Theoretical and Applied Genetics (2020) 133:3187–3199
https://doi.org/10.1007/s00122-020-03664-3

ORIGINAL ARTICLE

Improved *Brassica oleracea* JZS assembly reveals significant changing of LTR-RT dynamics in different morphotypes

Xu Cai[1] · Jian Wu[1] · Jianli Liang[1] · Runmao Lin[1] · Kang Zhang[1] · Feng Cheng[1] · Xiaowu Wang[1]

Received: 15 March 2020 / Accepted: 3 August 2020 / Published online: 9 August 2020

Abstract

Brassica oleracea is an important vegetable crop that has provided ancestor genomes of the two most important *Brassica* oil crops, *Brassica napus* and *Brassica carinata*. The current *B. oleracea* reference genome (JZS, also named 02–12) displays problems of large mis-assemblies, low sequence continuity, and low assembly integrity, thus limiting genomic analysis. We reported an updated assembly of the *B. oleracea* reference genome (JZS v2) obtained through single-molecule sequencing and chromosome conformation capture technologies. We assembled an additional 83.16 Mb of genomic sequences, and the updated genome features a contig N50 size of 2.37 Mb, representing an ~ 88-fold improvement. We detected a new round of long terminal repeat retrotransposon (LTR-RT) burst in the new assembly. Comparative analysis with the reported genome sequences of two other genomes of *B. oleracea* (TO1000 and HDEM) identified extensive gene order and gene structural variation. In addition, we found that the genome-specific amplification of Gypsy-like LTR-RTs occurred around 0–1 million years ago (MYA). In particular, the *athila*, *tat,* and *Del* families were extensively amplified in JZS around 0–1 MYA. Moreover, we identified that the syntenic genes were modified due to the insertion of genome-specific LTR-RTs. These results indicated that the genome-specific LTR-RT dynamics were associated with genome diversification in *B. oleracea.*

Introduction

Brassica oleracea is one of the most economically important *Brassica* species cultivated worldwide, mainly as a vegetable crop that includes cabbage, broccoli, and cauliflower subspecies/morphotypes (Kopsell and Kopsell 2006; Liu et al. 2014). *B. oleracea* also provided the ancestor genome of *Brassica napus* and *Brassica carinata*, both of which are cultivated as important oil crops. A high-quality genome assembly has been pursued for the improvement of the genetics and breeding in these crops. The whole-genome sequences of JZS (*B. oleracea* sp. *capitata*, heading type) and TO1000 (*B. oleracea* ssp. *Alboglabrata*, kale-like type), which were assembled using the next-generation sequencing data, were released in 2014 (Liu et al. 2014; Parkin et al. 2014). While NGS technology provided high-accuracy and high-throughput reads to assemble the genome at relatively low cost, the methods had several limitations, especially the short sequencing length that directly led to low assembly integrity, low sequence continuity, and a large number of gaps and assembly errors in the JZS current genome. Single-molecule real-time (SMRT) sequencing developed by Pacific BioSciences (PacBio) could offer longer sequencing reads, and these long sequencing reads could greatly improve the integrity of the genome assembly, especially the assembly of repeat regions (Jiao and Schneeberger 2017; Rhoads and Au 2015). In 2018, a high-quality HDEM (*B. oleracea* ssp. *botrytis italica*, broccoli type) genome assembled by long reads was released; both sequence continuity and assembly integrity had been greatly improved, in particular the sequences of transposable elements (Belser et al. 2018). The three *B. oleracea* accessions JZS (Jinzaosheng), HDEM and TO1000 used here belong to three different crops. JZS is a heading cabbage belonging to *B. oleracea* ssp. *Capitata*, HDEM is a broccoli belonging to *B. oleracea* ssp. *botrytis italica*, and TO1000 was derived from Chinese kale belonging to *B. oleracea* ssp. *alboglabra*.

Chromosome conformation capture (Hi-C) is a sequencing-based approach for determining genome 3D organization that has been used to anchor scaffolds to chromosomes

Communicated by Isobel A. P. Parkin.

Xiaowu Wang
wangxiaowu@caas.cn

[1] Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Haidian District, No.12. Zhongguancun South St, Beijing 100081, China

in the *Brassica* genome (Dudchenko et al. 2017; Wang et al. 2019; Zhang et al. 2018). Efficient Hi-C pipelines of anchoring scaffolds to chromosomes was an important factor in the rapid application of Hi-C technologies to plant genome assembly. Nowadays, the main software programs for anchoring scaffolds to chromosomes by Hi-C data are Lachesis, 3D-DNA and ALLHiC (Burton et al. 2013; Dudchenko et al. 2017; Zhang et al. 2019). Lachesis extracts the interaction information of the Hi-C reads alignments and then clusters the scaffolds and divides them into different groups, finally sorting and updating the scaffolds to the chromosome level. 3D-DNA not only has the functions possessed by the Lachesis software but also automatically performs the scaffolds error correction based on the interaction information and handles some misjoins introduced in the preliminary assembly. These misjoins are formed by two distant scaffolds being assembled together (Fig. S1), and these misjoins can directly affect the construction of pseudomolecules. ALLHiC is mainly used to phase and scaffold polyploid genomes based on Hi-C data. Currently, although some software can automatically correct misjoins in scaffolds (i.e., SALSA and 3D-DNA) (Dudchenko et al. 2017; Ghurye et al. 2017), these pipelines often determine breakpoints within gene bodies and generate many false-positive breakpoints.

More complete assembly allows the identification of new features of transposable elements (TEs). As a major group of plant TEs, long terminal repeat retrotransposons (LTR-RTs) are an important component of plant genomes, and the content varies greatly among different species (Morgante et al. 2005). After updating the *B. rapa* genome with PacBio technology, a more complete TE sequence was obtained that enabled detection of a new round of LTR-RT burst events in the *B. rapa* genome (Zhang et al. 2018). However, a small content and only a single round of LTR-RT burst was detected in the *B. oleracea* JZS reference genome (Liu et al. 2014). From previous reports, amplification of LTR-RTs could change the genomic size and structure as well as regulate the diversification of different species (Kim et al. 2017; Naito et al. 2006; Zhou et al. 2017). Lineage-specific LTR-RT bursts could rapidly change genome size, and the distinct amplification patterns of different transposable element families led to diversification of different species (Ammiraju et al. 2007; Du et al. 2010; Hawkins et al. 2006; Piegu et al. 2006; Vitte et al. 2007). However, we know little about how LTR-RTs regulate genome diversification among *Brassica* subspecies. In *B. rapa*, an LTR-RT insertion in the *FLOWERING LOCUS T* was associated with delayed flowering which directly regulated plant phenotype (Zhang et al. 2015). With PacBio technology, more complete LTR-RT sequences could be obtained, making it possible to comprehensively study LTR-RTs.

Here, we reported a new version of *B. olerecea* genome assembly (JZS v2) with a high level of assembly integrity and sequence continuity. The methods involved single-molecule sequencing and Hi-C technology. We assembled an additional 83.16 Mb of genomic sequences, and the updated genome features a contig N50 size of 2.37 Mb, representing an ~ 88-fold improvement. We corrected large mis-assemblies in the previous assembly and detected a new round of LTR-RT burst in the new assembly. Meanwhile, we developed a reliable misjoins correction pipeline based on Hi-C data. Then, by comparative analysis with the other two reported assemblies of *B. oleracea* (TO1000 and HDEM), we identified extensive gene order and gene structural variations. Meanwhile, we investigated LTR-RTs in the three genomes and found that the subspecies genome-specific amplification of Gypsy-like LTR-RTs occurred less than 1 (mostly 0.4) million years ago (MYA). In particular, the *athila*, *tat,* and *Del* families were highly amplified in JZS in the last 1 MYA. We further found that gene modifications had occurred by the genome-specific insertion of LTR-RTs, and these modified genes were enriched in cells, cell parts and endomembrane system functions. Our analyses unveiled extensive gene structural variations among *B. oleracea* subspecies genomes and showed that the genome-specific LTR-RT dynamics were associated with genome diversification in *B. oleracea*.

Results

Genome sequencing and assembly

We sequenced and de novo assembled the JZS genome through a combination of three technologies (Illumina, PacBio and Hi-C). We used 38.34 Gb (~ 61 ×) single-molecule real-time (SMRT) sequencing reads and 53.16 Gb (~ 84 ×) paired-end sequencing reads (Table S1) to assemble the JZS genome, resulting in a 561 Mb assembly with a contig N50 size of 2.37 Mb (Table 1). Compared to the JZS v1 assembly, we assembled an additional 83.16 Mb of genomic sequences (Table 1).

We developed a reliable pipeline to automatically correct misjoins in the original scaffolds. For construction of pseudomolecules, we first used 3D-DNA to correct misjoins in the original scaffolds. The analysis yielded a total of 1,018 breakpoints, of which 137 were inside the gene bodies. To improve integrity of predicted genes and reduce the number of false positive breakpoints, we developed a more reliable error correction pipeline that did not rely on the 3D-DNA misjoin correction module. This pipeline would not fragment the previously annotated gene sequences. After correction, we detected 70 reliable breakpoints, and the assembly contained 1,184 contigs with a contig N50 size of 2.37 Mb

Table 1 Assembly statistics of JZS v2 and other *B. oleracea* assemblies

Item	JZS v2	JZS v1 (Liu et al. 2014)	TO1000 (Parkin et al. 2014)	HDEM (Belser et al. 2018)
Number of contigs	1184	48,307	85,075	264
Contig N50 (Mb)	2.37	0.03	0.02	9.49
Contig N90 (kb)	442.17	6.22	3.41	2202.32
Contig sizes (Mb)	561.01	477.85	445.62	545.02
Assembly size (Mb)	561.16	515.37	488.62	554.98
Total sequences	649	1573	32,928	129
Ratio of anchored sequences (%)	96.17	74.71	91.46	95.29
Assembly ($G+C$) s (%)	36.74	33.71	32.92	35.92

(Table S2). The results showed that error correction by the newly developed pipeline significantly improved the construction of pseudomolecules (Fig. S2).

The quality of the JZS v2 genome was evaluated using four methods. First, 96.17% of the assembled sequences were anchored to nine chromosomes, which significantly improved the ratio of 74.71% in the JZS v1 assembly (Table 1). Second, approximately 97.1% of embryophyte genes were detected in the JZS v2 assembly by BUSCO (Waterhouse et al. 2018), similar to those in other published *B. oleracea* genomes (97.2% for JZS v1, 97.0% for TO1000, and 96.9% for HDEM) (Table S3), indicating the near-complete genome of JZS v2. Third, we further used LTR Assembly Index (Ou et al. 2018) to evaluate the genome continuity. The LAI value of JZS v2 was increased from 4.13 for JZS v1 to 10.13, indicating the high quality of repeat sequences. Fourth, comparison of the high-quality genome assembly of HDEM and whole genome Hi-C contact map indicated that large mis-assemblies existing in the JZS v1 assembly were extensively corrected in JZS v2 assembly (Fig. 1, Fig. S3).

Genome annotation

In the JZS v2 genome, 48.06% (269.66 Mb) of the assembly sequences were annotated as repetitive elements, approximately 47 Mb longer than that in the JZS v1 assembly (Fig. 2, Table S4). The most abundant repetitive sequence type was the LTR-RT. As a predominant group of plant TEs, LTR-RTs include two main suprfamilies, Gypsy-like and Copia-like LTR-RTs. The Gypsy-like and Copia-like LTR-RT represented approximately 11.31% (63.47 Mb) and 10.13% (56.83 Mb) of the JZS v2 genome sequences, and 21 Mb and 18.83 Mb longer in length than correspondingly length in the JZS v1 genome, respectively. Compared to the other two *B. oleracea* subspecies assemblies, there were differences in the composition of TEs (the repeat sequences of JZS v2, HDEM and TO1000 assemblies were 269.66 Mb, 255.31 Mb and 189.37 Mb, respectively) (Table S4).

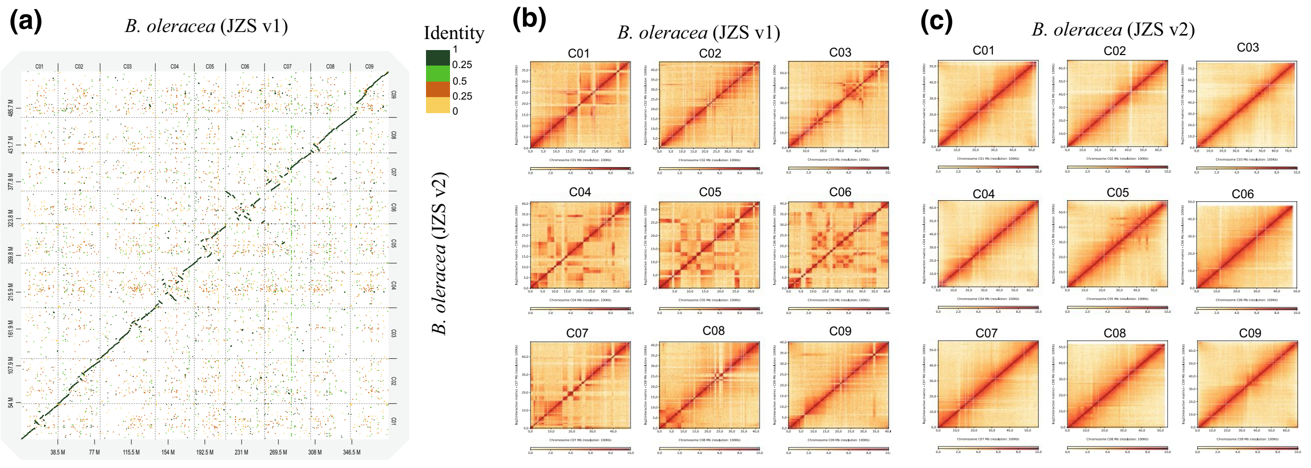
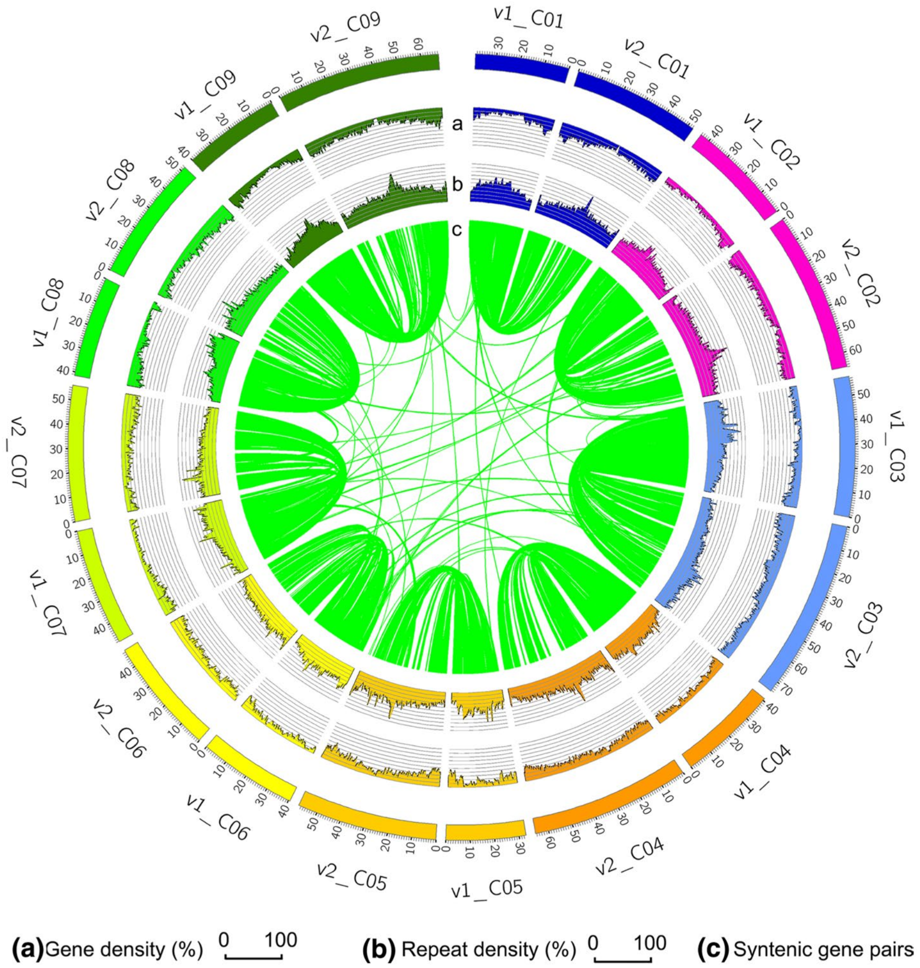


Fig. 1 Comparison between assemblies of JZS v1 and JZS v2. **a** Dot-plot alignment of the JZS v2 with JZS v1. **b**, **c** Whole genome contacts of Hi-C data of JZS v1 **b** and JZS v2 genomes (**c**)

Fig. 2 Genomic landscape between *B. oleracea* JZS v1 and JZS v2 assemblies. **a**, **b** Gene and transposable-element density in sliding windows of 500 kb and step size 50 kb. **c** Syntenic gene pairs between JZS v1 and JZS v2 assemblies



In total, 59,064 protein-coding genes were predicted in the JZS v2 genome by using a pipeline combining ab initio, protein-homology-based and RNA-seq-based gene prediction (Table 2). Approximately 97.5% of the embryophyte genes were detected in the JZS v2 gene dataset according to BUSCO, and the ratio was higher than in the previously reported *B. oleracea* annotated gene datasets (Table S5). Compared to the annotation of JZS v1, the number of genes in the JZS v2 genome was 13,306 more than in the JZS v1 annotated dataset, and the extra genes in JZS v2 were mainly WGD genes (Table S6). Then, we aligned gene sequences (include intron) in JZS v1 to JZS v2 assembly; the results indicated that more than 95% gene sequences in JZS v1 were well aligned to the JZS v2 assembly (Identity > = 0.95; Coverage > = 0.95). Similarly, nearly 90% of the gene sequences in JZS v2 were well aligned to JZS v1 assembly (Table S7). Further analysis revealed that nearly 90% of the annotated genes in the JZS v1 genome were syntenic with the JZS v2 genome, and nearly 80% of the non-syntenic genes in JZS v2 were supported by homologous evidence in other Brassicaceae species (Table S8 and S9), indicating the reliability of our annotation. In addition, the number of annotated genes in the JZS genome was less than that of HDEM genome. This might be due to the inclusion of some low-quality annotated genes in published *B. oleracea* genomes,

Table 2 Statistics of predicted genes among *B. oleracea* assemblies

Item	JZS v2	JZS v1	TO1000	HDEM
Number of genes	59,064	45,758	59,225	61,279
Number of genes on plus strand	29,556	23,020	29,537	30,419
Number of genes on minus strand	29,508	22,738	29,688	30,860
Multi-exon genes	51,231	33,336	46,129	45,232
Mean gene length (bp)	2270	1761	1750	1969
Gene density (gene/Mb)	105.25	88.79	121.21	110.42
Number of CDS	293,551	208,039	268,617	263,525
Mean CDS length (bp)	233	228	230	218
Percent coding sequences (%)	12.17	9.21	12.63	10.35
Number of intron	234,487	162,281	209,392	202,246
Mean intron length (bp)	280	204	200	313
Percent Intron sequences (%)	11.72	6.43	8.58	11.39

such as genes without start codons, genes ending without a stop codon, and gene lengths less than 50 bp etc. In the JZS v2, JZS v1, TO1000, and HDEM genomes, the low-quality genes accounted for 0% (0 of 59,064), 0.91% (416 of 45,758), 9.28% (5498 of 59,225), and 17.87% (10,950 of 61,279), respectively (Table S10).

Construction of JZS v2 sub-genomes and genome blocks

We reconstructed from JZS v2 three sub-genomes using the gene syntenic relationships to *Arabidopsis thaliana* (Table S11). There were 54,120 annotated genes in the three sub-genomes, and the LF sub-genome maintained more gene copies than the other two sub-genomes (Fig. S4). We calculated the syntenic relationship of JZS v2 and JZS v1 with *A. thalina*; the analysis showed that the continuity of the syntenic fragments in JZS v2 was higher than that of JZS v1, especially on chromosomes C04, C05, and C06 (Fig. S5). This indicated that JZS v2 could provide more complete sub-genomes and genome blocks. Then, we defined genome blocks in the JZS v2 genome based on sub-genomes information (Fig. S6, Table S12). In addition, we investigated the location of the centromeres of nine chromosomes in the JZS v2 genome. In total, we detected 16.17 Mb centromere sequences on nine chromosomes, and gene density in these regions was 29.00 genes per Mb, which was much lower than the gene density of the whole genome (105.25 genes per Mb) (Table S13).

Extensive gene order and gene structural variations among *B. oleracea* genomes

Extensive gene order and gene structural variation was detected among the three *B. oleracea* genomes. On the basis of the coding sequences of the 15,422 single-copy orthologous genes, we constructed the phylogeny for the three genomes with *B. rapa* as the outgroup (Fig. 3a). Then, we calculated syntenic genes of the three genomes. Approximately 83.97% (49,595 of 59,064), 83.44% (51,134 of 61,279) and 82.96% (49,132 of 59,225) syntenic genes were detected in JZS v2, HDEM and TO1000 assemblies (Table S14). Then, we calculated syntenic genes among JZS v2, TO1000, HDEM and *A. thaliana*; there were 2,272 genes that being lost in the JZS v2 genome but presenting in the HDEM or TO1000 genomes. Meanwhile, there were

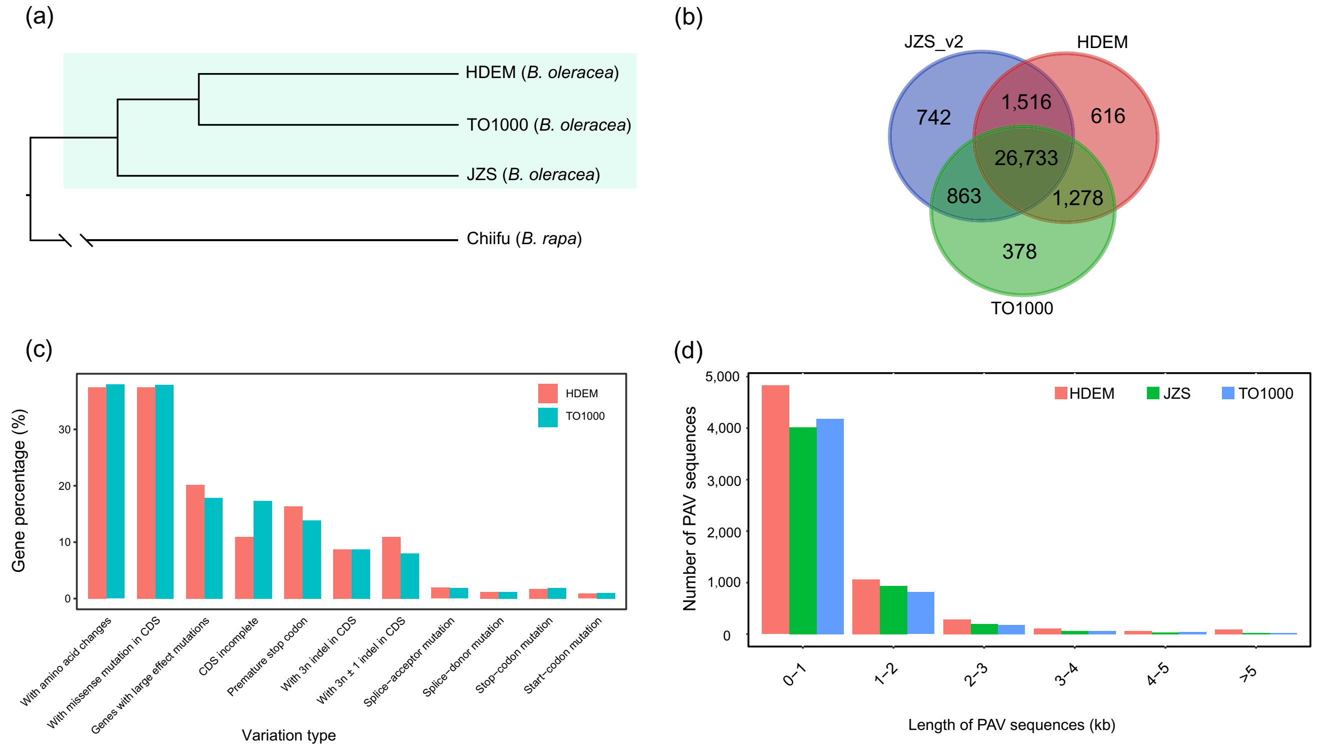


Fig. 3 Gene structural variations among JZS v2, TO1000, and HDEM genomes. **a** Phylogeny for JZS v2, TO1000, and HDEM genomes. The phylogenetic tree was constructed on the basis of 15,422 single-copy orthologous genes with Chiifu as outgroup. **b** Comparisons of syntenic genes among JZS v2, TO1000 and HDEM genomes. To ensure that the detected syntenic genes were reliable, *A. thaliana* was also included to calculate synteny. **c** Variations within genes between JZS v2 and two other *B. oleracea* genomes. **d** Length distribution of PAV segments between JZS v2, HDEM, and TO1000 genomes

1,983 and 2,874 syntenic genes being lost in the HDEM and TO1000 genomes (Fig. 3b, Table S15). Furthermore, we investigated variations within genes between JZS v2 and other two *B. oleracea* genomes; approximately 20.18% and 17.83% of the genes contained large-effect mutations such as gain or loss of stop codons, splice-donor or splice-acceptor or other major protein difference variations in HDEM and TO1000, respectively (Fig. 3c, Table S16). Also, we calculated syntenic genes similarity between two of the three genomes. The results showed that 27.94%, 27.57%, and 18.93% of pair genes in JZS v2 with HDEM, JZS v2 with TO1000, and HDEM with TO1000 genomes, respectively, have considerable structural variation (the similarity of paired protein sequences of syntenic genes less than 90%) (Fig. S7). Presence/absence variations (PAVs) were used to describe sequences that were present in some genomes but absent in others (Springer et al. 2009); we used a previously reported method (Sun et al. 2018) to detect PAV sequences in each of the three *B. oleracea* subspecies genomes. We identified 5,270 JZS v2 specific genomic segments (5.00 Mb in total), 6,438 HDEM specific genomic segments (7.11 Mb in total), and 5,307 TO1000 specific genomic segments (4.78 Mb in total). Most of the PAV segments were very short (0–1 kb), and very few PAV segments were longer than 5 kb (Fig. 3d and Table S17).

Different changing patterns of LTR-RT among *B. oleracea* genomes

We detected 9,755 (62.82 Mb in total), 2,648 (14.44 Mb in total), 1,706 (8.91 Mb in total), and 7,149 (47.84 Mb in total) intact LTR-RTs in JZS v2, JZS v1, TO1000, and HDEM genomes, respectively (Table S18). Compared with JZS v1, the 7,107 (48.38 Mb in total) extra intact LTR-RTs were specifically assembled in JZS v2 (Table S18, Fig. S8). We calculated the insertion times of all of the intact LTR-RTs in the JZS v2 genome and found a new round of LTR-RT burst event in the new assembly, mainly due to the recent (0–1 MYA) large-scale expansion of LTR-RTs. Approximately 52% of JZS v2 intact LTR-RTs were formed around 0–1 MYA, whereas only approximately 30% of intact LTR-RTs in the JZS v1 genome corresponded to this time (Fig. 4).

The changing patterns of LTR-RT around 0–1 (mostly 0–0.4) MYA were different between the JZS v2 and HDEM genomes (Fig. 5). Among the detected intact LTR-RTs, the number of those formed during 0–1 and 0–0.4 MYA in the JZS v2 genome was 2.20 and 3.78 times more than that of HDEM genome (5,075 and 2,310 LTR-RTs in JZS v2 and HDEM genomes during 0–1 MYA, 2,948 and 780 LTR-RTs in JZS v2 and HDEM genomes during 0–0.4 MYA). The rate of Gypsy-like LTR-RT amplification in the JZS v2 genome displayed a pattern of continuous increase, that was absent in HDEM (Fig. 5). During the last one million years, the number of Gypsy-like LTR-RTs in JZS v2 was 2.13 times that of the HDEM genome; however, it was 4.12 times during 0–0.4 MYA. Furthermore, the recent (0–1 MYA) amplification patterns of *athila*, *tat,* and *Del* families, members of the gypsy superfamily, were consistent with the amplification patterns of Gypsy-like LTR-RTs in the JZS v2 and HDEM genomes. In the JZS genome, the numbers of *athila*, *tat* and *Del* families amplified in the last 1 million years were 3.58, 5.70, and 4.08 times that of the HDEM genome, respectively (Fig. 5, Table S19, S20). These results revealed that the striking difference in the recent Gypsy-like LTR-RTs expansion might contribute to *B. oleracea* genome diversification.

The insertion of genome-specific LTR-RT was related to syntenic gene modification

The insertion of genome-specific LTR-RTs was closely related to syntenic gene modification, which might have contributed to *B. oleracea* genome diversification. In this work, syntenic gene modification referred to a syntenic gene that was modified due to the insertion of genome-specific

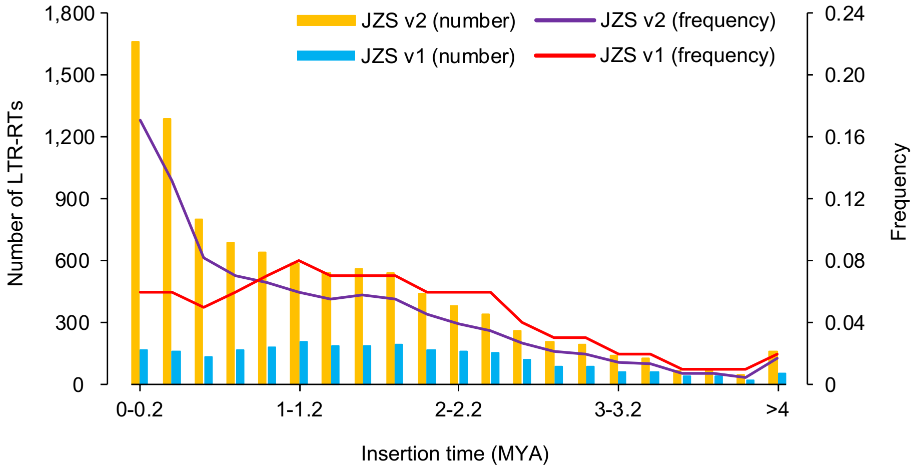


Fig. 4 Insertion time of all intact LTR-RTs in JZS v2 and JZS v1 assemblies

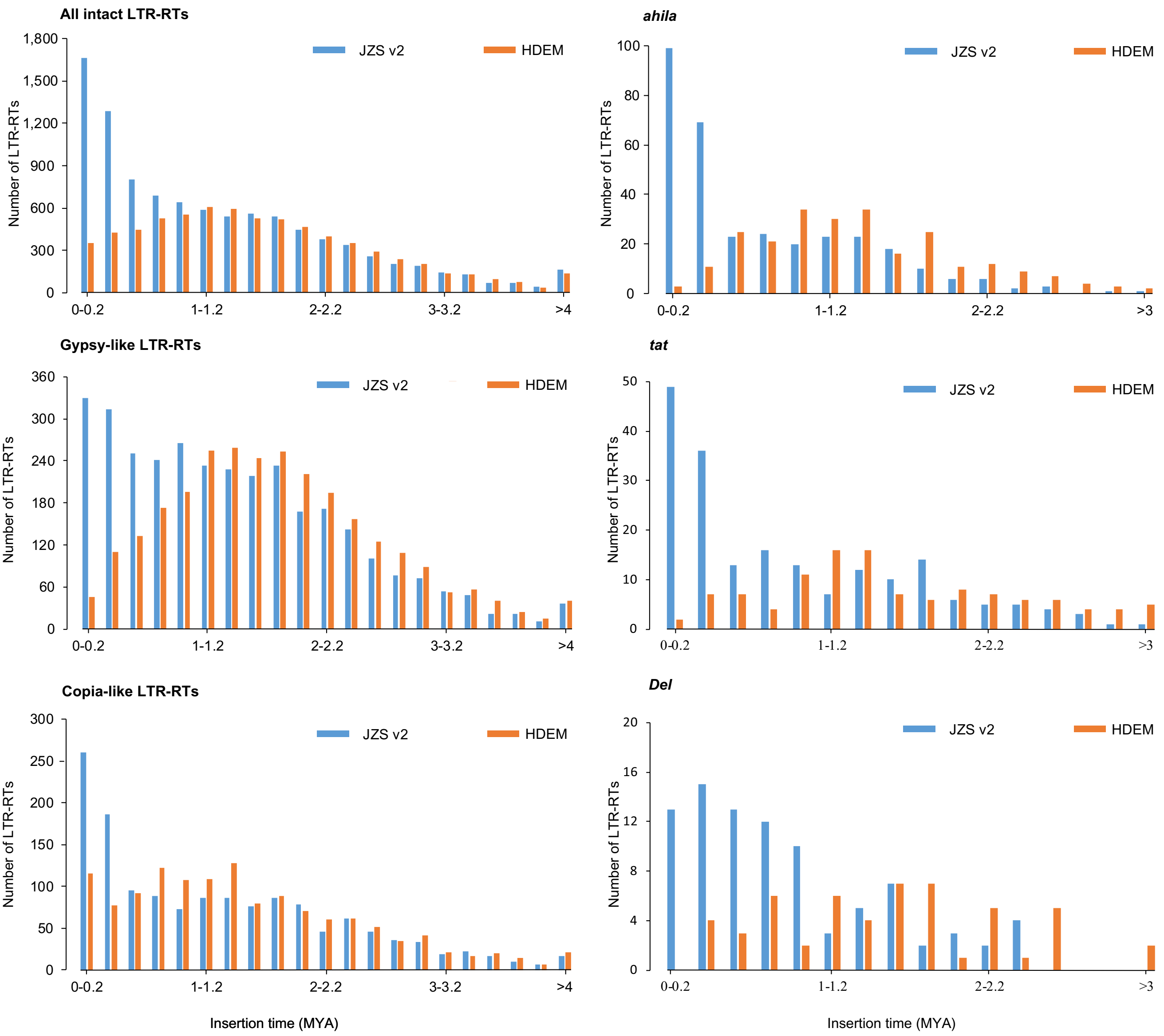


Fig. 5 Insertion time of LTR-RT in JZS v2 and HDEM genomes. The left and right panels depict the predicted insertion time of LTR-RT (all, Gypsy-like and Copia-like LTR-RTs) and three specific families (*athila*, *tat*, *del*) of the Gypsy superfamily

LTR-RTs (Fig. 6a, Table S21). In total, we detected the insertion of 79, 38, and 26 genome-specific LTR-RTs in the JZS v2, HDEM, and TO1000 genomes, respectively, that were closely related to 127 modified syntenic genes in the three genomes (Table S22). First, we identified syntenic gene pairs between *A. thaliana* and JZS v2, TO1000 and HDEM genomes and obtained a syntenic gene list. It was found that 2,272, 1,983 and 2,874 syntenic genes could not be detected in the JZS v2, HDEM, and TO1000 genomes, respectively (Fig. 3b, Table S13). Interestingly, we found the insertion of genome-specific LTR-RTs were associated with these undetectable genes. In the JZS v2, HDEM and TO1000 genomes, the insertion of 79, 38, and 26 genome-specific LTR-RTs were associated with 71, 38 and 23 modified syntenic genes, and 67 of 71, 34 of 38 and 19 of 23 modified genes were genome-specific modified syntenic genes, respectively (Fig. S9, Table S22). To investigate the functions of the modified gene, we used the corresponding syntenic gene in *A. thaliana* to represent the modified gene. According to the GO annotation of these modified genes (we used the syntenic genes in *A. thaliana* as representatives), 125 of the 127 modified genes had 158 GO terms; 34.66% (44 of 127), 33.86% (43 of 127), and 26.77% (34 of 127) genes had GO:0,008,150, GO:0,003,674 and GO:0,005,575, respectively (Table S23). GO enrichment revealed that these modified genes were mainly related to cells, cell parts, and endomembrane system functions (P value < 0.01) (Fig. 6b). These results indicated that the insertion of genome-specific

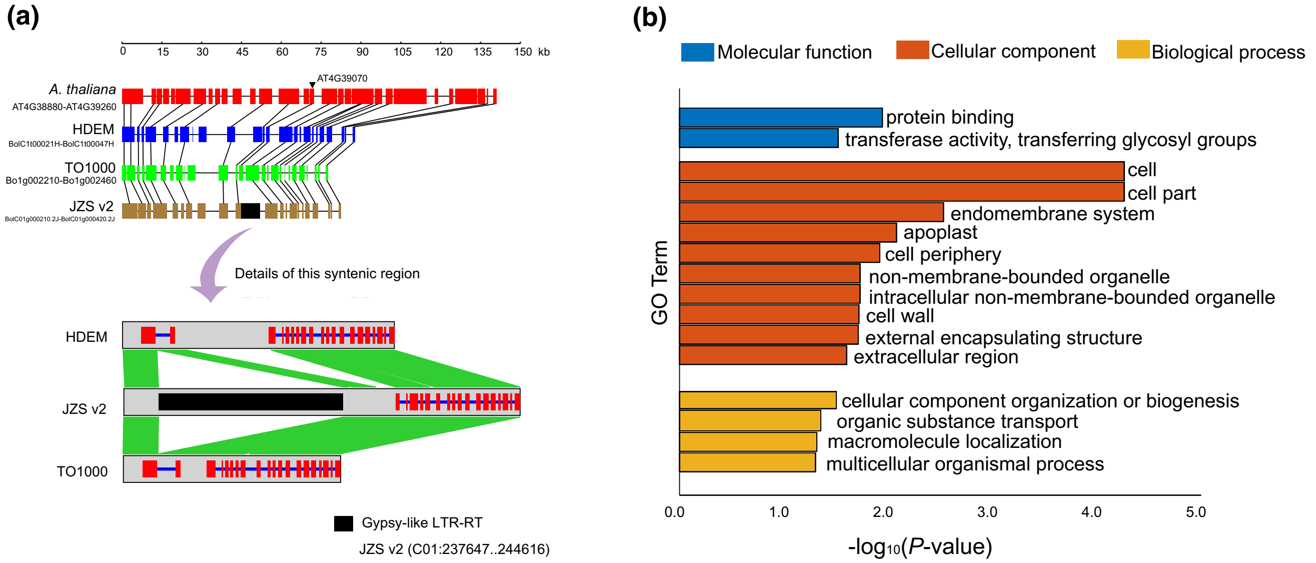


Fig. 6 The insertion of genome-specific LTR-RT related to gene modification. **a** An example of the insertion of a specific Gypsy-like LTR-RT related to gene modification in the JZS genome. **b** GO enrichment analysis of all modified genes in JZS v2, TO1000, and HDEM genomes. We used the syntenic genes in *A. thaliana* as representatives to investigate GO terms of modified genes in the three genomes

LTR-RTs could be closely related to syntenic gene modifications that might contribute to genome diversification in *B. oleracea*.

Discussion

B. oleracea is one of the diploid species in the famous "Brassica U's triangle" (Nagaharu, 1935), that includes many different subspecies/morphotypes: cabbage, kale, cauliflower, broccoli, kailan, Brussels sprouts, and kohlrabi (Cheng et al. 2016). As a reference genome for the heading type *B. oleracea* (cabbage), JZS has been widely used for genomic analysis (Liu et al. 2014). Here we reported a greatly improved assembly of the JZS genome, by taking an advantage of our newly developed misjoins correction pipeline in the process of anchoring scaffolds to chromosomes by Hi-C data. Although the whole genome Hi-C contact map revealed high-quality of our assembly, there is a small region on C05 that showed not clear contiguity. Further analysis revealed that it was mainly caused by the centromeric region (Fig. 1, Table S12). The high-quality reference genome of the heading type *B. oleracea* provided a more complete gene set and TE sequences, thus enabling comparisons of genome structures and making it possible to investigate diversification among subspecies genomes.

The large phenotypic differences among the seven subspecies make it interesting to investigate diversification among *B. oleracea*. It was reported that large genomic sequence structure variations and genetic variations were associated with subspecies diversification. In the rice genome, an AA-genome-specific inversion (~ 170 kb, bearing 14 orthologous genes) took place ~ 4.4 million years after the split with BB-genome species (Stein et al. 2018). In the maize genome, more than 10% of the annotated genes were nonsyntenic and more than 20% of the predicted genes had either large-effect mutations or large structural variations between B73 and Mo17 genomes (Sun et al. 2018). In our study, we also detected large genome-specific structure variation and extensive gene structure variation among the three subspecies. We detected three TO1000 specific inversions (on C01, C05, and C09) (Fig. S10). Despite the high quality of the TO1000 genome assembly, we still suspected that it might contain assembly errors. However, we had no other evidence to support the reliability of the TO1000 assembly (i.e., Hi-C data or BioNano data). In addition, approximately 17% of the annotated genes were nonsyntenic in each subspecies of *B. oleracea*, a value significantly higher than in other species (10.66% and 3.60% between two subspecies of maize and rice) (Sun et al. 2018). Although we also identified PAV sequences of the three *B. oleracea* genomes, we were unable to verify it experimentally, which made it impossible to calculate the error rate in calling PAVs.

LTR-RT compositions are often observed to differ among different subspecies within a species, thus supporting their importance in the formation of subspecies (Ammiraju et al. 2007; Hawkins et al. 2006; Vitte et al. 2007). In *Capsicum* genomes, the amplification of *athila* LTR-RTs, members of the gypsy superfamily, led to genome expansion in *C. baccatum* (Kim et al. 2017). In

Oryza species, lineage-specific massive LTR-RT bursts were detected in very recently diverged AA-genome *Oryza* species (Zhang and Gao 2017), and it was also reported that several lineage-specific transpositional bursts occurred in 13 domesticated and wild rice relatives (Stein et al. 2018). These results revealed that LTR-RTs were important drivers of speciation and diversification. In *Brassica*s, LTR-RTs played an important role of genome evolution. It was reported that the large proportion of LTR-RT and multiple rounds of LTR-RT bursts occurred in *Brassica* species, as well as uneven distribution and formation of LTR-RT hotspots on chromosomes (Cai et al. 2018; Yang et al. 2016; Zhang et al. 2018). However, the dynamics of subspecies-specific LTR-RT amplification are largely unknown in *Brassica*s. In the present study, we found genome-specific amplification of Gypsy-like LTR-RTs around 0–1 million years ago (MYA). In particular, the *athila*, *tat,* and *Del* families were extensively amplified in JZS during the last 0–1 MYA. These results reinforced the notion that the striking difference in the distribution of recent Gypsy-like LTR-RTs and the varied amplification patterns of the three specific families were closely related to *B. oleracea* genome diversification.

The insertion of genome-specific LTR-RT was closely associated with gene function. In cucumber, the glabrous mutation was controlled by a single recessive locus *csgl3*, and the loss-of-function of *CsGL3* in the mutation was due to the insertion of an LTR-RT in the 4th exon of *CsGL3* (Pan et al. 2015). In *B. rapa*, it was reported that a transposon insertion in the second intron of *BrFT2* (*BrFT2* was involved in flowering time regulation in *B. rapa*) was detected in one of the recombinant inbred line (RIL) parental lines; the *BrFT2* transcript was not present in the parental line that harbored the mutated allele, and RILs carrying only the mutated *BrFT2* allele showed delayed flowering (Zhang et al. 2015). In the bread wheat genome, enrichment of TE families in gene promoters was reported, and this was conserved between the A, B, and D subgenomes (Wicker et al. 2018), and it also has been reported that TEs were directly related to gene duplication and specific gene family expansion (Hoen et al. 2006; Kong et al. 2007). These results revealed the important role of the relationship between the insertions of LTR-RTs and gene functions. In this work, we found that the insertions of subspecies-specific LTR-RTs were related to gene modifications. In total, we found 143 LTR-RTs that were closely associated with 127 modified genes. We strongly believe that there were more modified genes associated with the insertions of LTR-RTs, since only genes that were colinear with *A. thliana* were used for detecting modified genes. However, we still could not explain how the genome-specific insertion of LTR-RTs induced the gene modifications that may have contributed to genomes diversification. LTR-RTs not only change the genomic structures through rapid self-replication to fuel the rapid turnover of intergenic regions but it may also induce gene modifications to direct changes in the functions of genes, thus driving genome diversification.

Materials and methods

Sample preparation and genome sequencing

B. oleracea sp. *capitata* homozygous line JZS (heading type) was used for sequencing and de novo assembly (Liu et al. 2014). High-quality genomic DNA was extracted from leaf tissues using a modified cetyltrimethylammonium bromide (CTAB) method (Allen et al. 2006), and then, the genomic DNA used for Illumina and PacBio library construction and sequencing. Libraries with an insert size of 20 kb for SMRT PacBio genome sequencing were constructed as previously reported (Pendleton et al. 2015), and these PacBio libraries were sequenced on the PacBio Sequel platform (Pacific Biosciences). Libraries for Illumina paired-end genome sequencing were constructed according to the standard manufacturer's protocol (Illumina). Illumina reads were generated from three paired-end sequencing libraries with insertion sizes of 250 bp, 350 bp, and 500 bp, and these three libraries were sequenced on an Illumina platform with a paired-end sequencing strategy. The Hi-C libraries of JZS were constructed following the pipelines described in a previous study (Grob et al. 2014), and the resulting libraries were submitted to an Illumina HiSeq 4000 sequencing device with 2 × 125 bp reads.

De novo assembly of PacBio and Illumina reads

A hybrid assembly strategy was used to complete the assembly of the JZS draft genome. Approximately 38 Gb (~ 61 ×) PacBio SMRT reads and 53 Gb (~ 84 ×) Illumina reads were used for scaffold assembly with MaSuRCA (Zimin et al. 2017). As recommended by the software developer, we used the raw Illumina and PacBio reads and the default parameters to hybrid assemble the JZS draft genome. This procedure resulted in a total assembly length of 561.11 Mb with an N50 length of 3.05 Mb (Table S2), and then BUSCO (Waterhouse et al. 2018) was used to perform a preliminary assessment of the assembly results.

Correction of misjoins in scaffolds

To detect misjoins in the hybrid assembled scaffolds, we developed a reliable misjoins correction pipeline (named MisjoinDetect) based on the Hi-C data. Our pipeline included the following three main steps. First, detection of regions of candidate misjoins. fastp (Chen et al. 2018) was used to filter low-quality Hi-C reads, and then, clean reads were mapped onto the initial assembled scaffold sequences by HiC-Pro (Servant et al. 2015). Meanwhile, scaffolds were divided into different segments according to the fixed bin size, and the interaction values between all of the fragments within each scaffold were extracted and used to form an interaction matrix. The regions of the candidate misjoins were defined based on the difference in the interaction values of adjacent bins. Second, we determined the locations of the breakpoints. The program first searched for gap information in the candidate area. If the gap existed, the gap area would be deleted, and the location would be defined as a breakpoint. (These errors were caused by different contigs being incorrectly connected by the de novo assembly software.) If it did not exist, the program would use the midpoint of the two genes in the middle of the candidate region as a potential breakpoint to ensure that the gene sequence was intact. Moreover, we provided an additional script that relied on a collinear list of genes from the related species. Based on the collinearity results, we could more accurately determine the location of the breakpoint to ensure a more complete syntenic region. Finally, clean scaffolds sequences and an Hi-C contact map were obtained for each corrected scaffold.

Construction of pseudomolecules and evaluation

The highly efficient pipelines developed by Aiden Lab (https://aidenlab.org) were used to anchor the 02–12 clean scaffolds onto the 9 chromosomes. Juicer (Durand et al. 2016) was used to align clean Hi-C reads to corrected scaffolds, and then, 3D-DNA (Dudchenko et al. 2017) was used to anchor corrected scaffolds onto chromosomes (-m haploid -e). Finally, Juicebox (Robinson et al. 2018) was used to visualize linked results by 3D-DNA, and we manually determined chromosome boundaries and some small errors. D-GENIES (Cabanettes and Klopp 2018) was used to perform synteny alignments between JZS assembly and HDEM. BUSCO (Waterhouse et al. 2018) was further used to evaluate the genome-assembly completeness, and 1440 single-copy orthologous genes were used as a dataset. Meanwhile, LAI index (Ou et al. 2018) was also used to evaluate the continuity of the assembly.

Gene prediction and function annotation

We used RepeatMasker (Tarailo-Graovac and Chen 2009) to mask the whole genome sequences, and then, gene prediction was based on the masked genomic sequences. The gene prediction process consisted of the following four steps. First, extraction of ab initio gene models. AUGUSTUS (https://github.com/Gaius-Augustus/Augustus) and GeneMark (Besemer and Borodovsky 2005) were used for de novo gene prediction. Second, GeneWise (Birney et al. 2004) with default parameters was used to detect homologous gene models. Third, detection of genetic models was supported by RNA-seq data. All of the transcriptome data were downloaded from NCBI (SRS472277, SRS472450, GSM1052958, GSM1052959, GSM1052960, GSM1052961, GSM1052962, GSM1052963, GSM1052964). Then, Trinity (Grabherr et al. 2011) and PASA (Haas et al. 2003) were used to predict genes. Finally, EVidenceModeler (Haas et al. 2008) was used to merge all of the gene model predictions. In our study, we filtered all low-quality gene models. Low-quality annotated genes included genes without start codons, genes ending without a stop codon, and gene lengths less than 50 bp. InterProScan (Hunter et al. 2009) was used to annotate motifs and domains, and we extracted gene ontology from output of InterProScan. Each of these annotation datasets were freely available from the BRAD database. https://brassicadb.org/brad/datasets/pub/Genomes/Brassica_oleracea/V2.0/.

Annotate transposable elements and identify LTR-RTs associated with syntenic gene modification

EDTA package (Ou et al. 2019) was used to construct a non-redundant TE library, and then, transposable elements (TE) were annotated and classified using RepeatMasker (Tarailo-Graovac and Chen 2009) with default parameters. Intact LTR-RTs in JZS v2, HDEM and TO1000 were identified using LTR_Finder (Xu and Wang 2007) with the parameters '-D 15,000—d 1000—L 7000—l 100—p 20—C—M 0.9′, and LTR_retriever (Ou and Jiang 2018) was further used to categorize the detected LTR-RTs into the subgroups of Copia-like and Gypsy-like LTR-RTs. The insertion time of the intact LTR-RT was extracted from outputs of LTR_retriever (the base substitution rate 1.3×10^{-8} was adopted in our work).

First, we calculated syntenic gene pairs between *A. thaliana* and JZS v2, TO1000 and HDEM genomes and obtained a syntenic gene list. Then, we identified whether the LTR-RT sequence was included in the region where the corresponding syntenic gene was not detected. If the LTR-RT could be detected, we extracted the upstream and downstream syntenic genes and manually checked.

Phylogenetic inference

First, we used OrthoFinder (Emms and Kelly 2015) to detect single copy genes among Chiifu, TO1000, HDEM and JZS genomes (Belser et al. 2018; Parkin et al. 2014; Zhang et al. 2018). The coding sequences (CDS) of 15,422 single-copy gene families within the four genomes were aligned at the nucleotide level using MAFFT (Katoh et al. 2005), and well-aligned regions were extracted using Gblock (v0.91b) (Talavera and Castresana 2007) with $-t=p$, $-b4=5$, $-b5=h$. Finally, we used RAxML (Stamatakis 2014) with PROTGAMMAWAG model and 500 bootstrap replicates to construct the tree.

Genome blocks and centromere detection in the JZS genome

SynOrths (Cheng et al. 2012b) was used to perform synteny analysis in this work. We used syntenic gene pairs between JZS v2 and *A. thaliana* to construct three sub-genomes and defined GBs in the JZS v2 genome. First, we conducted syntenic analysis between JZS v2 and *A. thaliana*, and then, the least fractionated (LF), the medium fractionated (MF1) and the most fractionated (MF2) subgenomes of JZS v2 were built by previously reported methods (Cheng et al. 2012a). Then, using the method of defining the genomic blocks (GBs) in *B. rapa* (Zhang et al. 2018), we defined GBs in the JZS v2 genome. Centromere regions in each chromosome were defined by using mummer (Kurtz et al. 2004) to map centromeric repeat sequences (CentBr, CRB, TR238, and PCRBr) (Koo et al. 2011; Lim et al. 2007) to genome sequences with parameters (–maxmatch—g 500—c 16—l 16).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00122-020-03664-3) contains supplementary material, which is available to authorized users.

Acknowledgments We would like to thank Ian Bancroft and Zhesi He for the help of the evaluation of JZS v2 assembly.

Author contribution statement XC wrote the manuscript. XW, JW and FC designed the experiments. XC performed the experiments. XW, JW, FC, JL, RL and KZ helped to improve the manuscript. All authors agree with the current statement.

Funding This work is supported by the National Program on Key Research Project (2016YFD0100307), the National Natural Science Foundation of China (NSFC grants 31630068), Central Public-interest Scientific Institution Basal Research Fund (No.Y2017PT52), China Agriculture Research System, the Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences, the Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Ministry of Agriculture, P.R. China.

Code availability The misjoins correction pipeline can be downloaded from https://github.com/caixu0518/MisjoinDetect.

Data availability The genome assembly and gene annotations are freely available through BRAD website (https://brassicadb.org/brad/datasets/pub/Genomes/Brassica_oleracea/V2.0/) or in the Genome Warehouse database (Members 2019) under bioproject number (PRJCA001832) and accession number GWHAASO00000000 (https://bigd.big.ac.cn/gwh).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interests.

References

Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF (2006) A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc 1:2320–2325

Ammiraju JS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, Walling JG, Ma J, Talag J, Brar DS, SanMiguel PJ, Jiang N, Jackson SA, Panaud O, Wing RA (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. Plant J Cell Mol Biol 52:342–351

Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, Genete M, Berrabah W, Chevre AM, Delourme R, Deniot G, Denoeud F, Duffe P, Engelen S, Lemainque A, Manzanares-Dauleux M, Martin G, Morice J, Noel B, Vekemans X, D'Hont A, Rousseau-Gueutin M, Barbe V, Cruaud C, Wincker P, Aury JM (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nature plants 4:879–887

Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33:W451–454

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988–995

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 31:1119–1125

Cabanettes F, Klopp C (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958

Cai X, Cui Y, Zhang L, Wu J, Liang J, Cheng L, Wang X, Cheng F (2018) Hotspots of Independent and Multiple Rounds of LTR-retrotransposon Bursts in *Brassica* Species. Hortic Plant J 4:165–174

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890

Cheng F, Sun R, Hou X, Zheng H, Zhang F, Zhang Y, Liu B, Liang J, Zhuang M, Liu Y, Liu D, Wang X, Li P, Liu Y, Lin K, Bucher J, Zhang N, Wang Y, Wang H, Deng J, Liao Y, Wei K, Zhang X, Fu L, Hu Y, Liu J, Cai C, Zhang S, Zhang S, Li F, Zhang H, Zhang J, Guo N, Liu Z, Liu J, Sun C, Ma Y, Zhang H, Cui Y, Freeling MR, Borm T, Bonnema G, Wu J, Wang X (2016) Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. Nat Genet 48:1218–1224

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X (2012a) Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. PLoS ONE 7:e36442

Cheng F, Wu J, Fang L, Wang X (2012b) Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species. Front Plant Sci 3:198

Du J, Tian Z, Bowen NJ, Schmutz J, Shoemaker RC, Ma J (2010) Bifurcation and enhancement of autonomous-nonautonomous retrotransposon partnership through LTR Swapping in soybean. Plant Cell 22:48–61

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL (2017) *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356:92–95

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell systems 3:95–98

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

Ghurye J, Pop M, Koren S, Bickhart D, Chin CS (2017) Scaffolding of long read assemblies using long range contact information. BMC genomics 18:527

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Grob S, Schmid MW, Grossniklaus U (2014) Hi-C analysis in Arabidopsis identifies the KNOT, a structure with similarities to the flamenco locus of Drosophila. Mol Cell 55:678–693

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31:5654–5666

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. Genome Biol 9:R7

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res 16:1252–1261

Hoen DR, Park KC, Elrouby N, Yu Z, Mohabir N, Cowan RK, Bureau TE (2006) Transposon-mediated expansion and diversification of a family of ULP-like genes. Mol Biol Evol 23:1254–1268

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37:D211–215

Jiao WB, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 36:64–70

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518

Kim S, Park J, Yeom SI, Kim YM, Seo E, Kim KT, Kim MS, Lee JM, Cheong K, Shin HS, Kim SB, Han K, Lee J, Park M, Lee HA, Lee HY, Lee Y, Oh S, Lee JH, Choi E, Choi E, Lee SE, Jeon J, Kim H, Choi G, Song H, Lee J, Lee SC, Kwon JK, Lee HY, Koo N, Hong Y, Kim RW, Kang WH, Huh JH, Kang BC, Yang TJ, Lee YH, Bennetzen JL, Choi D (2017) New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biol 18:210

Kong H, Landherr LL, Frohlich MW, Leebens-Mack J, Ma H, dePamphilis CW (2007) Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. Plant J Cell Mol Biol 50:873–885

Koo DH, Hong CP, Batley J, Chung YS, Edwards D, Bang JW, Hur Y, Lim YP (2011) Rapid divergence of repetitive DNAs in *Brassica* relatives. Genomics 97:173–185

Kopsell DA, Kopsell DE (2006) Accumulation and bioavailability of dietary carotenoids in vegetable crops. Trends Plant Sci 11:499–507

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12

Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY, Kwon SJ, Kim J, Choi BS, Lim MH, Jin M, Kim HI, de Jong H, Bancroft I, Lim Y, Park BS (2007) Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related *Brassica* species. Plant J Cell Mol Biol 49:173–183

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IA, Zhao M, Ma J, Yu J, Huang S, Wang X, Wang J, Lu K, Fang Z, Bancroft I, Yang TJ, Hu Q, Wang X, Yue Z, Li H, Yang L, Wu J, Zhou Q, Wang W, King GJ, Pires JC, Lu C, Wu Z, Sampath P, Wang Z, Guo H, Pan S, Yang L, Min J, Zhang D, Jin D, Li W, Belcram H, Tu J, Guan M, Qi C, Du D, Li J, Jiang L, Batley J, Sharpe AG, Park BS, Ruperao P, Cheng F, Waminal NE, Huang Y, Dong C, Wang L, Li J, Hu Z, Zhuang M, Huang Y, Huang J, Shi J, Mei D, Liu J, Lee TH, Wang J, Jin H, Li Z, Li X, Zhang J, Xiao L, Zhou Y, Liu Z, Liu X, Qin R, Tang X, Liu W, Wang Y, Zhang Y, Lee J, Kim HH, Denoeud F, Xu X, Liang X, Hua W, Wang X, Wang J, Chalhoub B, Paterson AH (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat commun 5:3930

Members BIGDC (2019) Database resources of the BIG Data Center in 2019. Nucleic Acids Res 47:D8–D14

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997–1002

Nagaharu U (1935) Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization. Jpn J Bot 7(7):389–452

Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, Tanisaka T, Wessler SR (2006) Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci USA 103:17620–17625

Ou S, Chen J, Jiang N (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res 46:e126

Ou S, Jiang N (2018) LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol 176:1410–1422

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, Jiang N, Hirsch CN, Hufford MB (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 20:275

Pan Y, Bo K, Cheng Z, Weng Y (2015) The loss-of-function GLABROUS 3 mutation in cucumber is due to LTR-retrotransposon insertion in a class IV HD-ZIP transcription factor gene *CsGL3* that is epistatic over CsGL1. BMC Plant Biol 15:302

Parkin IA, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL, Denoeud F, Belcram H, Links MG, Just J, Clarke C, Bender T, Huebert T, Mason AS, Pires JC, Barker G, Moore J, Walley PG, Manoli S, Batley J, Edwards D, Nelson MN, Wang X, Paterson AH, King G, Bancroft I, Chalhoub B, Sharpe AG (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biol 15:R77

Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W, Anantharaman T, Hastie A, Dai H, Fritz MH, Cao H, Cohain A, Deikus G, Durrett RE, Blanchard SC, Altman R, Chin CS, Guo Y, Paxinos EE, Korbel JO, Darnell RB, McCombie WR, Kwok PY, Mason CE, Schadt EE, Bashir A (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 12:780–786

Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza* australiensis, a wild relative of rice. Genome Res 16:1262–1269

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinf 13:278–289

Robinson JT, Turner D, Durand NC, Thorvaldsdottir H, Mesirov JP, Aiden EL (2018) Juicebox.js provides a cloud-based visualization system for Hi-C data. Cell systems 6:256–258

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 16:259

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5:e1000734

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, Wei S, Wang J, Liao Y, Wang M, Jacquemin J, Becker C, Kudrna D, Zhang J, Londono CEM, Song X, Lee S, Sanchez P, Zuccolo A, Ammiraju JSS, Talag J, Danowitz A, Rivera LF, Gschwend AR, Noutsos C, Wu CC, Kao SM, Zeng JW, Wei FJ, Zhao Q, Feng Q, El Baidouri M, Carpentier MC, Lasserre E, Cooke R, Rosa Farias DD, da Maia LC, Dos Santos RS, Nyberg KG, McNally KL, Mauleon R, Alexandrov N, Schmutz J, Flowers D, Fan C, Weigel D, Jena KK, Wicker T, Chen M, Han B, Henry R, Hsing YC, Kurata N, de Oliveira AC, Panaud O, Jackson SA, Machado CA, Sanderson MJ, Long M, Ware D, Wing RA (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. Nat Genet 50:285–296

Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X, Liu H, Ma X, Jiao Y, Wang B, Wei X, Stein JC, Glaubitz JC, Lu F, Yu G, Liang C, Fengler K, Li B, Rafalski A, Schnable PS, Ware DH, Buckler ES, Lai J (2018) Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet 50:1289–1295

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577

Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics Chapter 4:Unit 4 10

Vitte C, Panaud O, Quesneville H (2007) LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC genomics 8:218

Wang W, Guan R, Liu X, Zhang H, Song B, Xu Q, Fan G, Chen W, Wu X, Liu X, Wang J (2019) Chromosome level comparative analysis of *Brassica* genomes. Plant Mol Biol 99:237–249

Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543–548

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramirez-Gonzalez RH, De Oliveira R, International Wheat Genome Sequencing C, Mayer KFX, Paux E, Choulet F (2018) Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol 19:103

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–268

Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, Hu Z, Chen S, Pental D, Ju Y, Yao P, Li X, Xie K, Zhang J, Wang J, Liu F, Ma W, Shopan J, Zheng H, Mackenzie SA, Zhang M (2016) The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. Nat Genet 48:1225–1232

Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, Liang J, Cai C, Liu Z, Liu B, Wang F, Li S, Liu F, Li X, Cheng L, Yang W, Li MH, Grossniklaus U, Zheng H, Wang X (2018) Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. Hortic Res 5:50

Zhang QJ, Gao LZ (2017) Rapid and recent evolution of LTR Retrotransposons drives rice genome evolution during the speciation of AA-Genome Oryza species. G3: Genes Genomes Genet 7(6):1875–1885. https://doi.org/10.1534/g3.116.037572

Zhang X, Meng L, Liu B, Hu Y, Cheng F, Liang J, Aarts MG, Wang X, Wu J (2015) A transposon insertion in *FLOWERING LOCUS T* is associated with delayed flowering in *Brassica rapa*. Plant Sci Int J Exp Plant Biol 241:211–220

Zhang X, Zhang S, Zhao Q, Ming R, Tang H (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat plants 5:833–845

Zhou M, Hu B, Zhu Y (2017) Genome-wide characterization and evolution analysis of long terminal repeat retroelements in moso bamboo (*Phyllostachys edulis*). Tree Genet Genomes 13:1–12

Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marcais G, Yorke JA, Dvorak J, Salzberg SL (2017) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res 27:787–792