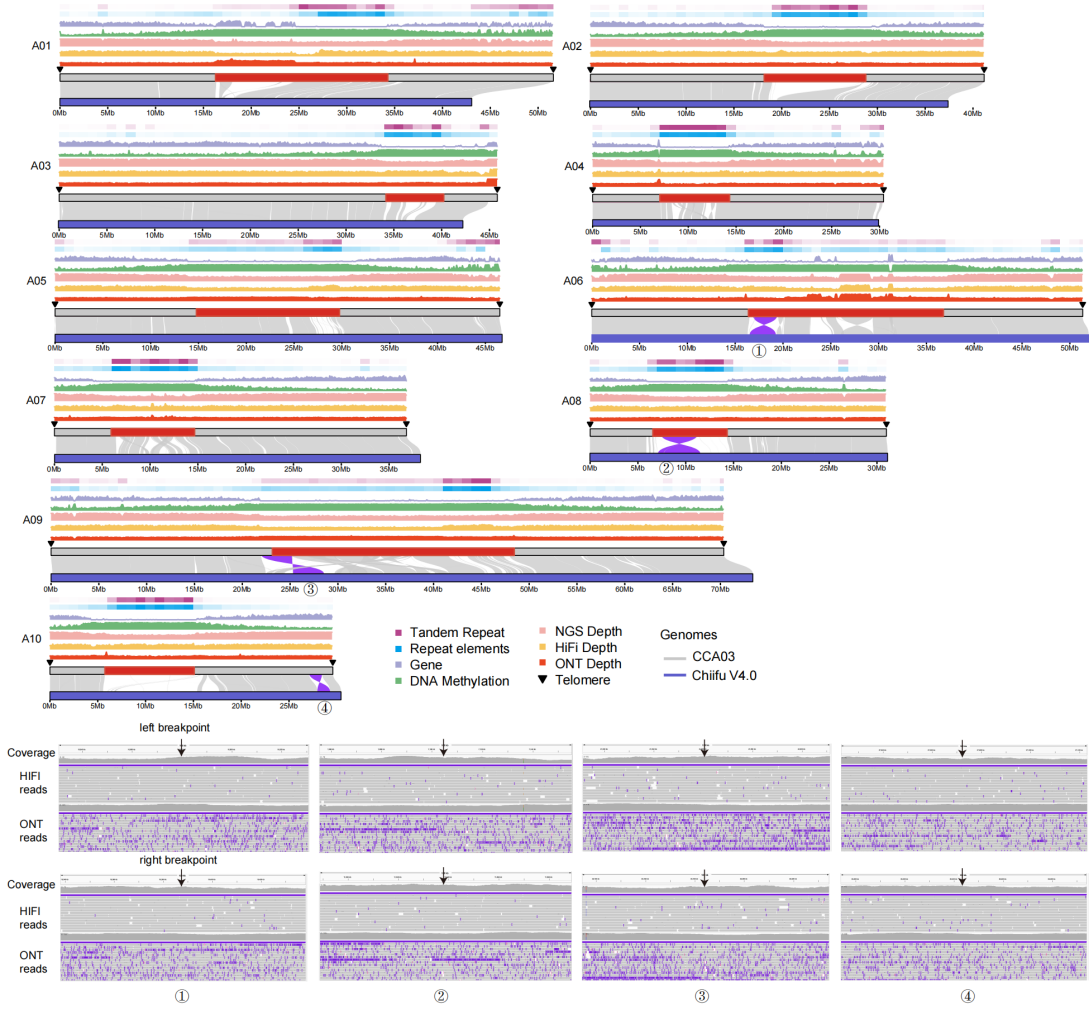


fig. S4. Comprehensive map on chromosomes of 11 *B. rapa* genomes.
(**A**-**K**) Genome-wide distribution of Tandem Repeats Density, Repeat elements Density, Gene Density, DNA Methylation Density, NGS Depth, HiFi Depth, ONT Depth, Telomere (solid triangle) from top to bottom in chromosomes A01-A10 of each accession as indicated. The gray regions in each of the chromosome map shows the syntenic alignment between CCA03 and the latest Chinese cabbage reference genome Chiifu V4.0. The two IGV graphs in panels A-K show manual validation of four large inversions marked ① to ④ on relevant chromosome(s), based on mapping the ONT reads and HIFI reads to the CCA03 genome assembly. Accessions include TU1 (**A**), MIZ1 (**B**), OI1 (**C**), PC1 (**D**), WTC1 (**E**), CX1 (**F**), TC1 (**G**), CC3 (**H**), CC2 (**I**), CC1 (**J**), and CCA03 (**K**).

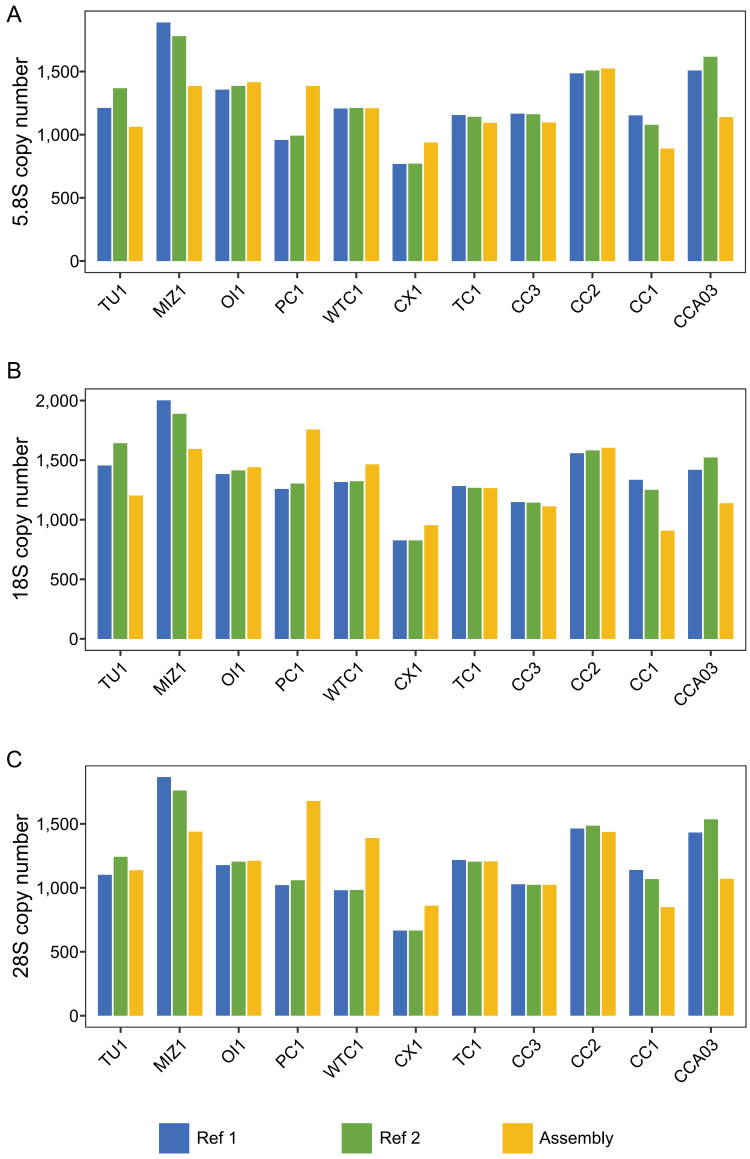


fig. S5. Copy number of 45S rDNAs by digital PCR in 11 *B. rapa* genomes.
Comparison of the copy number of (**A**) 5.8S, (**B**) 18S, (**C**) 28S rDNAs in the assemblies. Copy number was estimated from data of digital PCR-based assays with two single copy genes as reference (Ref 1 and Ref 2). Information about Ref 1 and Ref 2 is presented in table S26.

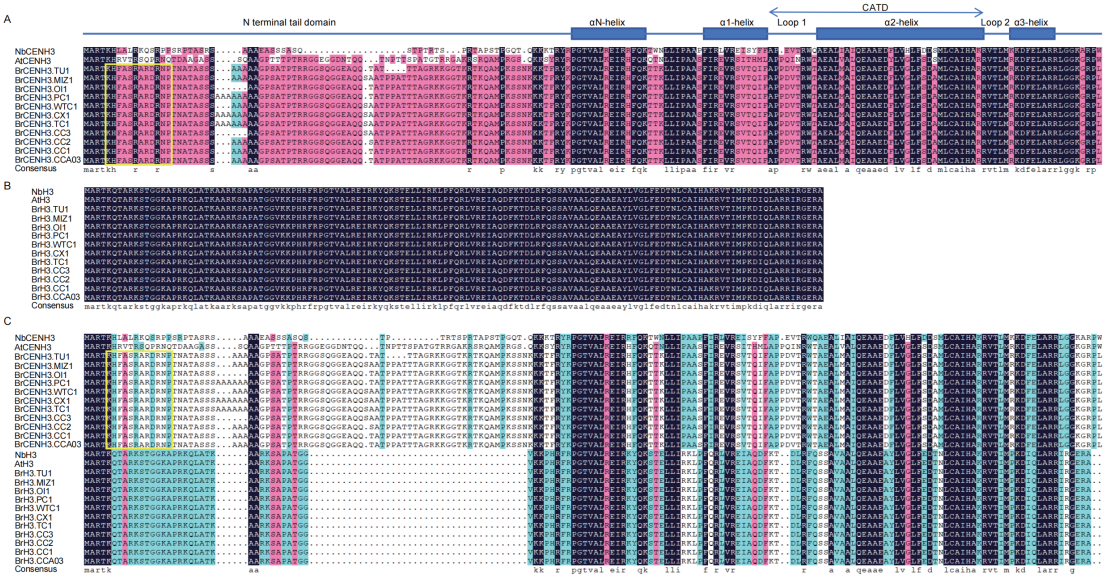


fig. S6. Antigenic peptide designing for production of anti-BrCENH3 antibody. (**A**) Multiple amino-acid sequence alignment of CENH3 proteins. Domains and structures are shown. (**B**) Multiple amino-acid sequence alignment of H3 proteins. (**C**) Comparisons among H3 and CENH3s. The BrCENH3-specific dodecapeptide used as an antigen to generate anti-BrCENH3 antibody is boxed in panel A and C.

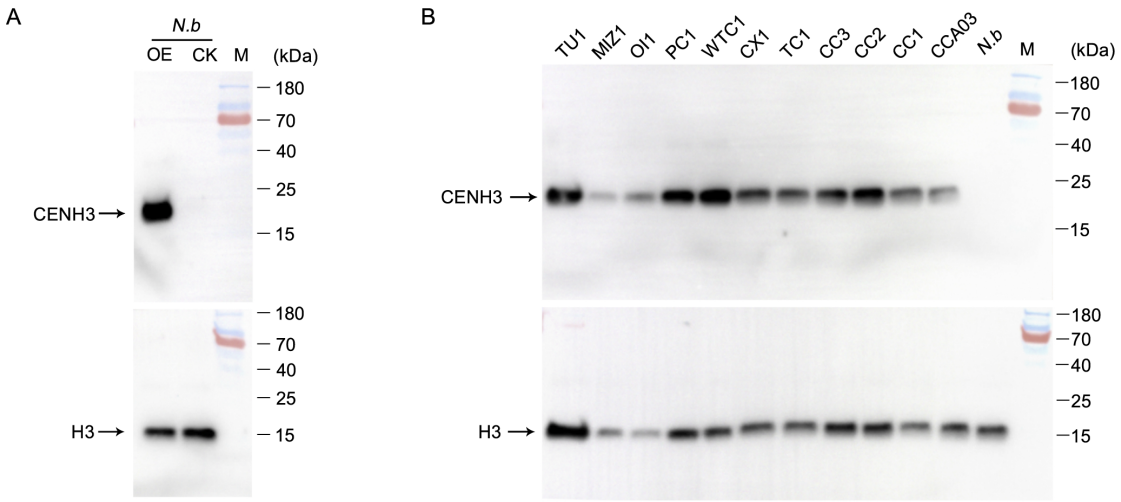


fig. S7. Western blot detection of BrCENH3.

(**A**) Transient assay. A CaMV 35S promoter-controlled BraCENH3 expression cassette 35S::BrCENH3 was constructed and cloned into *Agrobacterium tumefaciens* GV3101. Young leaves of *Nicotinana benthimiana* (*N.b*) were agroinfiltrated with *A. tumefaciens* GV3101 carrying 35S:BrCENH3. At 4 days post-agroinfiltration, total protein was extracted from these infiltrated leaf tissues (OE) and analyzed by western blot using the anti-BrCENH3 (upper panel) or anti-H3 (Lower panel) antibody. Protein samples extracted from leaf tissues agroinfiltrated with *A. tumefaciens* GV3101 carrying the empty gene expression vector was used as the negative control (CK). (**B**) Western detection of BrCENH3s and BrH3s in *B. rapa*. Total protein was extracted from leaf tissues of the eleven *B. rapa* subspecies/morphotypes as well as *N. benthimiana* (*N.b*) and analyzed by western blot using the anti-BrCENH3 (upper panel) or anti-H3 (Lower panel) antibody. Sizes (kDa) and positions of protein markers and positions of CENH3 and H3 are indicated.

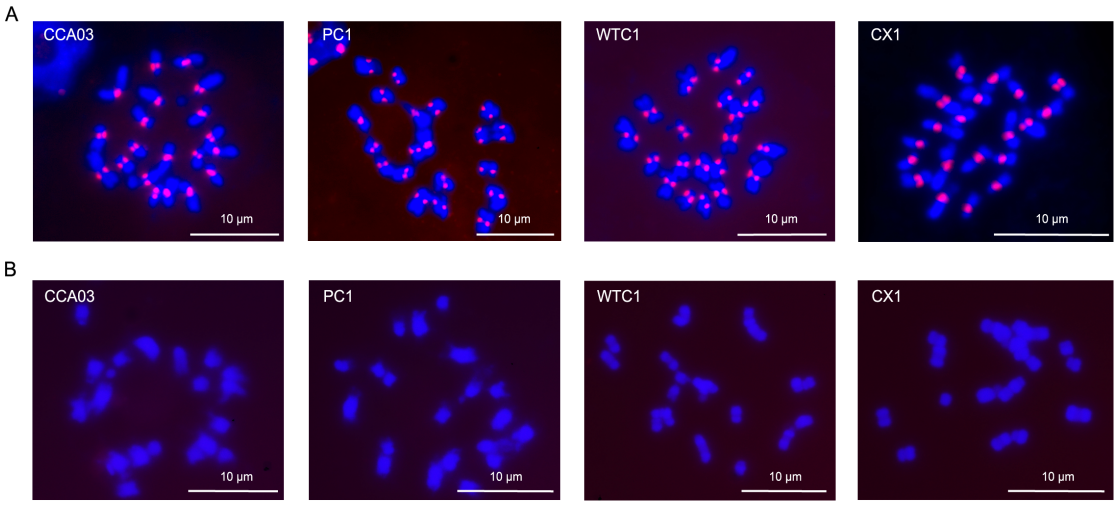
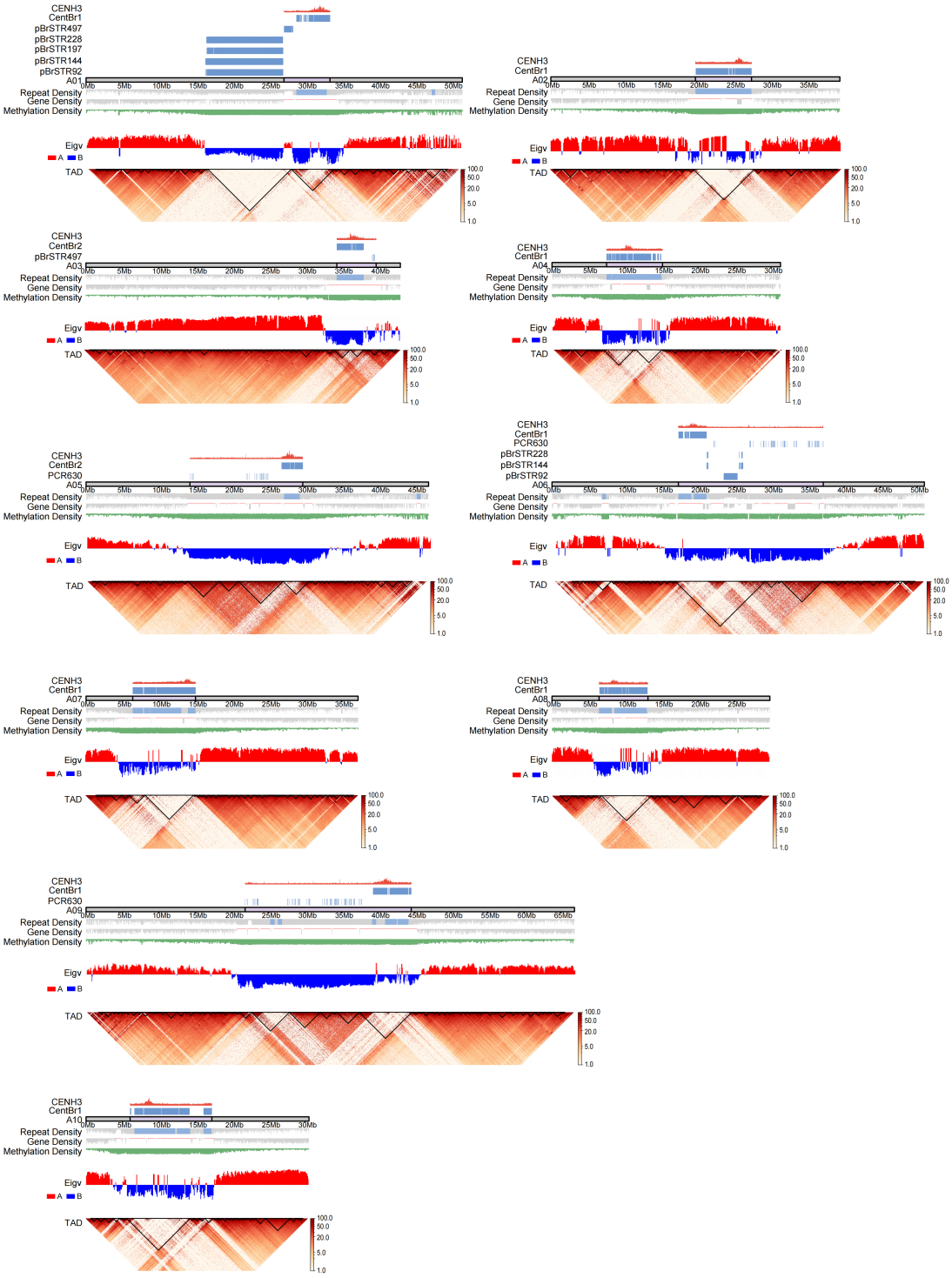
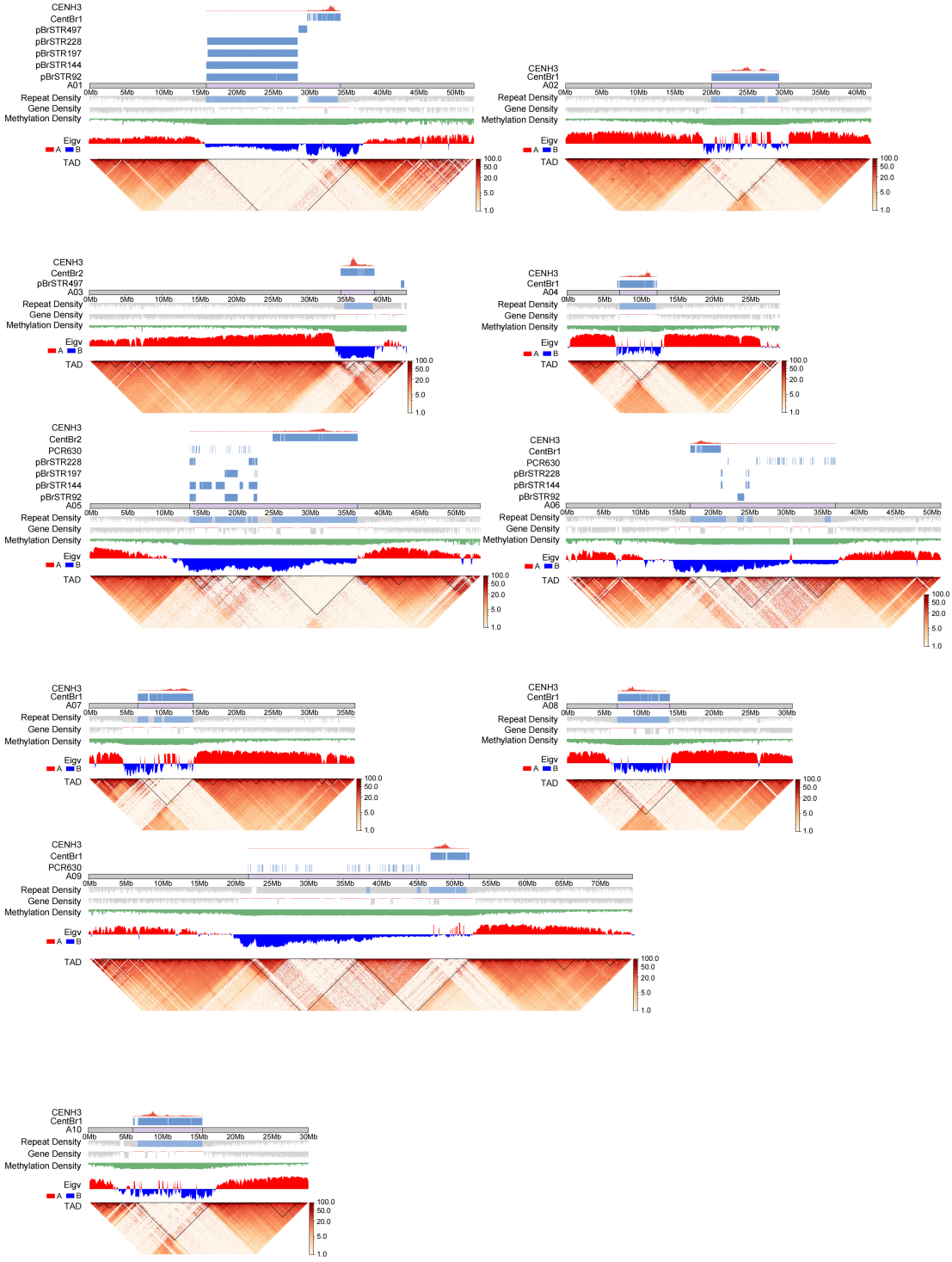


fig. S8. Immunostaining of chromosomes with anti-BrCENH3 antibodies.
Immunostaining was performed on sections of CCA03, PC1, WTC1 and CX1 root tips with the antibody specifically raised against BrCENH3. Chromosomes at metaphase (blue) were stained with 4′,6-diamidino-2-phenylindole (DAPI). Red fluorescence signals are visible at the primary constrictions of the chromosomes stained with the anti-BrCENH3 antibody (**A**), but not with the antisera purified from rabbits immunized with the carrier protein alone (**B**). Scale bar = 10 um.

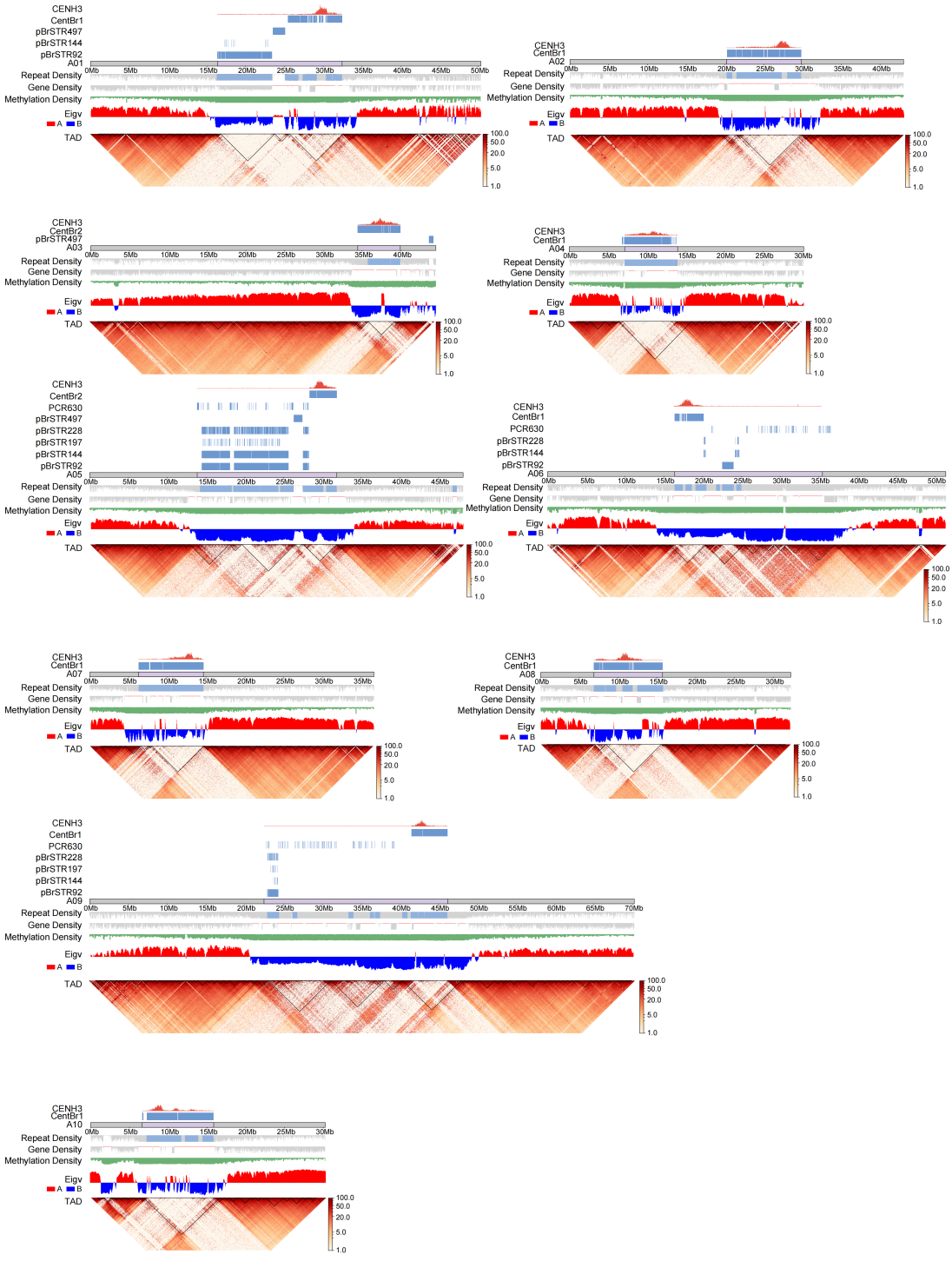
A Characterization of the (peri)centromeres A01-A10 in TU1



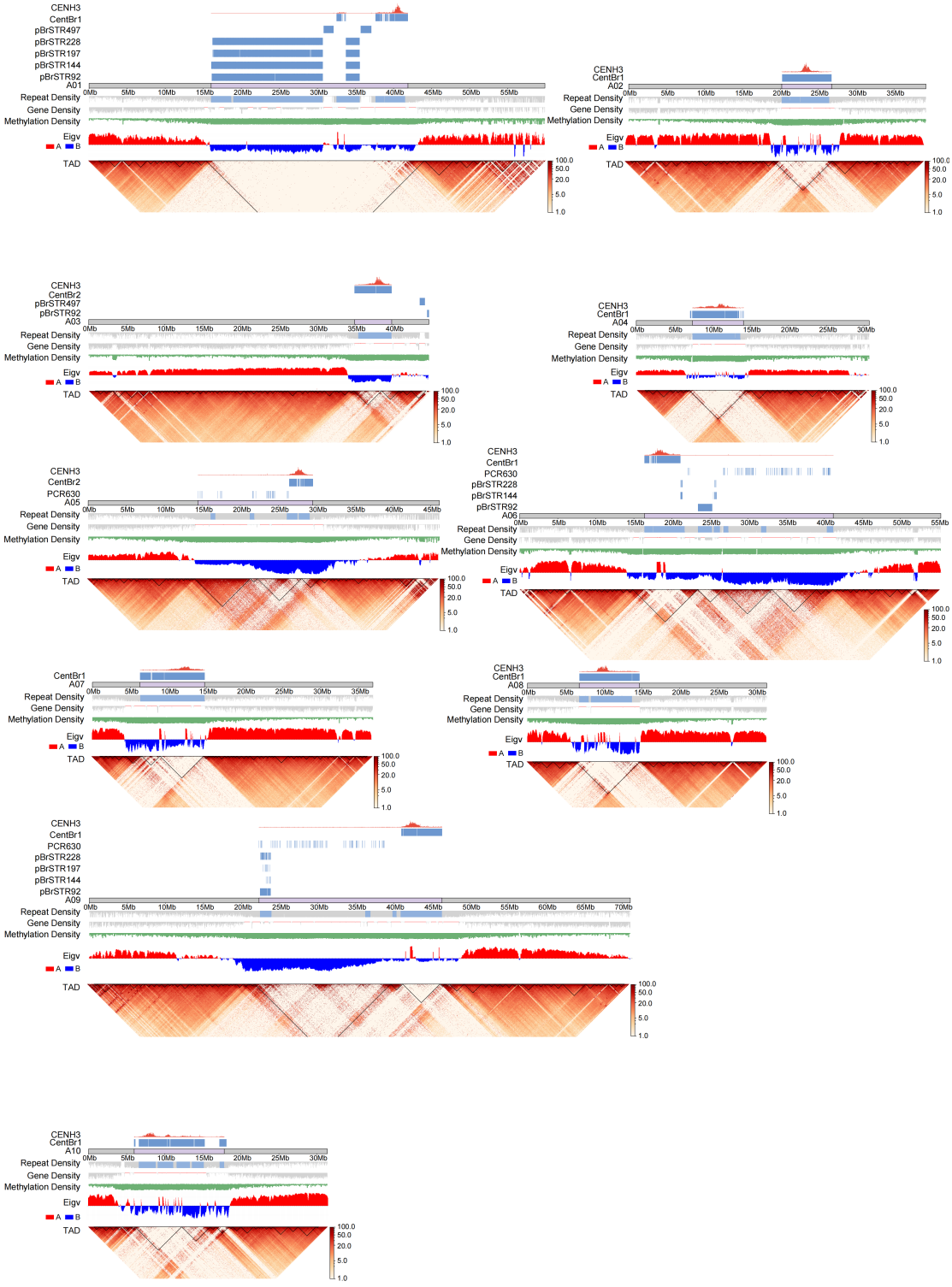
B Characterization of the (peri)centromeres A01-A10 in MIZ1



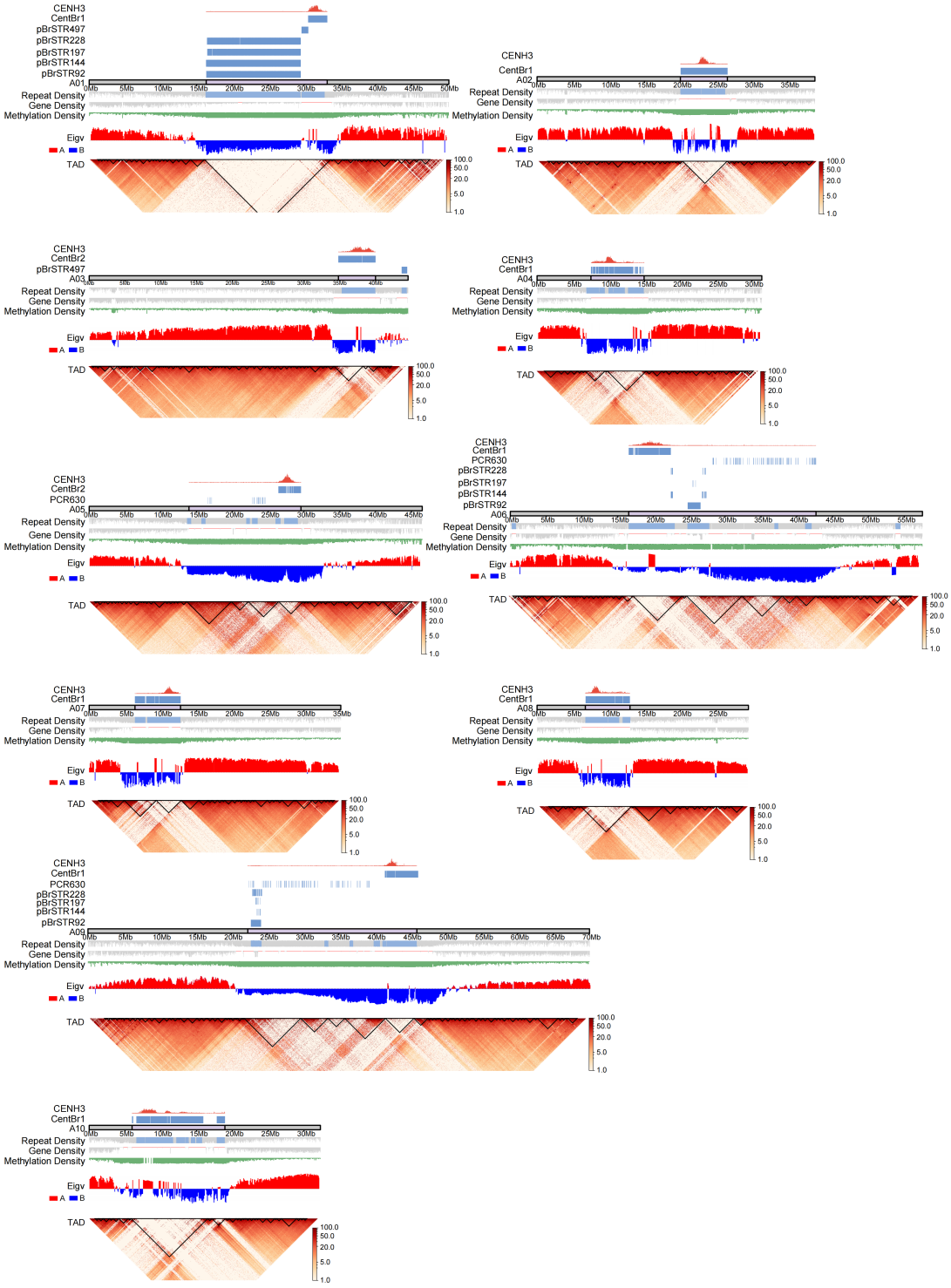
C Characterization of the (peri)centromeres A01-A10 in OI1



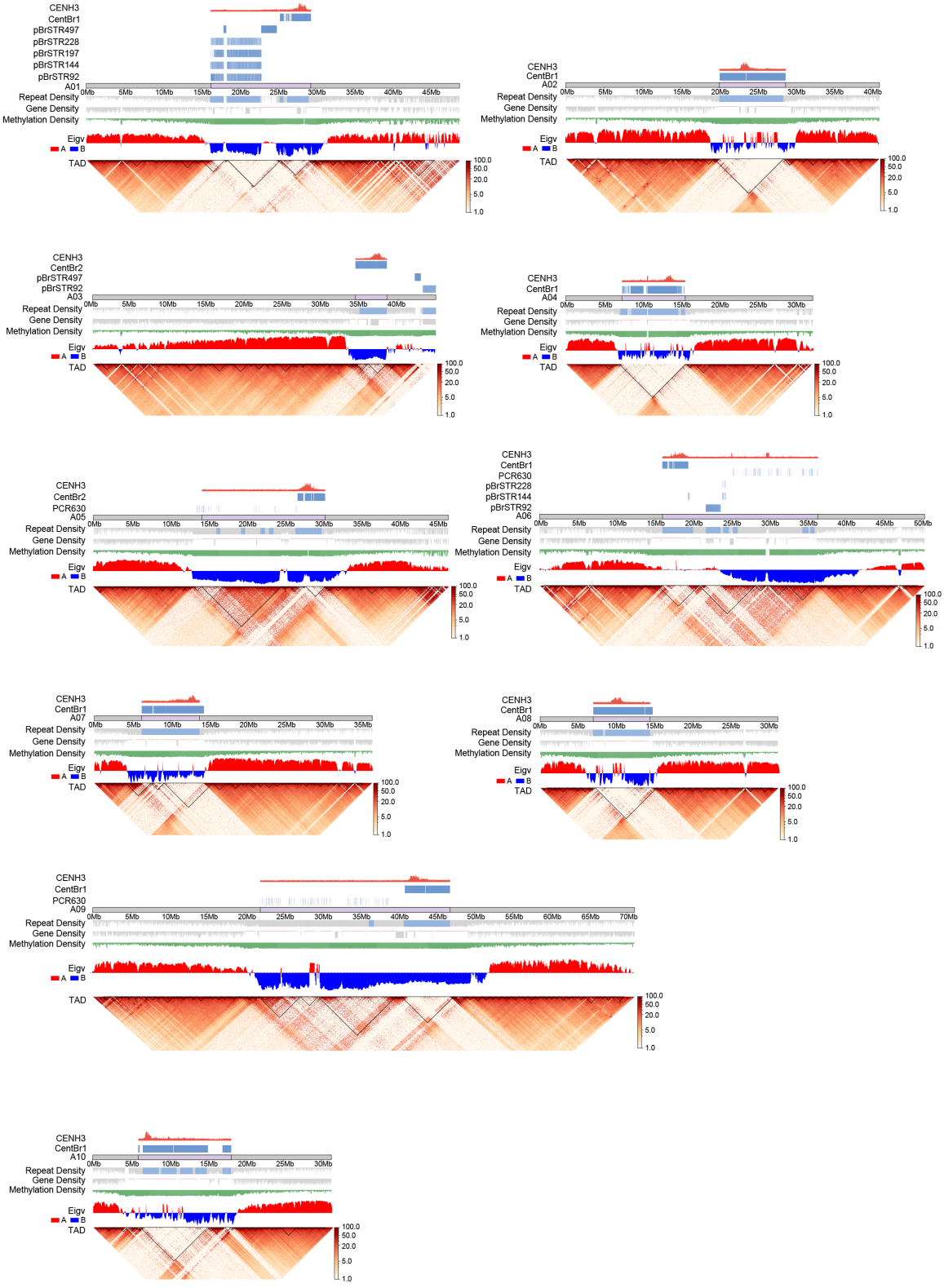
D Characterization of the (peri)centromeres A01-A10 in PC1



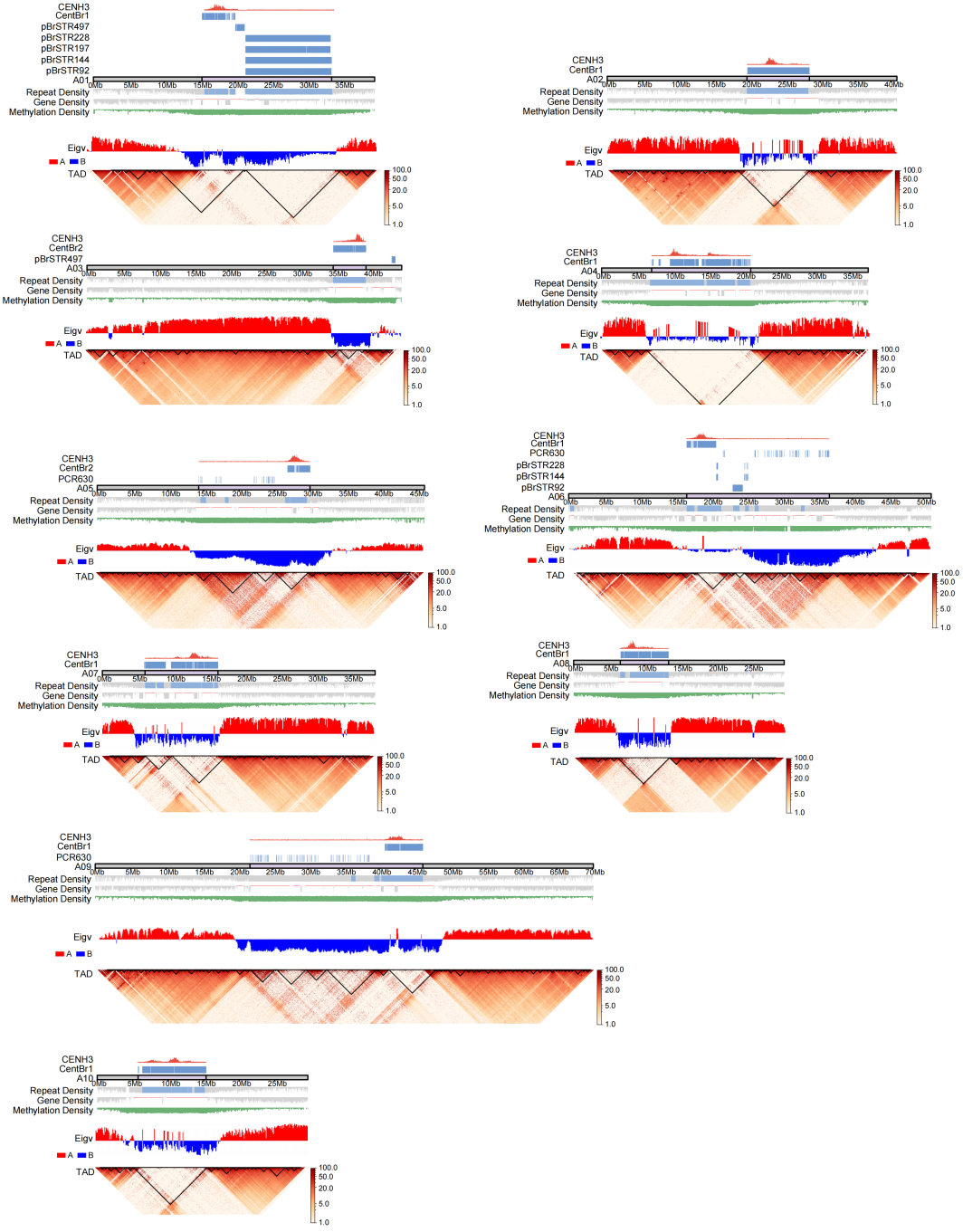
E Characterization of the (peri)centromeres in A01-A10 WTC1



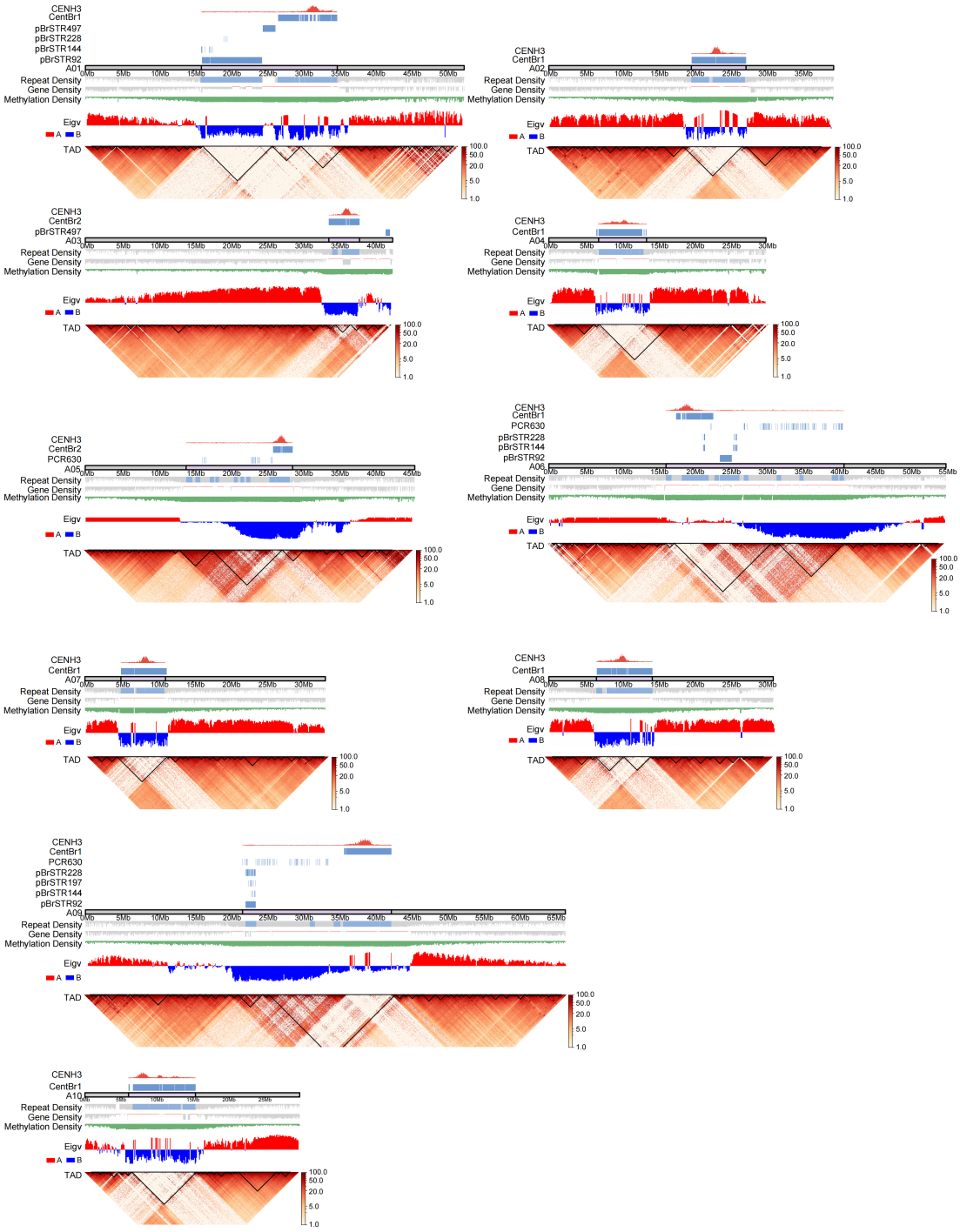
F Characterization of the (peri)centromeres A01-A10 in CX1



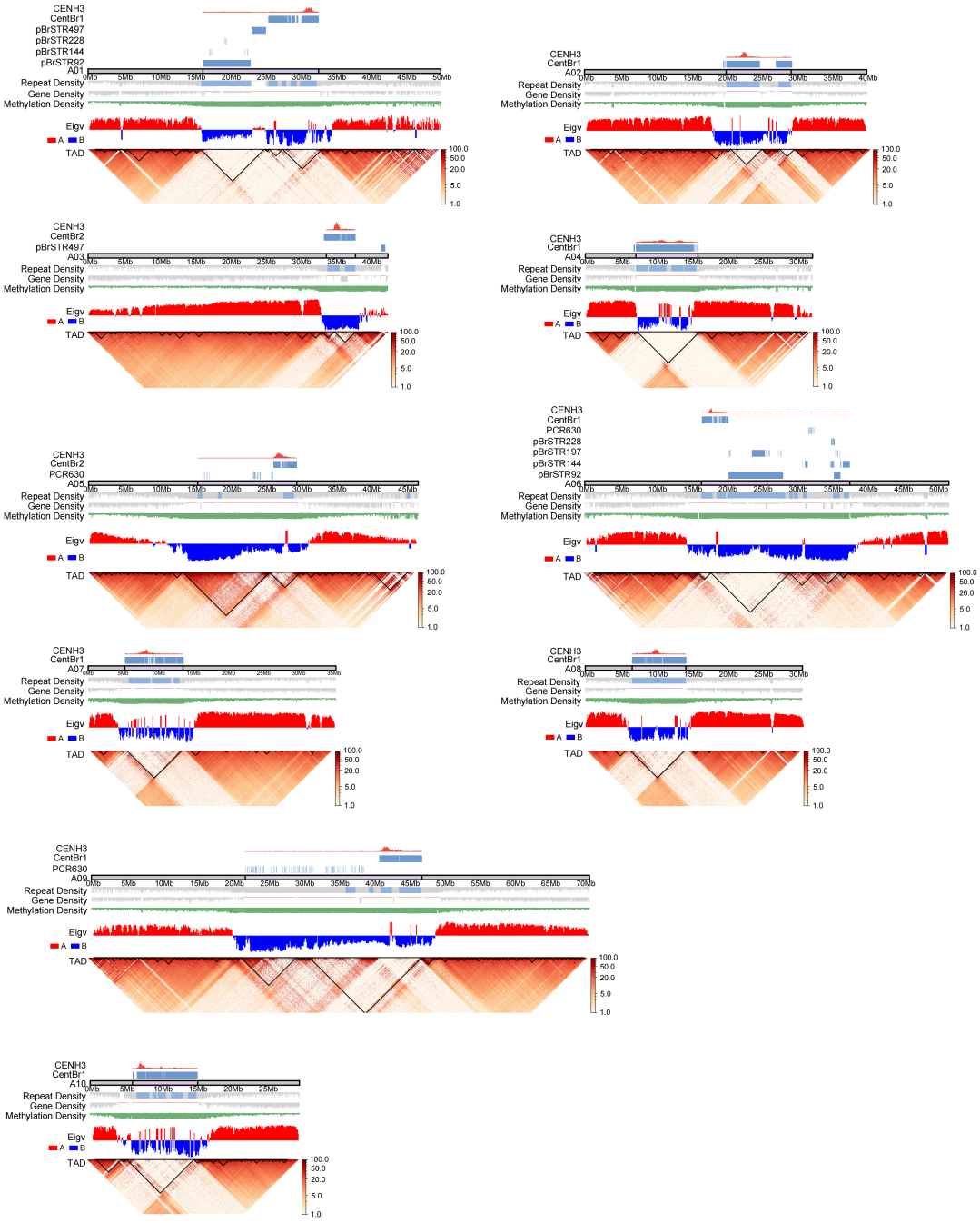
G Characterization of the (peri)centromeres A01-A10 in TC1



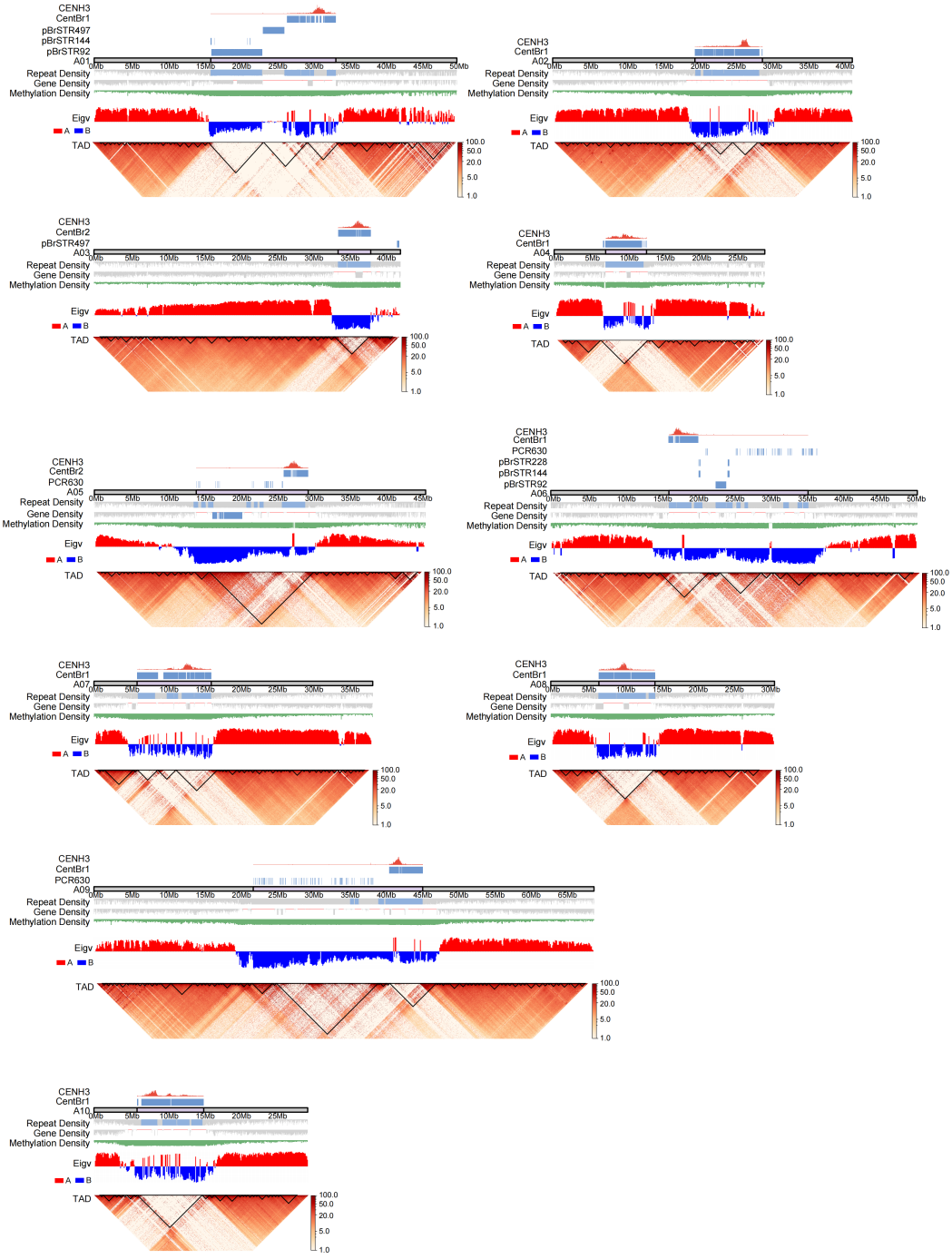
H Characterization of the (peri)centromeres A01-A10 in CC3



I Characterization of the (peri)centromeres A01-A10 in CC2



J Characterization of the (peri)centromeres A01-A10 in CC1



K Characterization of the (peri)centromeres A01-A10 in CCA03

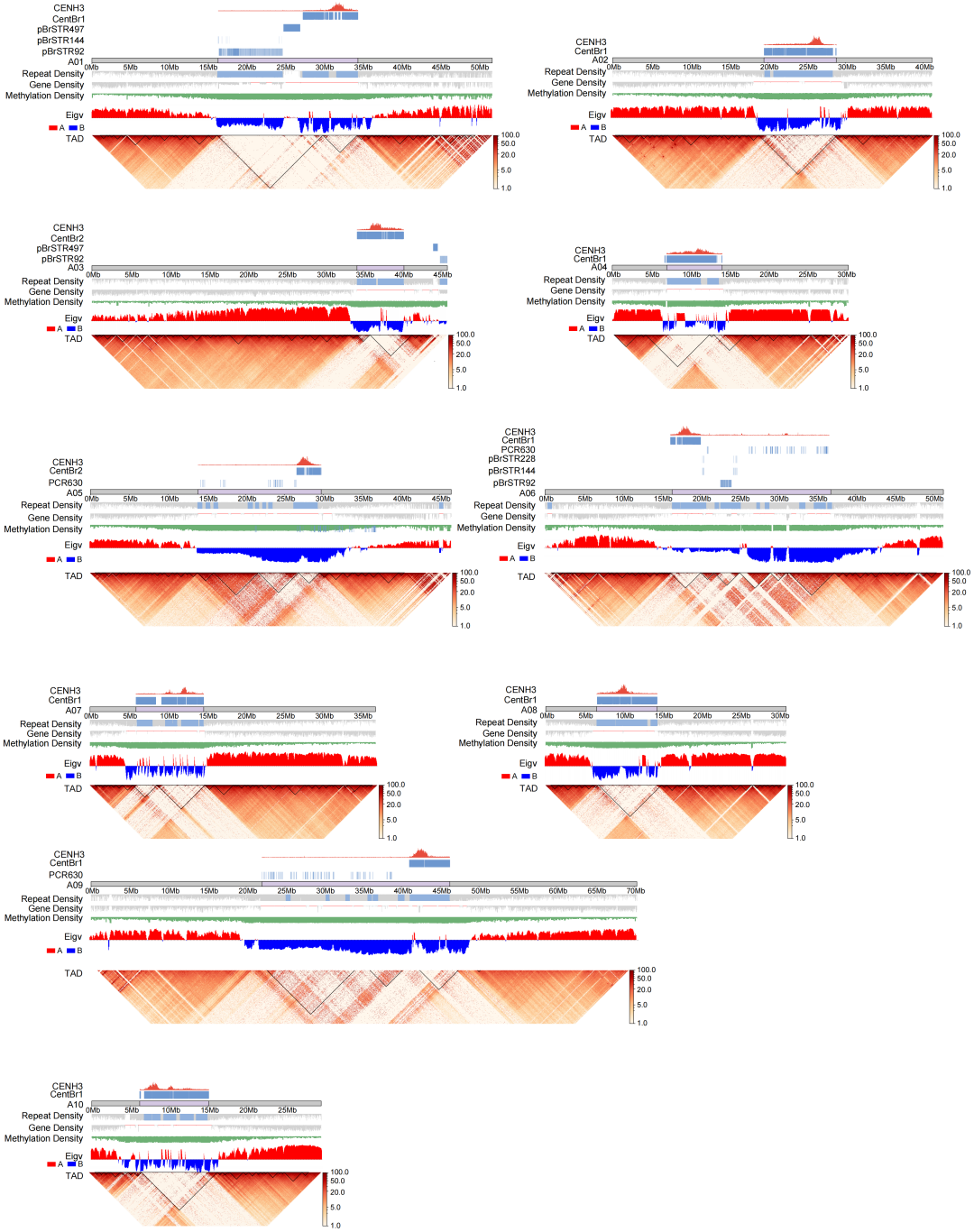


fig. S9. Characterization of the (peri)centromeres in 11 *B. rapa* genomes.
(**A**-**K**) CENH3 Coverage, CENH3 Density, satellite coverage, repeat elements density, gene density, DNA methylation density and A/B compartment in chromosomes A01-A10 of each *B. rapa* accession as indicated. Trapezoid heatmap displays the topologically associating domain (TAD) within the chromatin interaction map. Eigv, eigenvector value of correlation matrix. CENH3-rich region of per chromosome was centromere and pericentromere with high satellite coverage, high repeat elements density and low gene density was shown with purple block. Accessions include TU1 (**A**), MIZ1 (**B**), OI1 (**C**), PC1 (**D**), WTC1 (**E**), CX1 (**F**), TC1 (**G**), CC3 (**H**), CC2 (**I**), CC1 (**J**), and CCA03 (**K**).

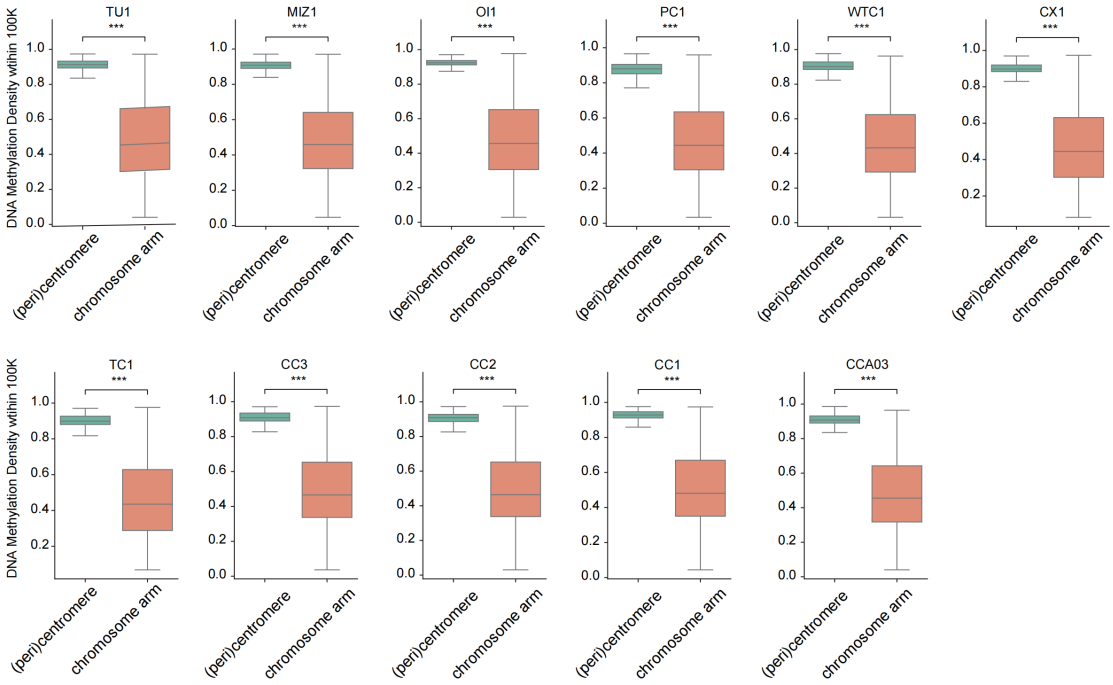


fig. S10. DNA methylation level between (peri)centromeres and chromosomal arms in 11 *B. rapa* subspecies/morphotypes.

Statistical significance of pairwise comparisons between chromosomal regions was assessed by Mann-Whitney U tests, with *** indicating $P < 0.001$.

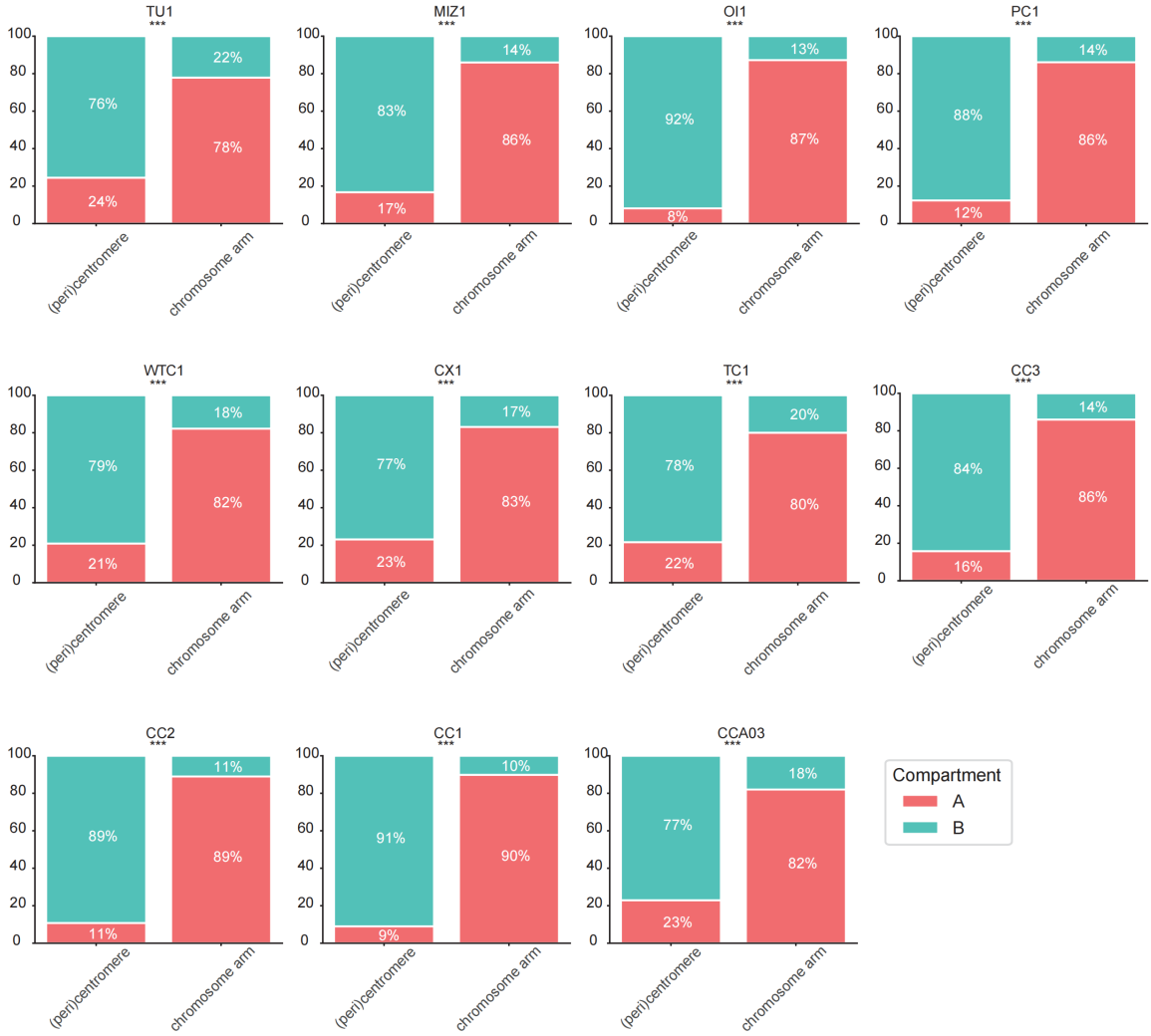


fig. S11. Distribution of A (red) and B compartment (blue) in (peri)centromeric regions and chromosomal arms in 11 *B. rapa* subspecies/morphotypes.
Statistical significance of pairwise comparisons between chromosomal regions was assessed by Mann-Whitney U tests, with *** indicating $P < 0.001$.

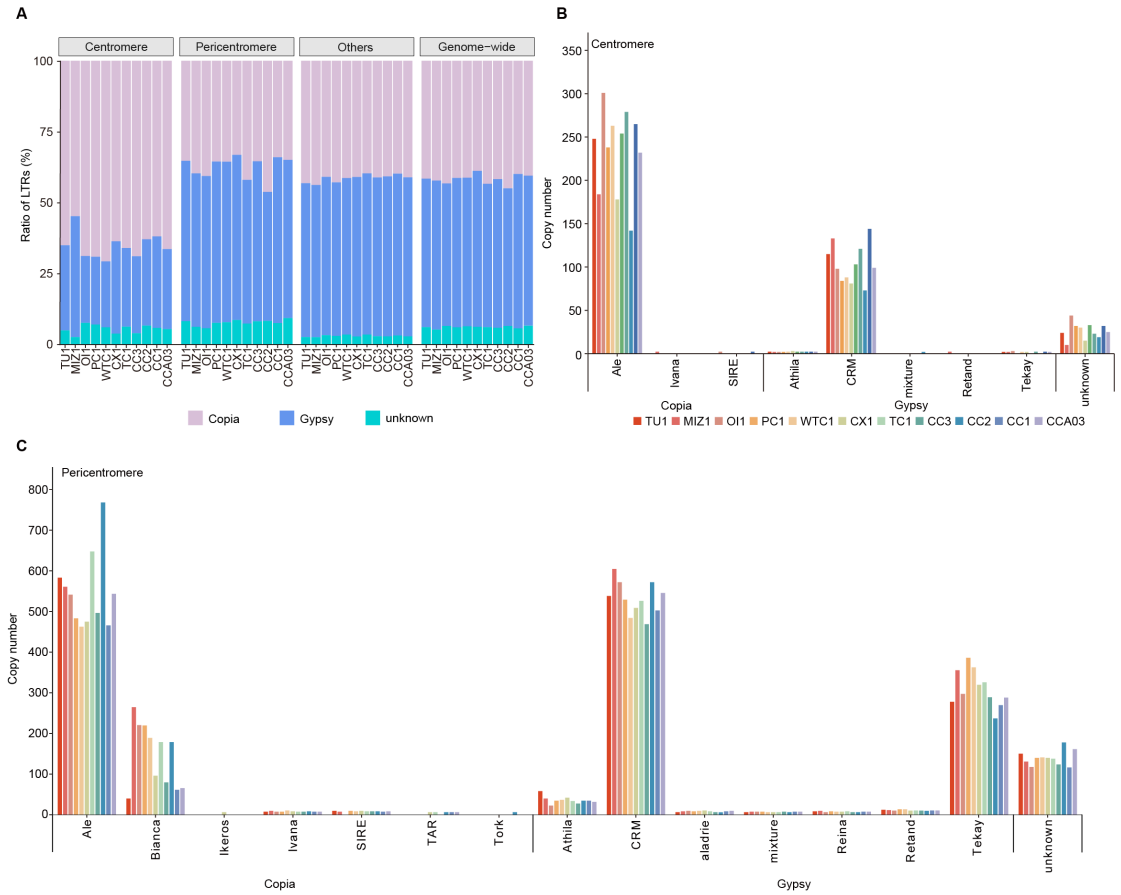


fig. S12. Annotation of LTRs in 11 *B. rapa* genomes.

(**A**) Percentage of LTRs in Centromere, Pericentromere, other genomic regions (Others), and Genome-wide. (**B**) Copy numbers of FL-LTR-RTs in centromere regions of 11 *B. rapa* genomes. (**C**) Copy numbers of FL-LTR-RT families in pericentromeric regions of 11 *B. rapa* genomes.

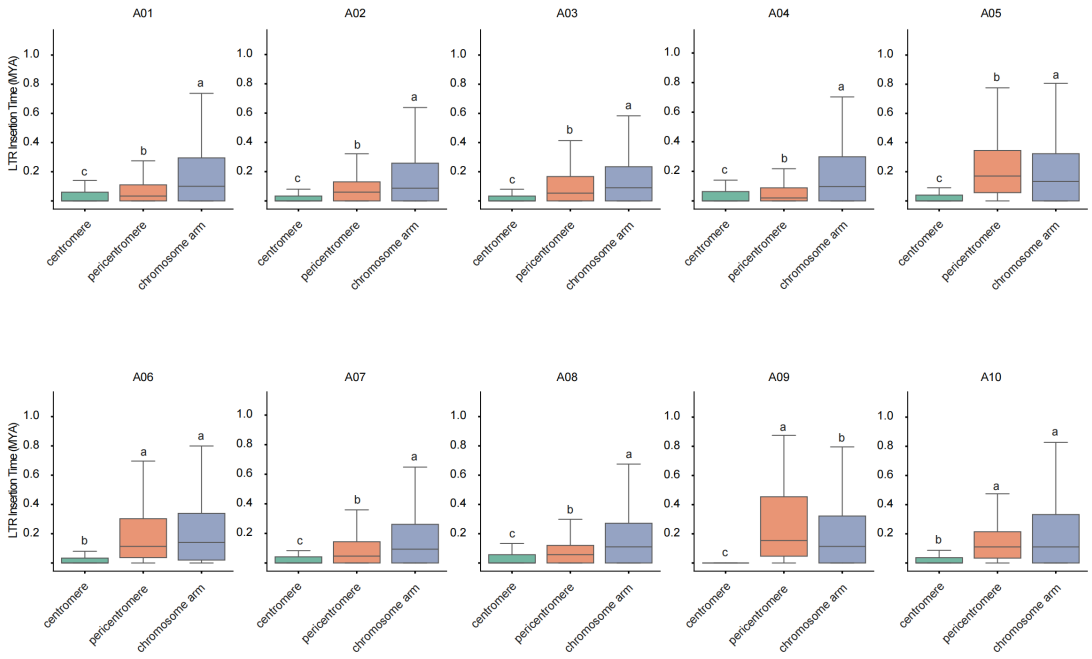


fig. S13. LTR retrotransposon insertion ages across chromosomal regions (A01-A10) in *B. rapa*.

Ages in million years ago (MYA) of LTR-retrotransposon insertions in centromeres, pericentromeres, and chromosome arms of 11 *B. rapa* accessions. Significance testing between chromosomal regions were performed using Mann-Whitney U tests. The different lowercase letters above the box plots represent significant differences ($P \leqslant 0.05$).

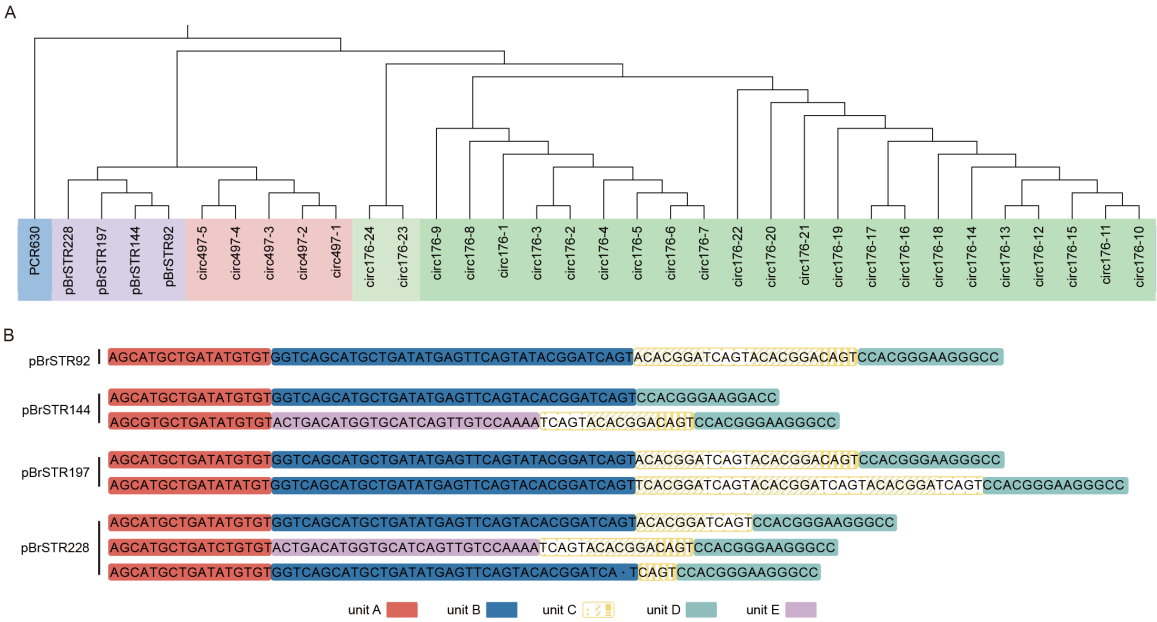


fig. S14. Phylogenetic analysis and classification of satellites.

(**A**) A phylogenetic relationship among seven types of satellites across 11 *B. rapa* genomes. Satellites within the same clades were shown in one color. (**B**) Satellite sequences of pBrSTR92, pBrSTR144, pBrSTR197 and pBrSTR228. Units A-E in each satellite are shown with different colors. Unit C consists of three subunits which are framed by boxes with different patterns. The sequences of PCR630, pBrSTR497 and CentBr are shown in table S17.

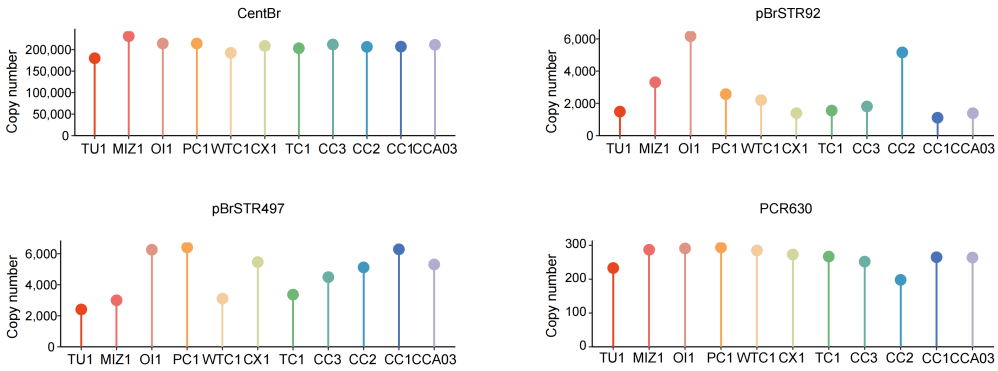


fig. S15. Copy numbers of satellites in 11 *B. rapa* genomes.

Copy numbers of four satellite repeats (CentBr, pBrSTR92, pBrSTR497, and PCR630) are shown across 11 *Brassica rapa* genomes. Bars are color-coded by subspecies/morphotype, as indicated on the x-axis.

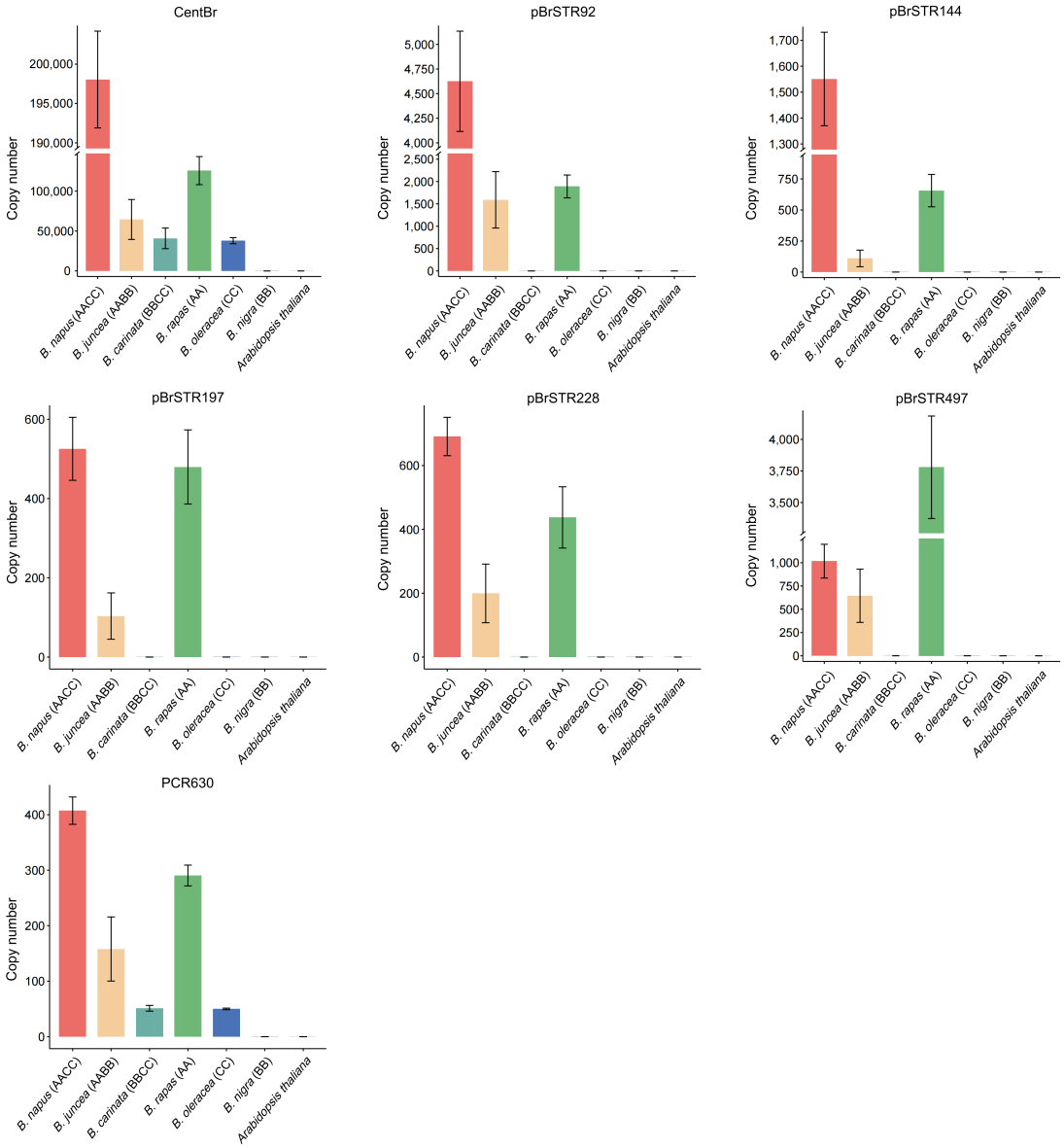


fig. S16. Copy numbers of 5 newly characterized and 2 known satellites across 6 *Brassica* sp., three diploids (AA, BB, CC) and three allotetraploids (AABB, AACC, BBCC) as well as *Arabidopsis thaliana*.

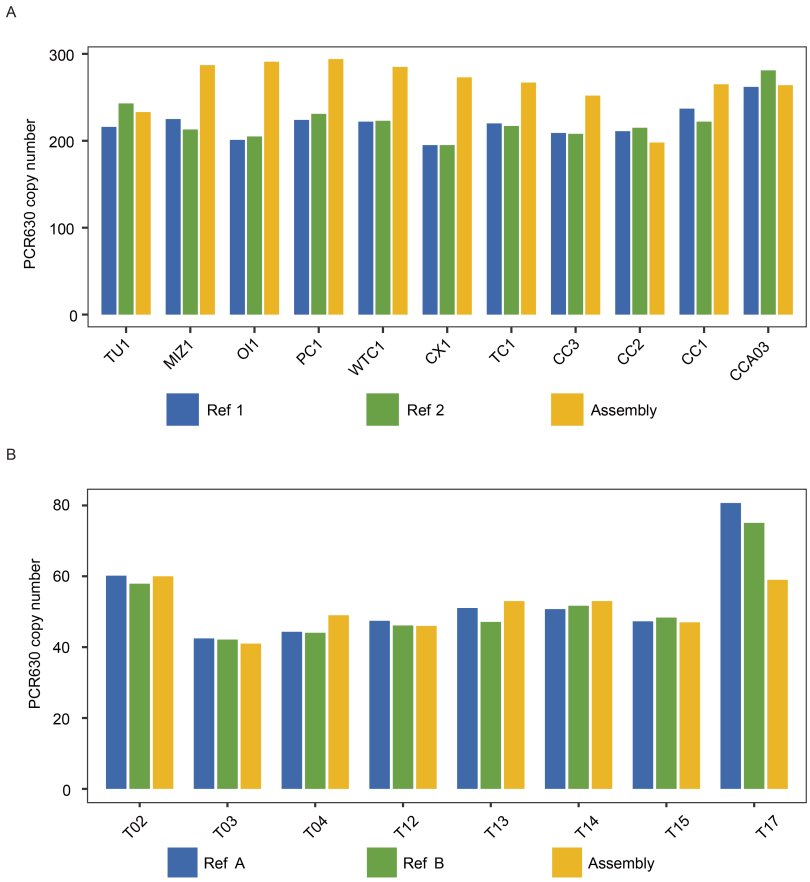


fig. S17. Digital PCR-based analysis of satellites.
Comparison of the copy number of satellite PCR630 in the 11 *B. rapa* (AA) (**A**) and 8 *B. oleracea* (CC) (**B**) genomes. Digital PCR-based estimation was obtained with single copy genes as reference (AA: Ref 1 and Ref 2; CC: Ref A and Ref B). The CC genomes are not complete, with high probability of missing satellite DNAs in their assemblies (*30*). This may result in overcalculating copy number changes in AA vs CC based on the in silico assay. We have thus re-examined the completeness and assembly quality of the published CC genomes (*30*) used for comparisons in our study. The total genome sizes of those CC assemblies range from 539.87 to 584.16 Mb with an average contig N50 of 19.18 Mb. On average, 98% contig sequences are anchored to the nine CC pseudochromosomes. The average BUSCO complete score is 98.70%, indicating high completeness and quality of the CC assemblies. We then performed digital PCRs on 11 AA (A) and 8 CC (B) (*30*) accessions to quantify satellite copy numbers. The PCR630 copy numbers measured by digital PCR vs in silico assay are close. The ratio between the PCR630 copy numbers obtained from digital PCR in AA vs CC is about 4.24. Unfortunately, digital PCR cannot be carried out to estimate the CenBr copy number due to that CentBr is of high sequence polymorphism, and no primers could be designed for digital PCR to capture all CenBrl variants. Nevertheless, both digital PCR and in silico assays show that AA genomes contain more CentBr and PCR630 than the CC counterparts.

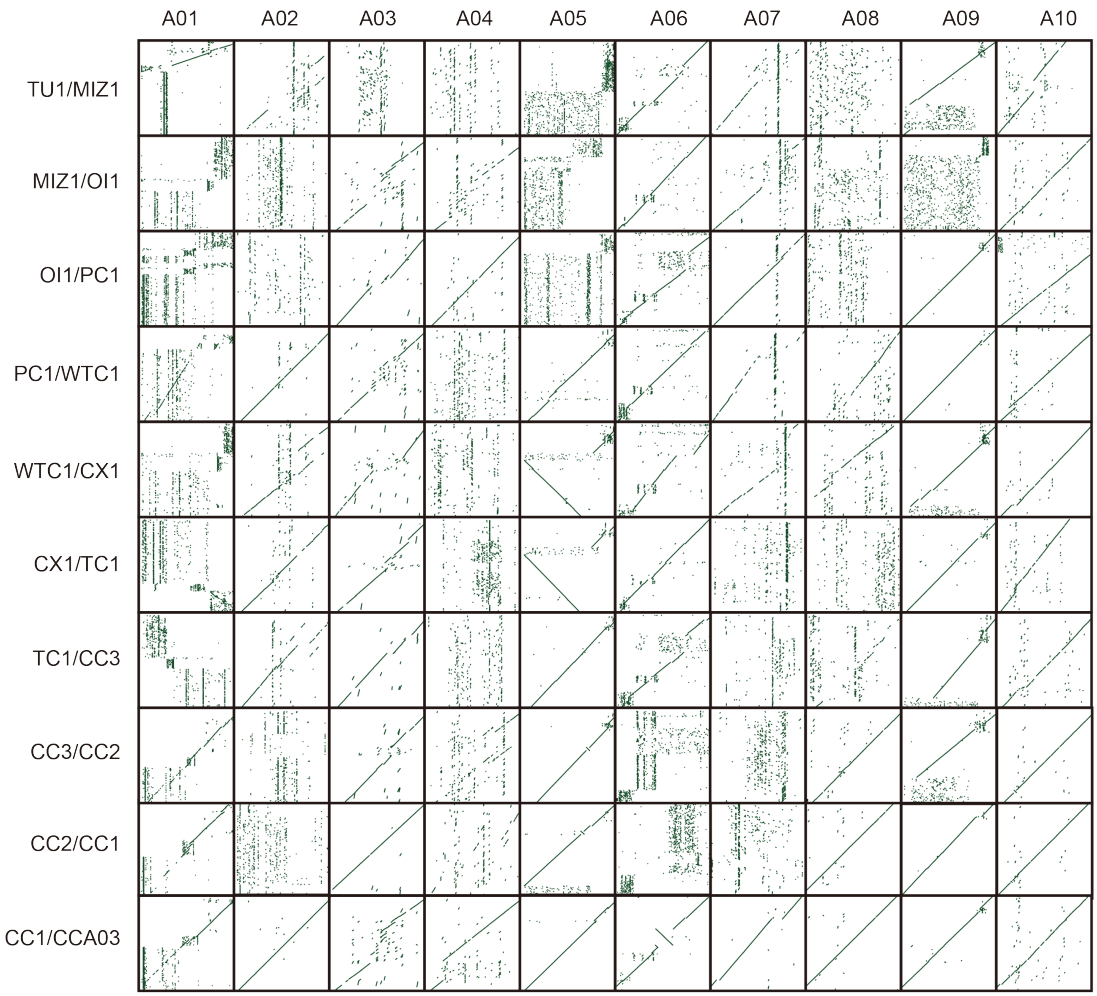


fig. S18. Dot-plots show sequence synteny in the (peri)centromeric regions of 10 chromosomes A01-A10 for each of 10 pairs of the 11 *B. rapa* accessions as indicated.

Dot-plots were produced using the DGENIE software and alignments with mashmap (v2.0). The pair of relevant *B. rapa* accessions are indicated on the left of the graphics. These pairs were chosen based on the position on the phylogenetic tree (Fig. 1A).

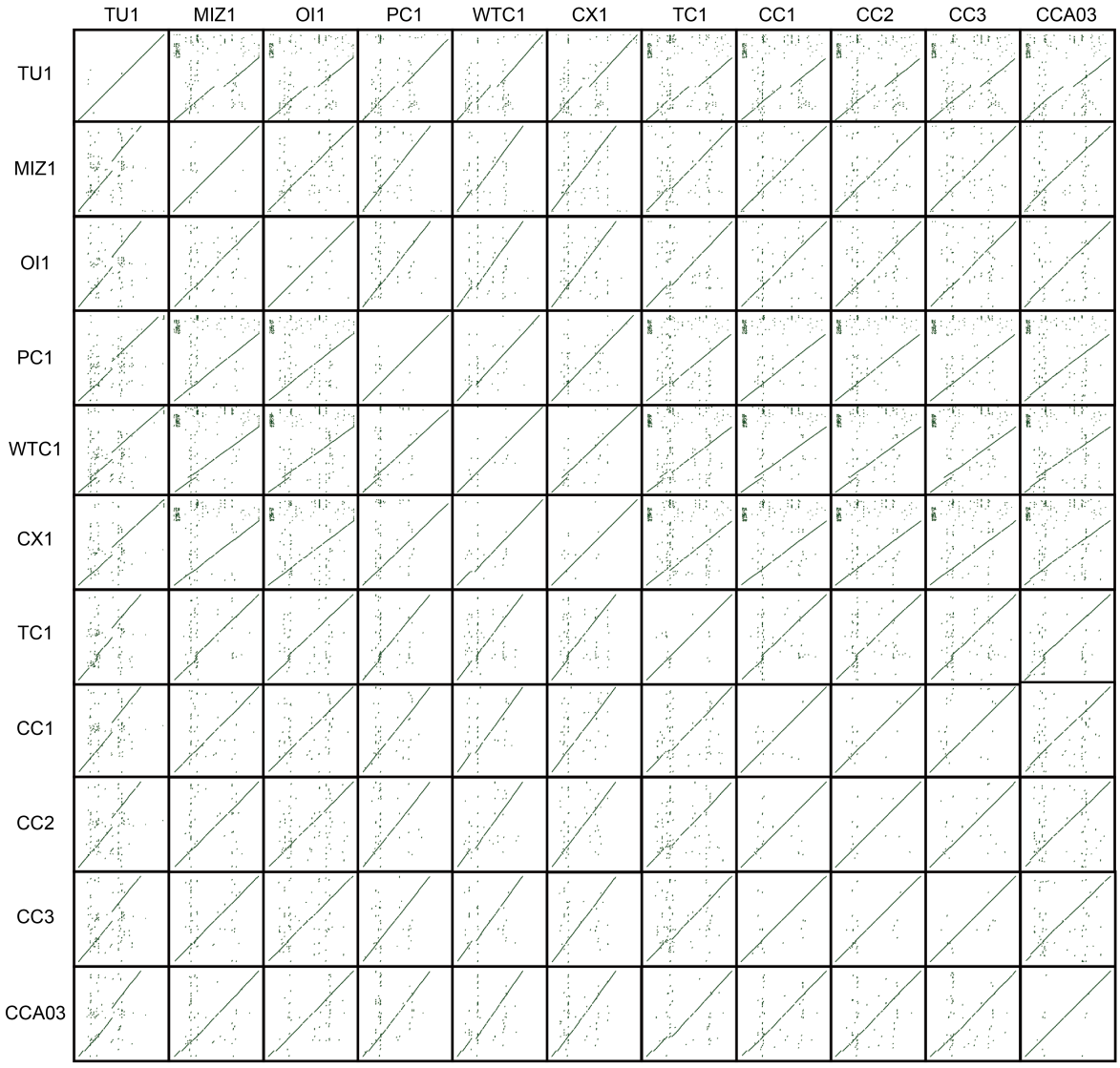


fig. S19. Dot-plots show sequence synteny in the (peri)centromeric regions of chromosome A10 for pairwise comparison between the 11 *B. rapa* accessions. Accessions include TU1, MIZ1, OI1, PC1, WTC1, CX1, TC1, CC1, CC2, CC3 and CCA03.

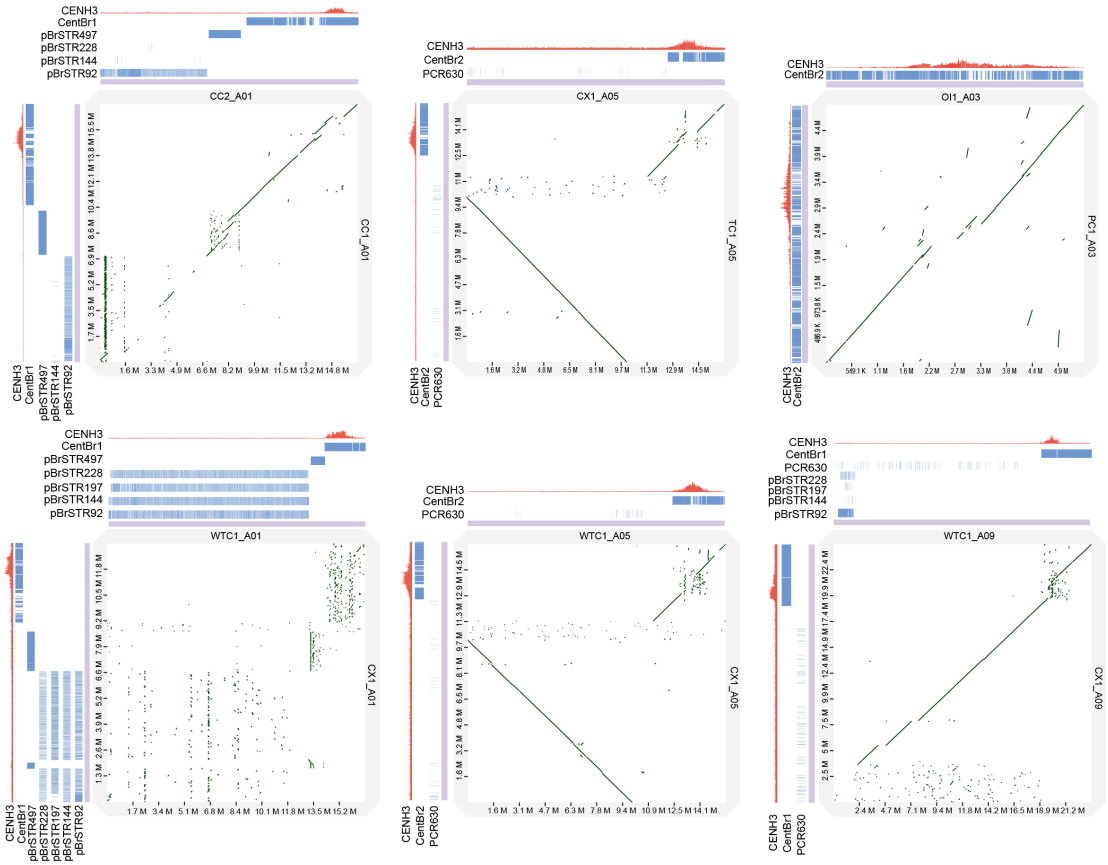


fig. S20. Association of subspecies-specific satellites with (peri)centromeric synteny. CENH3 peaks (red) and satellite coverage (blue) in chromosomes of different *B. rapa* accessions as indicated. Example dot-plots show sequence syntenies in the (peri)centromeric regions of chromosomes for six pairwise comparisons chromosome A01 of CC1 vs CC2, chromosome A05 of TC1 vs CX1, chromosome A03 of PC1 vs OI1, chromosome A01 of CX1 vs WTC1, chromosome A05 of CX1 vs WTC1, and chromosome A09 of CX1 vs WTC1. Horizontal vs vertical axis represents nucleotide positions at the scale of million (M) bases in chromosomes.

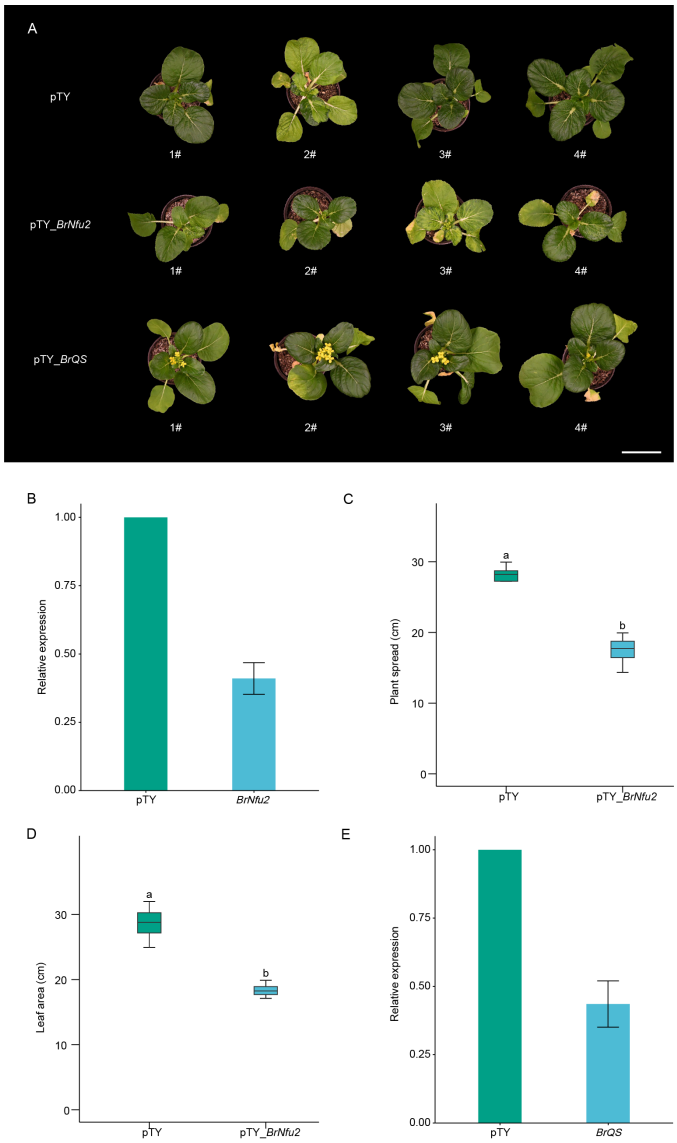


fig. S21. Knock-down of *BrNfu2* and *BrQS* by virus-induced gene silencing (VIGS) affects plant development in CX1.

(**A**) Phenotypic changes caused by *BrNfu2* (reduced size) and *BrQS* (early flowering) gene silencing in CX1 plants. Plants were photographed at 33 days post virus inoculation by bombardment. Relative expression of *BrNfu2* in pTY (empty vector) and gene silencing plants (**B**), plant spread (**C**) and leaf area (**D**). (**E**) Relative expression of *BrQS* in pTY (empty vector) and gene silencing plants. Different lowercase letters above the box plots represent significant differences ($P < 0.05$).

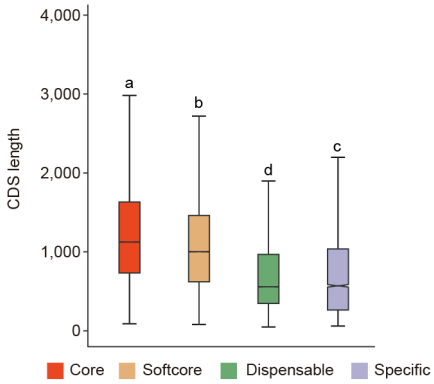


fig. S22. Distribution of protein coding-sequence (CDS) lengths for different types of genes.

The box-plots show the CDS lengths of genes in core, softcore, dispensable, and specific gene families. Different lowercase letters above the box plots represent significant differences ($P < 0.05$).

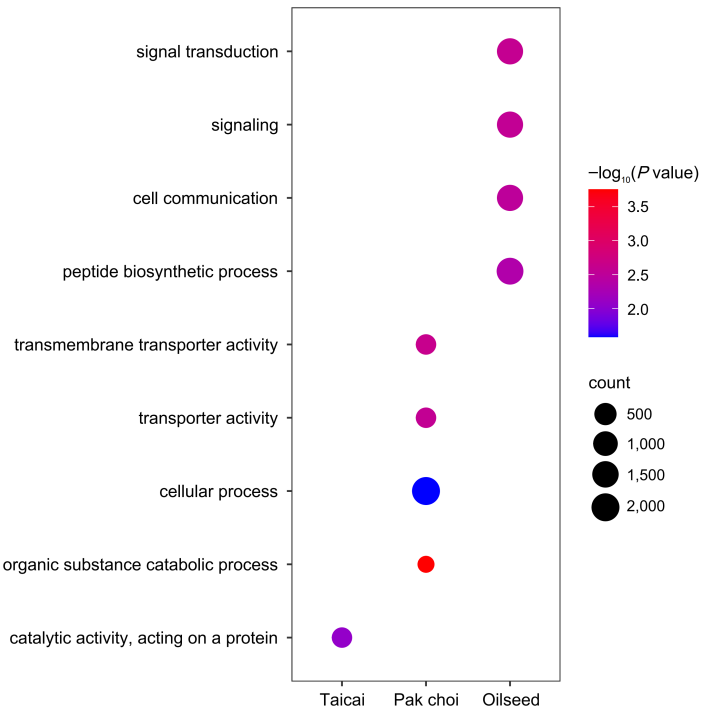


fig. S23. Functional analysis (Gene Ontology) of specific genes on different morphotype/subspecies Taicai, Pak choi and Oilseed.

The *P* value indicates the significance of enrichment for the GO terms; a lower *P* value corresponds to a more significant enrichment result. Morphotype/subspecies-specific genes that were not significantly enriched are not shown here.

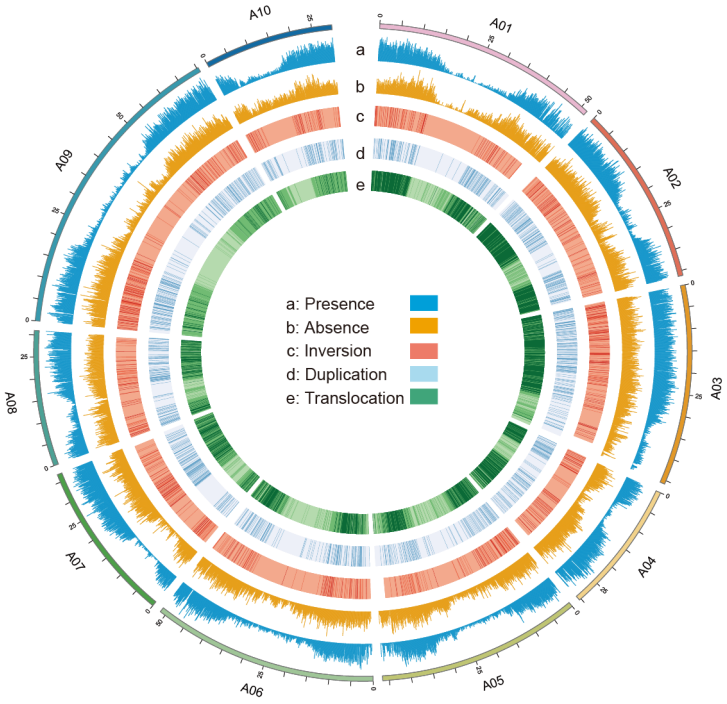


fig. S24. The SV component coverage map with a window of 100kb.
Each circle represents a different variation type in ten chromosomes.

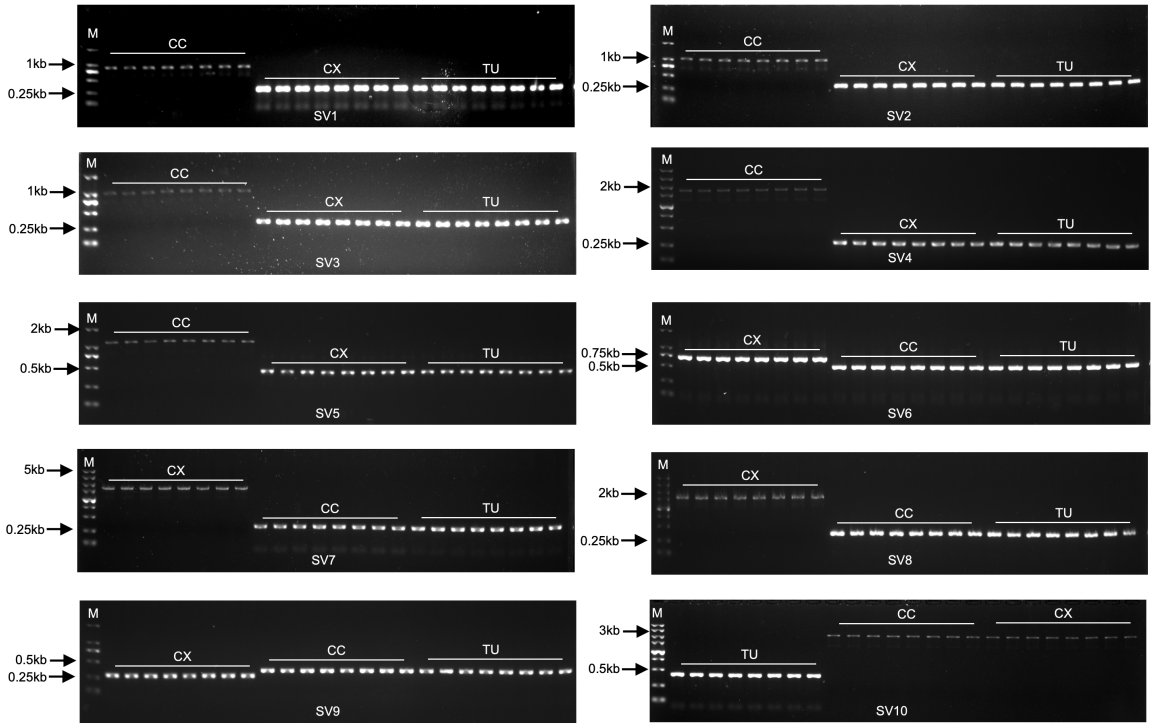


fig. S25. PCR amplification for SVs validation.

Ten PAVs were randomly selected for wet-experimental validation. Different bands show the presence or absence of 8 Chinese cabbage (CC), 8 Caixin (CX) and 8 Turnip (TU). SV1-5 is present in CC, but absent in other morphotypes. SV6-8 are present in CX, but not in other morphotypes. SV9 is absent in CX, but appears in other morphotypes. SV10 is absent in TU, and present in other morphotypes. Subspecies/morphotypes, SVs, as well as size and positions of DNA markers are indicated.

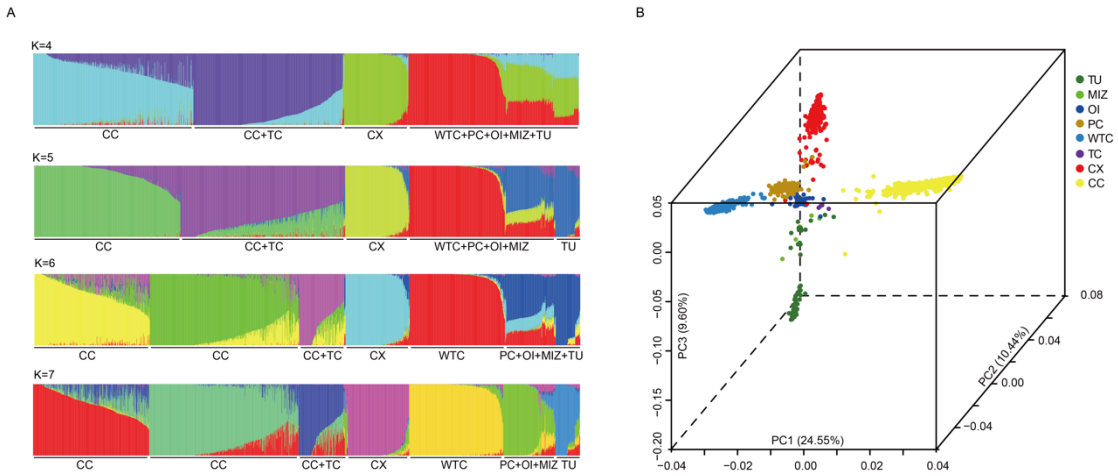


fig. S26. Population structure and principal component analysis (PCA) based on SVs in 1,720 *B. rapa* accessions.

(A) Structure of 1,720 *B. rapa* accessions based on SVs. Bar-plots showing the inferred ancestral components at K = 4 to 7. Each vertical bar represents a group of accessions. Colored segments within each bar indicate the proportional contributions from different ancestral population clusters. (B) Three-dimensional PCA of *B. rapa* accessions based on SVs. Each dot represents a group colored by morphotypes, including Turnip (TU), Mizuna (MIZ), Oilseed (OI), Pak choi (PC), Wutacai (WTC), Caixin (CX), Taicai (TC), and Chinese cabbage (CC). PC1, PC2, and PC3 explain 24.55%, 10.44%, and 9.60% of the total genetic variation, respectively.

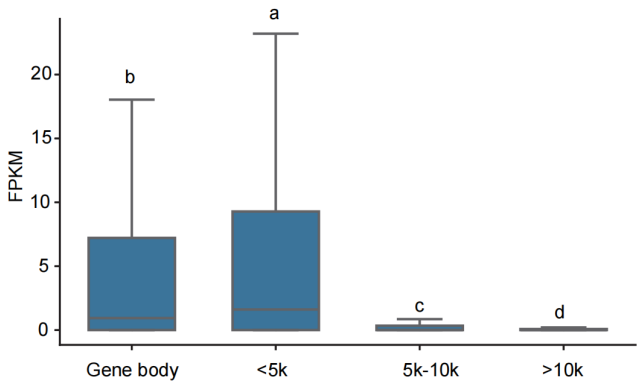


fig. S27. Gene expression corresponding to differential distances from SVs.
Significance testing between chromosomal regions were performed using Mann-Whitney U tests. The different lowercase letters above the box plots represent significant differences (P<0.05). FPKM: Fragments Per Kilobase of exon per Million mapped fragments.

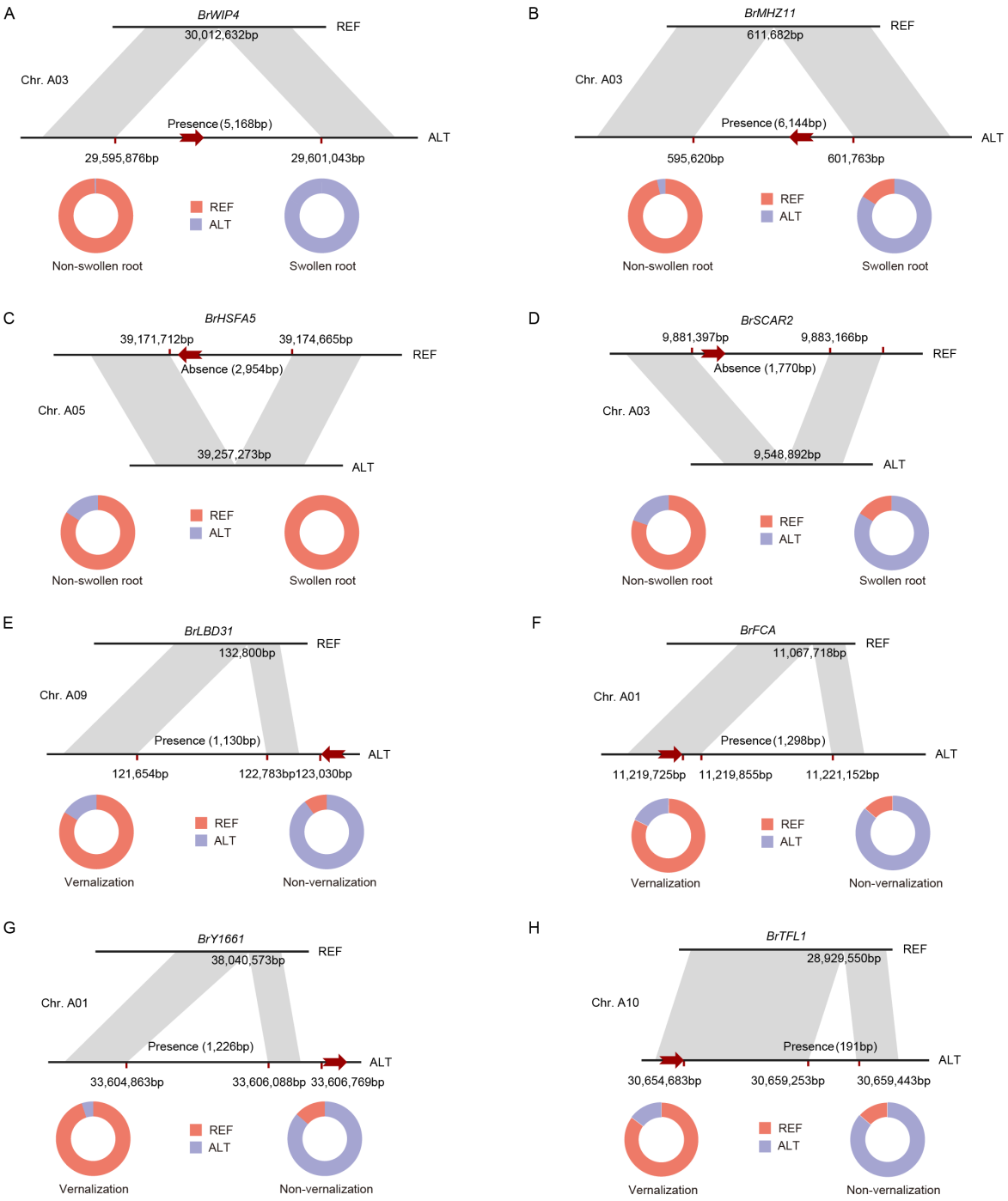


fig. S28. Presence-absence variation leads to gene function variation in different subspecies.

(**A**) In Turnip, a 5,168 bp fragment was identified, encompassing the entire coding region of *BrWIP4* gene. This fragment was present in 61/74 (82.43%) of Turnip accessions, but were only found in 9/1,642 (0.55%) non-Turnip accessions. (**B**) In Turnip, a 6,144 bp fragment was identified, encompassing the entire coding region of *BrMHZ11* gene. This fragment was present in 62/74 (83.78%) of Turnip accessions, but were only found in 59/1,642 (3.59%) non-Turnip accessions. (**C**) All 74 Turnip accessions did not contain a 2,954 bp fragment, whereas 1,381/1,642 (84.10%) of non-Turnip accessions contained this segment. This segment includes *BrHSFA5*. (**D**) 62/74

(83.78%) Turnip accessions did not contain a 1,770 bp fragment, whereas 1,318/1,642 (80.33%) of non-Turnip accessions contained this segment. This segment includes *BrSCAR2*. (**E**) In Caixin, a 1,130 bp fragment was identified, encompassing the region of *BrLBD31* downstream. This fragment was present in 198/220 (90.00%) of Caixin accessions, but were only found in 242/1,496 (16.18%) non-Caixin accessions. (**F**) In Caixin, a 1,298 bp fragment was identified, encompassing the region of *BrFCA* downstream. This fragment was present in 190/220 (86.36%) of Caixin accessions, but were only found in 267/1,496 (17.85%) non-Caixin accessions. (**G)** In Caixin, a 1,226 bp fragment was identified, encompassing the region of *BrY1661* upstream. This fragment was present in 190/220 (86.36%) of Caixin accessions, but were only found in 74/1,496 (4.95%) non-Caixin accessions. (**H**) In Caixin, a 191 bp fragment was identified, encompassing the region of *BrTFL1* downstream. This fragment was present in 190/220 (86.36%) of Caixin accessions, but were only found in 220/1,496 (14.71%) non-Caixin accessions. REF represents the same sequence as the reference genome CCA03, and ALT indicates that the sequence is different from REF. Gray boxes represent the genomic collinearity region. Red arrows show related genes.

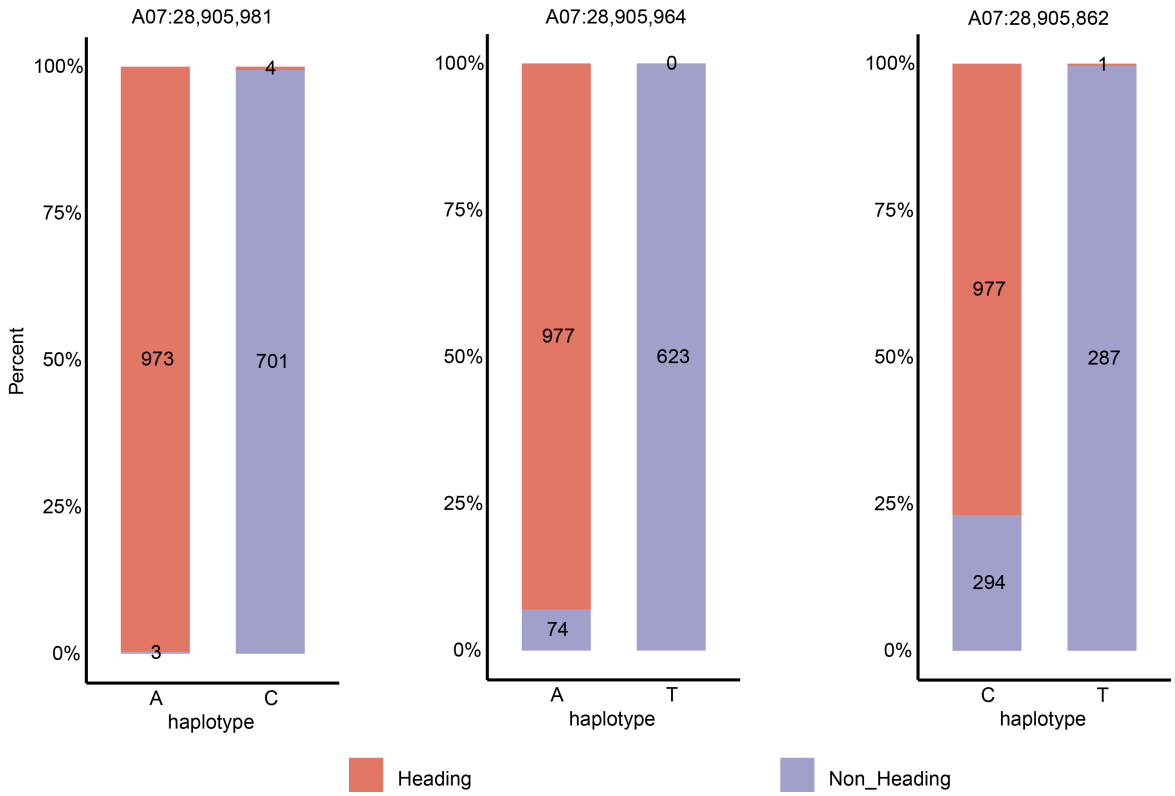


fig. S29. Haplotype analysis of three non-synonymous SNPs in *BrLH1*.
"Heading" denotes accessions forming leafy heads, whereas "Non-Heading" refers to those without head formation. The three panels display the genotype-to-phenotype associations for the three non-synonymous SNPs located at A07:28,905,981, A07:28,905,964, and A07:28,905,862 within the *BrLH1* gene. Each bar represents the proportion of accessions with or without heading in each haplotype. In our *B. rapa* population, the A-to-C SNP at A07:28,905,981, the A-to-T SNP at A07:28,905,964, and the C-to-T SNP at A07:28,905,862 reduces the proportion of heading accessions from 99.7% to 0.6%, 93.0% to 0%, and 96.9% to 0.3%, respectively. The three SNPs collectively exhibit substantial mutation effects and mutations at these SNPs are closely associated with the traits of heading and non-heading in *B. rapa*.

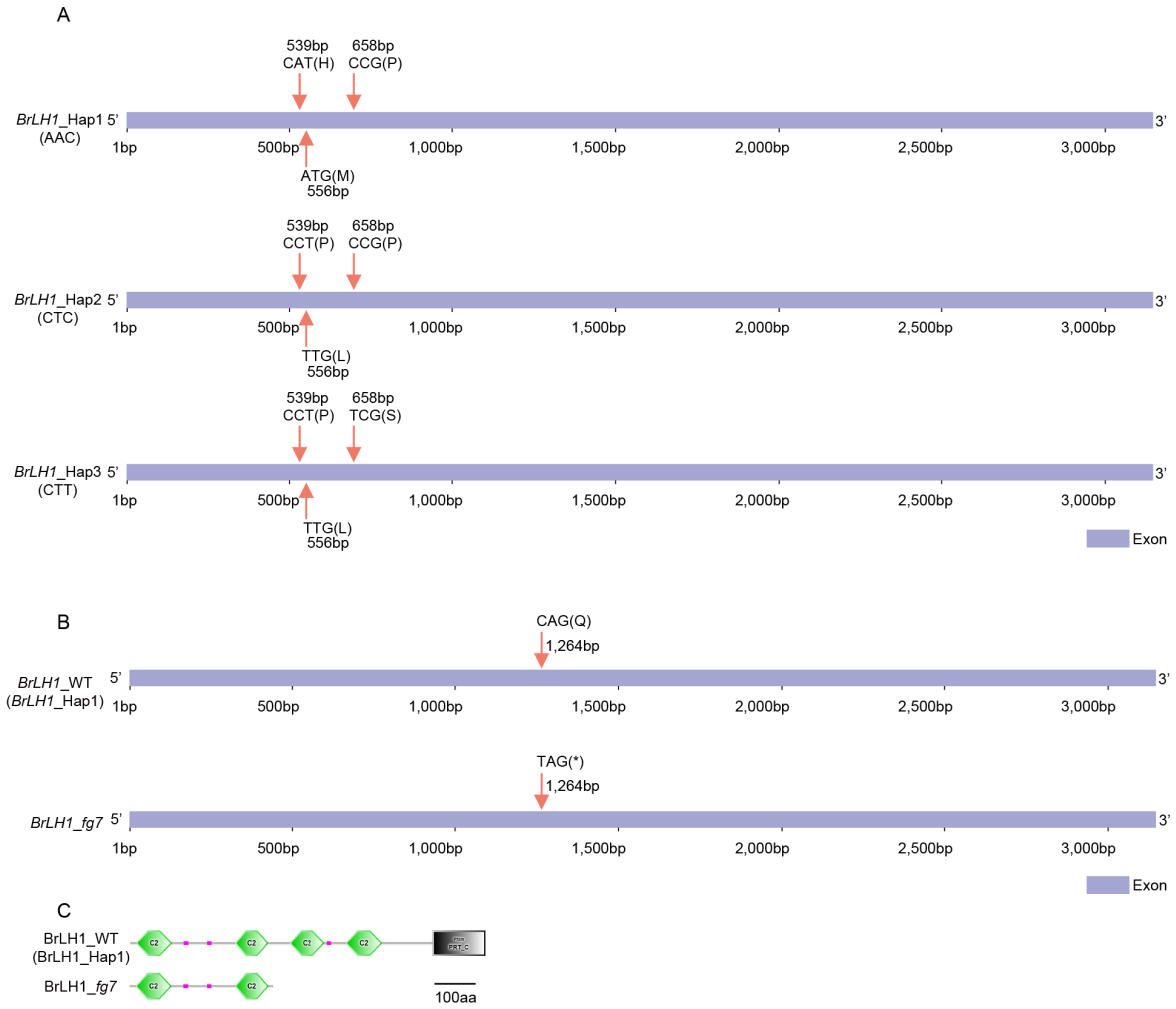


fig. S30. Identification of candidate gene at the loci for *fg7*.

(**A**) Pangenomic and genetic basis of heading vs non-heading in *B. rapa* subspecies. (**B, C**) Forward genetic analysis of candidate gene *BrLH1* responsible for non-heading phenotype in the Chinese cabbage *fg7* mutant. A C-to-T SNP was found in the *BrLH1* exon at nucleotide 1,264bp. This SNP leads to introduce a premature stop codon in the *BrLH1* coding sequence (B), resulting in translational truncation of the wild-type 1,048 amino-acid (AA) BrLH1 protein to a 421-AA polypeptide (C). The codon for Q in wild-type *BrLH1*, the premature stop codon (*) in the *fg7* mutant *BrLH1* (*mBrLH1*), and relevant domains in BrLH1 and mBrLH1 proteins are indicated. C2 presents Protein kinase C conserved region 2. PRT_C is the functional domain Plant phosphoribosyltransferase C-terminal. Pink point represents the low complexity region. The domain was predicted in SMART (https://smart.embl.de/).

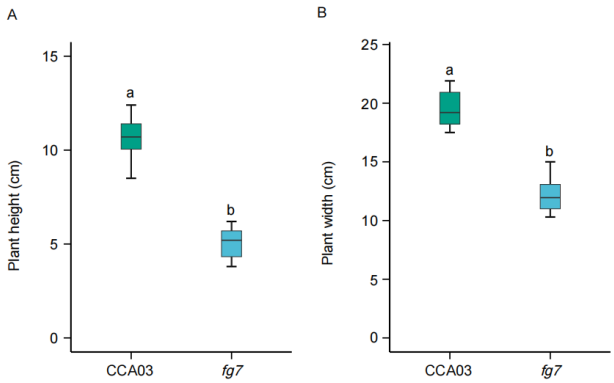


fig. S31. *BrLH1* genotype to phenotypes in *B. rapa*.
Plants height (**A**) and width (**B**) of wild type and mutant *fg7*. Different lowercase letters above the box plots represent significant differences ($P < 0.05$).

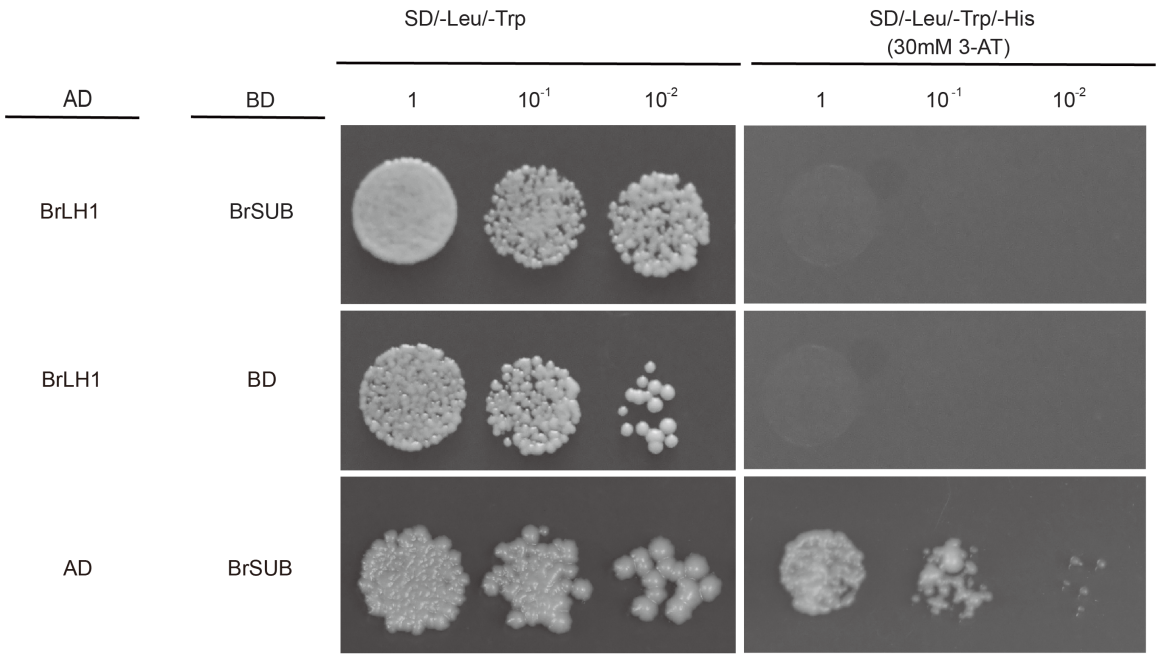


fig. S32. BrLH1 interacts directly with BrSUB in the yeast two-hybrid system.

The construct combinations were co-introduced into the yeast strain AH109. Transformants diluted to different concentrations were grown on the -Leu-Trp (lacking leucine and tryptophan) control plates and the -Leu-Trp-His (lacking leucine, tryptophan and histidine) with 30mM 3-AT (3-Amino-1,2,4-Triazole) selective plates for 3-4 days. The combination of BrLH1/pGBKT7 (empty vector) and pGADT7 (empty vector)/BrSUB as a negative control (BD: pGBKT7; AD: pGADT7).

Taken together, our results reveal how *BrLH1* and its protein product may genetically and biochemically modulate trait development in *B. rapa*.

(i) *BrLH1* belongs to the *SLM* gene family. It encodes a protein containing multiple C2 domains and transmembrane domains, known as MCTP (Multiple C2 domains and Transmembrane Region Proteins). In *B. rapa*, only a single copy of *BrLH1* exists and no other related homologs are found. The *BrLH1* homolog *QUIRKY* (*QKY*) in *Arabidopsis thaliana* is involved in regulating plant cell morphogenesis and cell patterning (*69*). *QKY* mutants exhibit various developmental defects, including disordered root epidermal cell patterns, abnormal floral organ development, twisted stems, and twisted leaves In *Arabidopsis* (*70,71*). QKY directly interacts with the STRUBBELIG (SUB) receptor-like kinase to stabilize SUB at the plasma membrane and maintain its level at the cell surface, which is essential for controlling how tissues develop. In additional, QKY also interacts with SCRAMBLED (SCM), which controls the CAPRICE protein movement to regulate root epidermal cell patterning. SUB mutants also display multiple developmental defects, such as twisted leaves and stems (*72*). Through yeast two-hybrid screening, we found that BrLH1 interacts with BrSUB (fig. S32). Thus, BrLH1 may involve BrSUB to modulate leaf heading in *B. rapa*.

(ii) *BrLH1* consists of 3,147 bp. The three non-synonymous SNPs locate at nucleotide 539, 556, and 658 in the gene, respectively. We notice that in the Chinese cabbage population, the A-to-C SNP at A07:28905981, the A-to-T SNP at A07:28905964, and the C-to-T SNP at A07:28905862 reduces the proportion of heading accessions from 99.7% to 0.6%, 93.0% to 0%, and 96.9% to 0.3%, respectively (fig. S29 and fig. S30A). The three SNPs exhibit substantial mutation effects and mutation at each SNP is closely associated with the traits of heading and non-heading in *B. rapa* (Fig. 5B, fig. S29 and fig. S30A).

(iii) The *fg7* mutant plants were dwarf (Fig. 5C). Why the accessions without AAC haplotype are not dwarf remains to be elucidated. *BrLH1* consists of 3,147 bp. Likely the C-to-T mutation at nucleotide 1,264 in *BrLH1* in *fg7* alters a sense codon to a non-sense stop codon (fig. S30, B to C), leading to premature termination of protein translation to produce a truncated 421-AA polypeptide rather than the wild-type 1048-AA BrLH1 (fig. S30C). Such truncated polypeptide may completely lose its biological functions, leading to non-heading and plant growth reduction (i.e., the dwarf phenotypes) in *fg7*. In contrast, the three SNPs in *BrLH1* in natural populations without the AAC haplotype are all sense mutations, which cause only single amino-acid changes in BrLH1 protein. Such single-amino mutation may not impose any effect on BrLH1 to control plant growth but affects BrLH1 to modulate leaf heading. Consequently, these accessions cannot form leaf heads although they properly grow to normal plant sizes.