Horticultural Plant Journal

Available online at www.sciencedirect.com
The journal's homepage: http://www.keaipublishing.com/en/journals/horticultural-plant-journal

Nuclear and chloroplast genome diversity revealed by low-coverage whole-genome shotgun sequence in 44 *Brassica oleracea* breeding lines

Sampath Perumal[a,b,c,1], *Nomar Espinosa Waminal*[a,d,1], *Jonghoon Lee*[a,e,1], *Hyun-Jin Koo*[a], *Boem-soon Choi*[f], *Jee Young Park*[a], *Kyounggu Ahn*[e], and *Tae-Jin Yang*[a,*]

[a] *Department of Agriculture, Forestry and Bioresources, Plant Genomics and Breeding Institute, and Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul 08826, Republic of Korea*
[b] *Global Institute for Food Security, University of Saskatchewan, 421 Downey Road, Saskatoon, SK S7N 4L8, Canada*
[c] *Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon S7N 0X2, Canada*
[d] *Chromosome Research Institute, Department of Life Science, Sahmyook University, Seoul 01795, Republic of Korea*
[e] *Joeun Seeds. Co., Ltd, Goesan-Gun, Chungcheongbuk-Do 367-833, Republic of Korea*
[f] *Phyzen Genomics Institute, 13, Seongnam-daero 331beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do 13558, Republic of Korea*

Received 25 June 2020; Received in revised form 3 August 2020; Accepted 13 October 2020

Available online 26 February 2021

ABSTRACT

Whole-genome shogun sequence (WGS) data generated by next-generation sequencing (NGS) platforms are a valuable resource for crop improvement. We produced 5–6 × WGS coverage of 44 *Brassica oleracea* breeding lines representing seven subspecies/morphotypes: cabbage, broccoli, cauliflower, kailan, kale, Brussels sprout, and kohlrabi to systematically evaluate the nuclear and chloroplast (Cp) diversity in the 44 *B. oleracea* breeding lines. We then exploited the impact of low-coverage NGS by evaluating nuclear genome diversity and assembly, annotation of complete chloroplast (Cp) genomes and 45S nuclear ribosomal DNA (45S nrDNA) sequences, and copy number variation for major repeats. Nuclear genome diversity analysis has revealed a total of 496 463 SNPs and 37 493 indels in the nuclear genome across the 44 accessions. Interestingly, some SNPs showed subspecies enrichment at certain chromosomal regions. The assembly of complete Cp genomes contained 153 361–153 372 bp with 37 variants including SNPs and indels. The 45S nrDNA transcription unit was 5 802 bp long with a total of 31 SNPs from the 44 lines. The phylogenetic tree inferred from the nuclear and Cp genomes coincided and clustered broccoli, cauliflower, and kailan in one group and cabbage, Brussels sprout, kale, and kohlrabi in another group. The morphotypes diverged during the last 0.17 million years. The Cp genome diversity reflected the unique cytoplasm of each subspecies, and revealed that the cytoplasm of many breeding lines was replaced and intermingled via inter-subspecies crosses during the breeding process instead. The polymorphic Cp markers provide a classification system for the cytoplasm types in *B. oleracea*. Furthermore, copy numbers of major transposable elements (TEs) showed high diversity among the 44 accessions, indicating that many TEs have become active recently. Overall, we demonstrated a comprehensive utilization of low-coverage NGS data and might shed light on the genetic diversity and evolution of diverse *B. oleracea* morphotypes.

Keywords: *Brassica oleracea* subspecies; Chloroplast genome; Nuclear ribosomal DNA; Genetic diversity; Phylogeny

[*] Corresponding author.
E-mail address: tjyang@snu.ac.kr

[1] These authors contributed equally to this work.

Peer review under responsibility of Chinese Society for Horticultural Science (CSHS) and Institute of Vegetables and Flowers (IVF), Chinese Academy of Agricultural Sciences (CAAS)

https://doi.org/10.1016/j.hpj.2021.02.004

1. Introduction

Next-generation sequencing (NGS) technologies have allowed voluminous sequence data to be generated in a relatively short time and at an affordable price. This has rekindled interest in whole-genome sequencing and resequencing analysis. NGS has resulted in decoding the whole-genome sequences of more than 470 plant species to date (Zhang et al., 2011; as of Nov 2020, http://www.plabipd.de/). The bottleneck for analysis has mainly shifted from generating sequences to developing various application methods that utilize these sequences to address relevant biological questions for crop improvement (Zhang et al., 2011).

Although the in-depth assembly of whole nuclear genomes using NGS provides concrete information for plant breeding (Pop et al., 20019; Zhang et al., 2011; Le Nguyen et al., 2019), much relevant information, such as the identification and estimation of genomic repeat elements (REs) (Macas et al., 2007; Natali et al., 2013), the assembly of organelle genomes (Kim et al., 2015b) and SNP detection (Lee et al., 2015), can be obtained even with only low-coverage NGS data (Rasheed et al., 2017; Scheben et al., 2017). The acquisition of suitable tools and knowledge to exploit low-coverage sequences is necessary, to maximize their potential in providing appropriate information to enhance crop improvement.

Brassica is the most economically important genus among 51 genera in the Brassiceae tribe of the Brassicaceae family (Rakow et al., 2004). Approximately 39 species comprise the genus (http://www.theplantlist.org/tpl1.1/search?q=brassica), with many considered to be weeds. Six species, including three diploids (*B. rapa*, AA; *B. nigra*, BB; and *B. oleracea*, CC) and three allotetraploids (*B. juncea*, AABB; *B. carinata*, BBCC; and *B. napus*, AACC), whose relationship is described as U's triangle (Nagaharu, 1935), have received more attention owing to their economic impact as sources of vegetable, condiments, fodder, and oil (Rakow et al., 2004; Cheng et al., 2015a).

Interspecific hybridization, reticulated evolution, and several rounds of whole-genome duplication events within the Brassicaceae family have promoted high inter- and intra-specific genetic and phenotypic diversity and up to over eight-fold genome size variation within the family (Lysak et al., 2005; Marhold and Lihová, 2006; Panjabi et al., 2008; Cheng et al., 2015b; Tank et al., 2015). These attributes make species of this family good models for polyploid genome evolution studies. In addition, beneficial traits from wild relatives of cultivated species, such as male sterility and resistance to pests and diseases, provide a platform for crop improvement and phylogenetic studies (Rakow et al., 2004).

Brassica oleracea is an important vegetable crop with wide morphological diversity. It is generally subdivided into six groups (Snogerup, 1980) that include kales (var. *acephala*) such as stem kale, cabbages (var. *capitata*, var. *sabauda*, var. *bullata*), including head cabbages and Brussels sprout, kohlrabi (var. *gongylodes*), inflorescence kales (var. *botrytis*, var. *italica*), which include cauliflower and broccoli, branching bush kales (var. *fruticosa*), and Chinese kale (var. *alboglabra*), which is used as a leaf vegetable (Rakow et al., 2004). The publication of the *B. oleracea* reference genomes have promoted applied and comparative research that utilizes whole-genome sequence (WGS) data (Chalhoub et al., 2014; Liu et al., 2014). Understanding the genomic effects and mechanisms that influence intra-specific variations is advantageous for crop improvement. Relevant information can be derived from low-coverage sequencing data, which can facilitate the development of crops with better qualities, and address different production demands in the changing global climate (Liu et al., 2014).

In this study, we described several uses of low-coverage NGS data for genomic studies in *B. oleracea*. We provide the complete assembly and diversity of chloroplast genomes and nuclear ribosomal DNA (nrDNA) sequences, genome-wide SNP calling, quantification of major repetitive DNAs, and discussed the potential of low coverage NGS data for comprehensive genome analysis.

2. Materials and methods

2.1. Plant materials and sequencing

A total of 44 *Brassica oleracea* accessions consisting of seven subspecies (20 cabbages, five each for broccoli, cauliflower, kohlrabi, and kale, three kailan, and one Brussels sprout) were used in this study (Table S1). We used different *B. oleracea* subspecies with diverse morphotypes such as heading-type (kailan, broccoli, cauliflower), leafy-type, stem-modifications (Brussels sprout) and tuber-forming (kale, kohlrabi). All plant materials were provided by Joeun Seeds Co. (Chungcheongbuk-Do, South Korea) and Asia Seeds Co. (Seoul, South Korea). Each breeding line was bred by cross-hybridization within or between subspecies and then genetically fixed by several rounds of self-pollination.

The whole-genome sequence data of IM_Bol_01 to 04 were obtained from previous research (Lee et al., 2015) and the remaining 40 accessions were sequenced as follows. Genomic DNAs were extracted from at least 2 g samples of young leaves, following the modified cetyltrimethylammonium bromide (CTAB) protocol (Allen et al., 2006). The quality and quantity of the DNA were examined using a NanoDrop ND-1000 (NanoDrop, USA). More than 5 μg extracted DNA was randomly sheared and quantified using the Truseq DNA PCR-free kit (Agilent, USA) according to the manufacturer's protocol. Sequencing of the shotgun libraries was performed using Illumina Hi-seq 2000 and NextSeq 500 platforms (Table S1). Fragmentation, library construction and sequencing using Hi-seq 2000 and Nextseq 500 were carried out by Macrogen (Seoul, South Korea) and Lab Genomics, Inc. (Seongnam, South Korea), respectively. All raw sequencing reads of 44 accessions were deposited to a public database called National Agricultural Biotechnology Information Center (NABIC, http://nabic.rda.go.kr) (Seol et al., 2016) and accession numbers were listed in the Table S1.

2.2. SNP calling, genotyping and kinship analysis

Raw paired-end (PE) sequences were quality filtered using the program clc_quality_trim V4.3.0.114910 (http://www.clcbio.com) with a based phred quality score > 33 and length > 250 bp. Filtered reads were aligned to the reference sequences of using the Burrows–Wheeler alignment tool (BWA) 32 with default parameters, and properly paired reads were sorted by the parameter of maximum insert size (800 bp). Read grouping and removal of PCR duplicates were performed using Picard (http://picard.sourceforge.net). The genome analysis toolkit (GATK 3.4) was used to perform local realignment of reads to correct misalignments due to the presence of indels. UnifiedGenotyper from the GATK package was used to detect variants such as SNPs and Indels (McKenna et al., 2010). Finally, the combined variants from all 44 accessions were filtered with VariantFiltration from the GATK with read depth of at least 5× to reduce the false positive variants (de Summa et al., 2017).

Kinship analysis was performed using high-quality SNPs from 44 accessions using Genome association and prediction integrated tool (GAPIT), a GWAS statistical method based on the mixed linear model (MLM), with the Efficient Mixed Model Association (EMMA) algorithm. The relationships between individuals were calculated by the VanRaden method using the mean and average cluster algorithm (Lipka et al., 2012; Tang et al., 2016). The genotyping data converted into HapMap formats were used for GWAS (Genome-wide association study) in GAPIT.

2.3. *Assembly and comparative analysis of complete chloroplast genomes and 45S nrDNA sequences*

Approximately 6 × physical coverage of WGS data (mean of 4.0 Gb) of the 645 Mb estimated genome size of *B. oleracea* (Parkin et al., 2014) was used for complete chloroplast (Cp) and 45S nrDNA assembly (Table S1). *De novo* assembly was performed to obtain Cp and nrDNA sequences following the dnaLCW pipeline using a CLC assembly cell (v. beta 4.6, Denmark) (Kim et al., 2015b). The principal contigs containing Cp genomes and nrDNA were ordered based on the reference chloroplast genome sequence (GenBank accession No. KR233156) and that of 45S nrDNA (GenBank accession No. X52322.1) (Seol et al., 2015). The 44 assembled chloroplast genome sequences were aligned and compared using the MAFFT program (http://mafft.cbrc.jp/alignment/server/) (Katoh and Toh, 2008). The maximum likelihood phylogenetic tree was created for Cp genome sequences and 45S nrDNA by MEGA 6.0 (Tamura et al., 2013) with 1 000 bootstrap replicates.

2.4. *Divergence time analysis using Cp and nrDNA sequences*

Divergence time was estimated using the Bayesian method as implemented in the BEAST2 program (Drummond et al., 2012; Bouckaert et al., 2014) based on the complete Cp and nrDNA sequences from 44 accessions. The generalized time-reversible model, GTR+I+G substitution model, was used to construct the tree topology and divergence time. The Markov Chain Monte Carlo (MCMC) algorithm was performed for 10 000 000 generations and we sampled every 1000 generations, with Yule tree prior, with an uncorrelated lognormal relaxed clock model. Tracer (v 1.6) was used to analyze the BEAST output after discarding 10% generations as burn-in and the remaining BEAST runs were used for the posterior possibilities. Tree annotator was used to estimate the divergence time. *B. rapa* was constrained to be the outgroup and the age of divergence time between *B. rapa* and *B. oleracea* was constrained by a normal distribution with a mean of 4

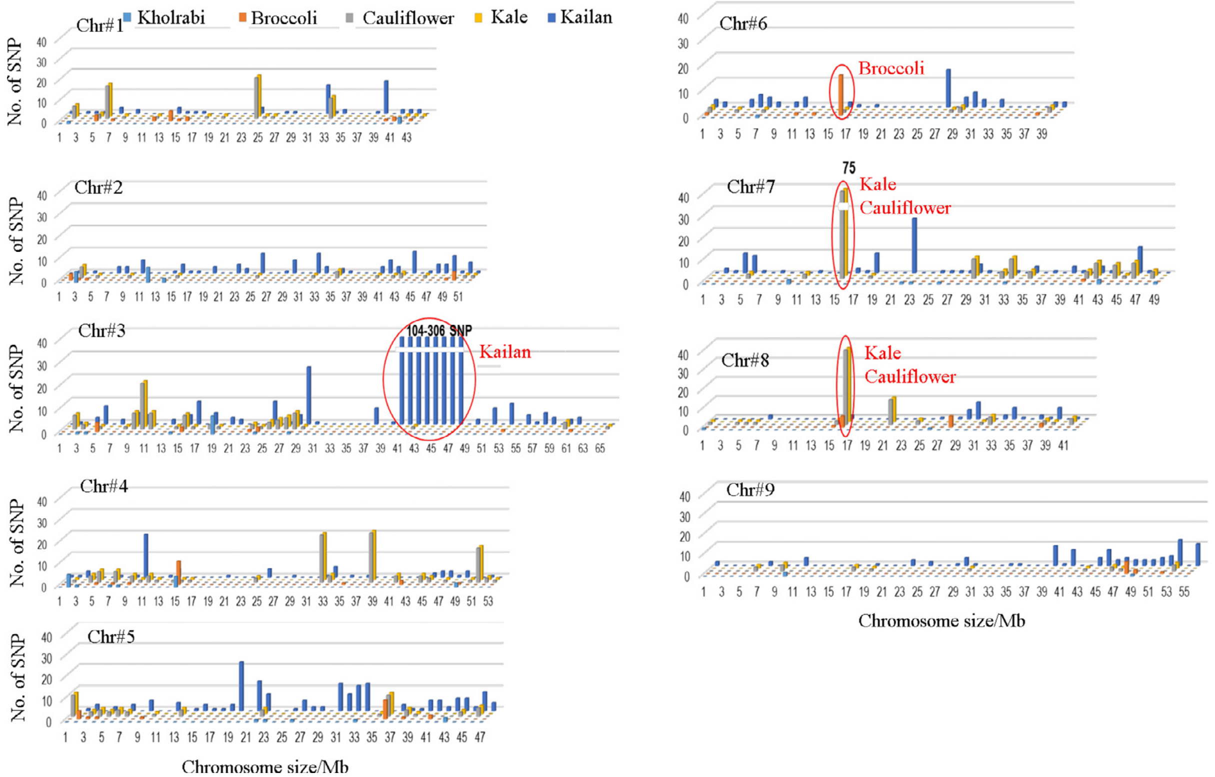

Fig. 1 Chromosome-wide distribution of subspecies-specific single nucleotide variants identified in seven *Brassica oleracea* subspecies

The numbers of SNPs are shown in 1-Mb sliding windows. Chromosome blocks containing species-specific variations are highlighted with red circles and their function word art can be found in Fig. S1.

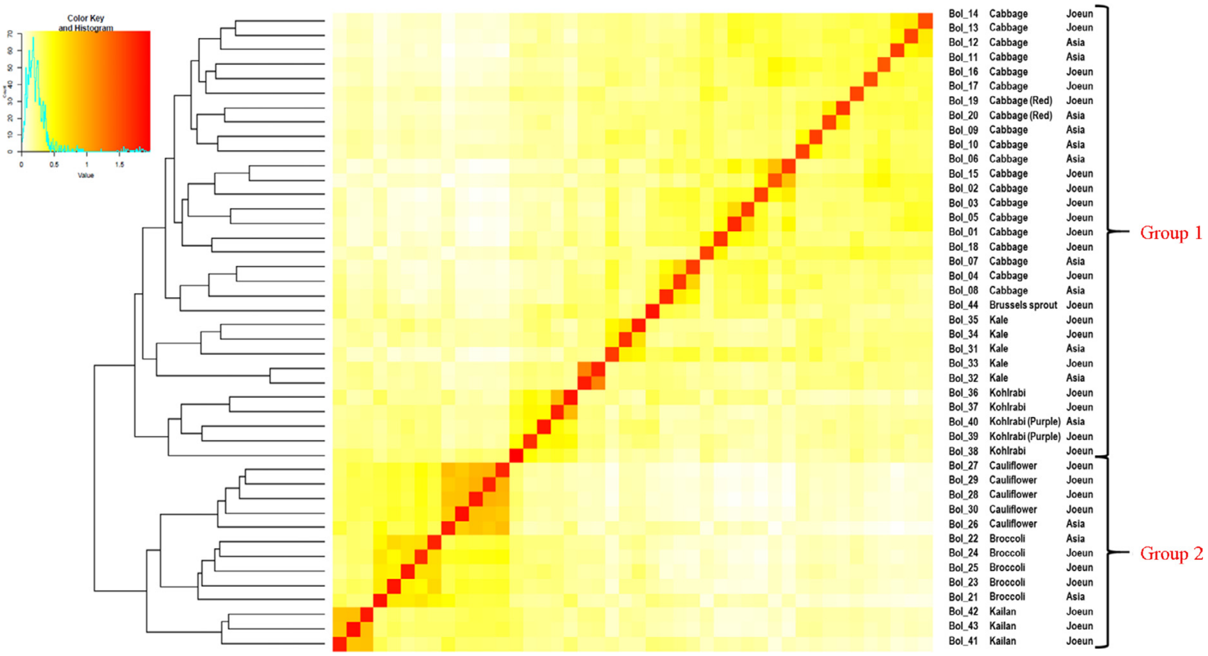

Fig. 2 Whole-genome high-quality SNP-based kinship analysis and heat map among the 44 *Brassica oleracea* accessions
Two major groups were identified based on kinship analysis using 44 *B. oleracea* accessions.

million years (my) and standard deviations of 0.5 my (Yang et al., 2006).

2.5. *Structural variants of Cp and nrDNA sequences and PCR analysis*

Structural variants such as SNP and indel variations were analyzed within the Cp and nrDNA sequences for the 44 accessions. To identify highly reliable variants, variants were manually curated for Cp and nrDNA sequences. Highly informative variants that differentiate multiple subspecies were validated by PCR analysis by designing specific primers. Each PCR reaction contained 10 ng template DNA, 10 pmol · L^{-1} primers, 0.5 μmol · L^{-1} dNTPs, 2 U *Taq* polymerase (TaKaRa, Japan) in a final volume of 20 μL. The PCR cycle consisted of 10 min at 95 °C, followed by 36 cycles of 30 s at 94 °C, 30 s at 55–62 °C and 30 s at 72 °C, with a final extension at 72 °C for 5 min. The amplicons were visualized on a 2% agarose gel or fragment analyzer to estimate the product size.

2.6. *Quantification of major repeats and miniature transposable elements*

Based on a previous study, nine families of repeats were identified as major repeats in the *B. oleracea* genome (Waminal et al., 2016). These repeats were used as a reference to quantify the major repeats in the 44 accessions using a CLC genome assembler (ver. 4.06, Denmark) with parameters of 200 to 500 bp autonomously controlled distance. Genomic abundance in terms of mean read depth (RD) as well as the length of the contig (LC) were performed using a CLC-reference assembly approach. Likewise, 40 miniature inverted-repeat transposon families, including 5, 15 and 20 terminal repeat retrotransposon in miniature (TRIM), short interspersed element (SINE) and miniature inverted-repeat transposable element (MITE) families, were quantified in the 44 accessions using the CLC-reference assembly.

3. Results

3.1. *Discovery of whole-genome diversity in 44 breeding lines*

We mapped each set of paired reads onto the nine pseudochromosomes of the reference genome sequence, and called out sequence polymorphisms, which were categorized as either SNPs or indels. In total, 12 681 369 SNPs and 1 513 964 indels were detected and genotyped to at least one accession. This value was substantially reduced to 496 463 and 37 493, respectively, when all 44 accessions were strictly accounted for (Table S2). Approximately 11% of SNPs and 4% of indels were detected in genic regions (Table S3). Chromosome 3 contained the highest number of detected SNPs and indels, whereas chromosomes 8 and 6 possessed the lowest number of SNPs and indels, respectively.

Comparison of different *B. oleracea* subspecies with the reference assembly revealed that kailan and kohlrabi had the highest and lowest number of SNPs and indels, respectively, although we could not obtain sufficient information for Brussels sprout due to a lack of accession replicates (Table S4). The distribution of SNPs from each subspecies on the reference chromosome assembly revealed subspecies-specific SNP enrichment at certain chromosomal regions (Fig. 1). For example, kailan-specific SNPs were enriched at the 39–45 Mb region of chromosome 3, Broccoli-specific SNPs were concentrated in a 16-Mb region on chromosome 6, whereas kale and cauliflower possessed enrichment in a 16-Mb region on chromosomes 7 and 8. We found that transcrip-

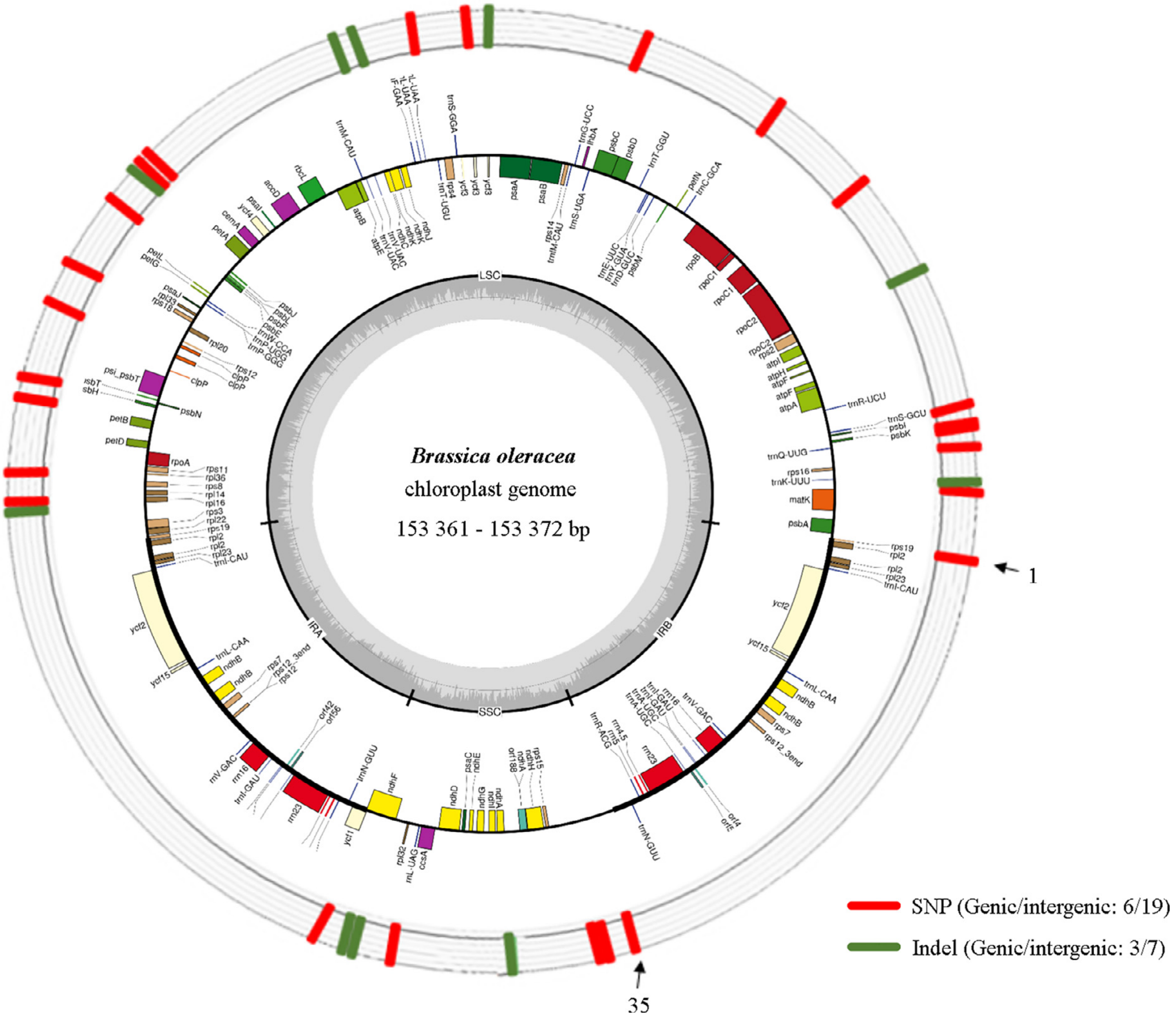

Fig. 3 Circular and variations map of the *Brassica oleracea* chloroplast genome
All 35 variations including SNPs and indels are denoted in the outer circle.

tion factors and transcription regulators were highly abundant in those subspecies-specific SNP enrichment regions (Fig. S1; Table S5).

3.2. *SNP-based kinship analysis among 44* B. oleracea *accessions identifies two distinct groupings*

Kinship analysis based on 496 463 high-quality SNPs across all 44 *B. oleracea* accessions (Table S2) divided the seven subspecies into two subgroups (Fig. 2). Group 1 includes subspecies selected for heading, lateral buds, leaves, and tubers, such as cabbage, Brussels sprout, kale, and kohlrabi. Kale (a leaf vegetable) showed the closest kinship to cabbage (a wrapped-leaf vegetable). Group 2, by contrast, includes subspecies selected for flowers, such as broccoli, cauliflower and kailan, also known as Chinese broccoli. Regardless of SNP datasets used for kinship analysis (i.e. ≥ 1 to ≥ 40 in Table S2), a clear classification of the seven subspecies into two subgroups was observed. Similarly, a clear grouping was obtained using principle component analysis (Fig. S2).

3.3. *The Cp genome, 45S nrDNA assembly and polymorphisms*

We successfully obtained the Cp genome sequences of all 44 *B. oleracea* accessions using approximately 6 × whole nuclear genome coverage by the dnaLCW method. The Cp genomes were conserved across all the 44 breeding lines and sizes varied from 153 361 bp (IM_Bol_34, kale) to 153 372 bp (IM_Bol_42, Kailan) (Table 1, Fig. 3). Multiple sequence alignment of Cp genomes revealed an 11-bp sequence difference between IM_Bol_34 and IM_Bol_42. In total, 35 sequence variations, which include 26 SNPs and 9 indels, were detected among the 44 accessions (Table 2, Fig. 3). Out of these 35 variations, 26 and 9 were localized in intergenic and genic regions, respectively. Three short sequence inversions were observed in trnH-GUG~psbA, psbK~psbI, and petA~psbJ intergenic regions, and a variation in T mononucleotide copy number was observed in the rps15 ~ trnN-GUU region.

Among the 35 polymorphisms, 18 were accession-specific, 2 were specific to dual accessions (Nos. 6 and 35 in Table 2), and 15 showed no specificity. Accessions Bol_07, Bol_10, and

Table 1 Assembly summary for chloroplast and 45S nrDNA sequences of the 44 *B. oleracea* accessions

No.	Morphotype	Chloroplast genome			45S nrDNA		
		Length/bp	Coverage (×)	Accession number	Length/bp	Coverage (×)	Accession number
1	Cabbage	153 366	520	MN396804	5802	2606	MN401690
2	Cabbage	153 366	272	MN396805	5802	1701	MN401691
3	Cabbage	153 366	418	MN396806	5802	428	MN401692
4	Cabbage	153 366	501	MN396807	5802	137	MN401693
5	Cabbage	153 365	709	MN396808	5802	991	MN401694
6	Cabbage	153 365	733	MN396809	5802	1640	MN401695
7	Cabbage	153 364	472	MN396810	5802	1644	MN401696
8	Cabbage	153 366	357	MN396811	5802	976	MN401697
9	Cabbage	153 366	673	MN396812	5802	1966	MN401698
10	Cabbage	153 365	674	MN396813	5802	2602	MN401699
11	Cabbage	153 366	736	MN396814	5802	3257	MN401700
12	Cabbage	153 366	804	MN396815	5802	1122	MN401701
13	Cabbage	153 366	387	MN396816	5802	1162	MN401702
14	Cabbage	153 366	605	MN396817	5802	1239	MN401703
15	Cabbage	153 366	611	MN396818	5802	1832	MN401704
16	Cabbage	153 366	386	MN396819	5802	1154	MN401705
17	Cabbage	153 366	335	MN396820	5802	2199	MN401706
18	Cabbage	153 366	316	MN396821	5802	3074	MN401707
19	Cabbage (Red)	153 364	508	MN396822	5802	2235	MN401708
20	Cabbage (Red)	153 366	712	MN396823	5802	595	MN401709
21	Broccoli	153 364	710	MN396824	5802	797	MN401710
22	Broccoli	153 365	820	MN396825	5802	1467	MN401711
23	Broccoli	153 364	611	MN396826	5802	1871	MN401712
24	Broccoli	153 364	627	MN396827	5802	1167	MN401713
25	Broccoli	153 365	824	MN396828	5802	1169	MN401714
26	Cauliflower	153 364	577	MN396829	5802	1304	MN401715
27	Cauliflower	153 365	541	MN396830	5802	754	MN401716
28	Cauliflower	153 365	1039	MN396831	5802	4468	MN401717
29	Cauliflower	153 365	945	MN396832	5802	3419	MN401718
30	Cauliflower	153 365	910	MN396833	5802	1536	MN401719
31	Kale	153 365	811	MN396834	5802	3295	MN401720
32	Kale	153 367	657	MN396835	5802	1652	MN401721
33	Kale	153 364	698	MN396836	5802	2497	MN401722
34	Kale	153 361	782	MN396837	5802	1673	MN401723
35	Kale	153 364	770	MN396838	5802	1381	MN401724
36	Kohlrabi	153 364	629	MN396839	5802	1250	MN401725
37	Kohlrabi	153 364	800	MN396840	5802	1358	MN401726
38	Kohlrabi	153 364	542	MN396841	5802	2119	MN401727
39	Kohlrabi	153 365	500	MN396842	5802	2467	MN401728
40	Kohlrabi	153 365	615	MN396843	5802	1232	MN401729
41	Kailan	153 364	567	MN396844	5802	1608	MN401730
42	Kailan	153 372	588	MN396845	5802	2713	MN401731
43	Kailan	153 365	479	MN396846	5802	4252	MN401732
44	Brussels sprout	153 366	673	MN396847	5802	1110	MN401733

Bol_44 contained three polymorphisms each, followed by Bol_42, Bol_34, Bol_25, with two, and Bol_9, Bol_24, Bol_26, Bol_30, Bol_32, Bol_36, and Bol_43 each with one. No subspecies-specific polymorphism was identified. To demonstrate the specificity and marker potential of these polymorphisms, we designed primers for five target regions, including four derived cleaved amplified polymorphic sequences (dCAPS) primer sets for SNPs and one primer set for an indel (Table 3), and PCR results confirmed their applicability in discriminating specific accessions (Figs. 3–5).

Following the similar pipeline used for Cp assembly, 45S nrDNA transcriptional unit sequences from the 44 accessions were assembled. The 45S nrDNA transcriptional unit with conserved structures (18S-ITS1-5.8S-ITS2-26S) and a size of 5 802-bp was obtained and characterized. We identified 30 SNPs and a single base-pair indel among the 44 accessions. Overall, polymorphisms were more frequent in the ITS regions than in the coding regions, although 4 and 10 SNPs were identified within 18S and 26S RNAs, respectively (Table S6, Fig. S3).

3.4. *Phylogenetic analysis of Cp genomes reveals divergence time of subspecies*

Phylogenetic analysis using the whole-chloroplast genomes of the 44 accessions (Fig. S4) revealed a similar two-group classification to that obtained using the whole-genome SNP kinship analysis (Fig. 2). Some accessions—four of cabbage (Bol_7, 8, 9, 19), two of kale (Bol_31, Bol_33), three of broccoli (Bol_21, 22, 25), one of cauliflower (Bol_29), two of kohlrabi (Bol_37, Bol_38) and one of kailan (Bol_42)—were grouped into different clades, suggesting that these accessions might result from inter-subspecies hybridization events during the breeding process that resulted in cytoplasm replacement (Fig. S4). In addition, molecular divergence times between *B. oleracea* subspecies were estimated based on the complete Cp genome sequences of the selected accessions representing each subspecies (Fig. 6). Tree topologies with inferred speciation dates clearly identified two major divergence periods of *B. oleracea* subspecies. The trees indicated that group 1 and group 2 diverged about 0.17 million years ago (mya).

Table 2 Sequence variations based on the complete chloroplast genome of 44 *Brassica oleracea* accessions

No.	Type	Position	Locus	Region	Specificity	Alleles	Appearance[b]
1	SNP[a]	264	trnH ~ psbA	Intergenic		tgtt/aaca	28/16
2	SNP	3778	matK ~ trnK	Intergenic		a/t	29/15
3	Indel	4290	trnK ~ rps16	Intergenic		-/a	28/16
4[c]	SNP	6172	rps16 ~ trnQ	Intergenic	Bol_10	c/t	43/1
5[c]	SNP	7252	psbK ~ psbI	Intergenic	Bol_42	t/a	43/1
6	SNP	7350	psbK ~ psbI	Intergenic	Bol_07, Bol_43	ttta/taaa/ttaa/aaaa/ttt-	17/15/10/1/1
7	SNP	8258	trnS ~ trnR	Intergenic	Bol_32	a-/at/ta	41/2/1
8	Indel	15 624	rpoC2	Genic		t/-	34/10
9	SNP	20 982	rpoC1	Genic	Bol_36	c/t	43/1
10	SNP	26 564	rpoB ~ trnC	Intergenic		at/a-/tt	28/14/1
11	SNP	34 359	psbC ~ trnS	Intergenic	Bol_44	a/c	43/1
12	Indel	42 466	trnL	Genic		a/-	41/3
13	SNP	43 676	ycf3	Genic	Bol_10	t/a	43/1
14[c]	SNP	46 463	trnL	Genic	Bol_07	a/c	43/1
15	Indel	49 411	ndhC ~ trnV	Intergenic	Bol_44	-/a	43/1
16	Indel	50 428	trnV ~ trnM	Intergenic	Bol_10	a/-	43/1
17	SNP	61 559	petA ~ psbJ	Intergenic	Bol_34	ta/at	43/1
18	SNP	62 107	petA ~ psbJ	Intergenic	Bol_34	-tga/atgt	43/1
19	SNP	62 120	petA ~ psbJ	Intergenic		ac/tc/a	41/3
20	SNP	64 428	psbE ~ petL	Intergenic		t/a	42/2
21	SNP	68 235	rpl20 ~ rps12	Intergenic		a/t	42/2
22	SNP	70 596	clpP	Genic	Bol_25	t/a	43/1
23	SNP	74 510	petB	Genic	Bol_26	g/a	43/1
24	SNP	75 697	petB ~ petD	Intergenic		g/c	28/16
25	SNP	79 478	rps8 ~ rpl14	Intergenic	Bol_25	a/t	43/1
26	SNP	81 165	rpl16 ~ rps3	Intergenic	Bol_30	g/t	43/1
27	Indel	81 551	rpl16 ~ rps3	Intergenic		t/-	42/2
28	SNP	109 711	ndhF	Intergenic	Bol_44	c/t	43/1
29	Indel	111 544	ndhF ~ rpl32	Intergenic		a/-	39/5
30	Indel	111 782	ndhF ~ rpl32	Intergenic		t/-	42/2
31	SNP	113 669	ccsA	Genic		t/a	39/5
32[c]	Indel	120 026	ndhA	Genic	Bol_42	-/tatatg	43/1
33	SNP	124 383	rps15 ~ trnN	Intergenic	Bol_07	t/a	43/1
34	SNP	124 821	rps15 ~ trnN	Intergenic		$t^{20}/t^{19}/t^{17}$	34/9/1
35	SNP	153 362	rps15 ~ trnN	Intergenic	Bol_9, Bol_24	gc/tc/ga	42/1/1

Note:

[a] Inverse complement.

[b] Number of accessions having a particular allele. For example, in 28/16–28 accessions having “tgtt” type and 16 having “aaca” type.

[c] Polymorphic markers used for marker validation.

Table 3 Primers used for the validation of SNP and indel regions of the chloroplast genome in 44 *Brassica oleracea* accessions

Variant no.[a]	Type	Locus	Primer sequence (5´–3´)	Size/bp	Restriction enzyme
A 4	SNP	rps16 ~ trnQ	F: TTTATAGTAAGATGAAAATCCGTTGACT**GA**T R: AAAAGTGTCGGATGAATGAAAAA	156	*Mbo*I
B 5	SNP	psbK ~ psbI	F: CATTCCTTTAGTTTGTGTAATTGATTC R: ATATTTTTACATATATAGAATTTAATG**CC**G	177	*Hpa*II
C 14	SNP	trnL	F: GCCATCCTGTCCAATGAATTACT R: TTGAGCCAATAAAAACTGAGAAAATTGC**T**T	182	*Mse*I
D 26	SNP	rpl16 ~ rps3	F: CTTATTATGACCATCCCTCATGG R: TGATGAGTTCAGATATTAATGAACTCT**T**TA	194	*Mse*I
E 32	Indel	ndhA	F: CAAATTCTTTGTGTATTTTGGTGTTT R: GGCTTGAAGCGGGTTAAAA	151	

Note:

[a] Variants number corresponds with Table 2.

A phylogenetic tree based on the 45S nrDNA transcriptional unit region produced three groups (Fig. S5), which differed from those obtained with nuclear and Cp genomes (Fig. 2, Fig. S4). Cabbage, broccoli, Brussels sprout, kohlrabi, and kale were formed a clade termed group 1 and kailan and cauliflower were together in a clade as group 2. Group 3 represents two kale accessions, and one of each broccoli and cauliflower. Broccoli accessions did not cluster together but were placed within groups 1 and 3.

3.5. *Quantification of major repeats in 44* B. oleracea *accessions*

A total of nine major repeats were previously identified using the dnaLCW method (Waminal et al., 2016), which could be classified as either tandem repeats (TRs) or transposable elements (TEs). The former group included CentBo1, CentBo2, 5S nrDNA, 45S nrDNA, BoSTR-a, and BoSTR-b, whereas the latter included centromeric retro-transposons of *Brassica* (CRB), and a

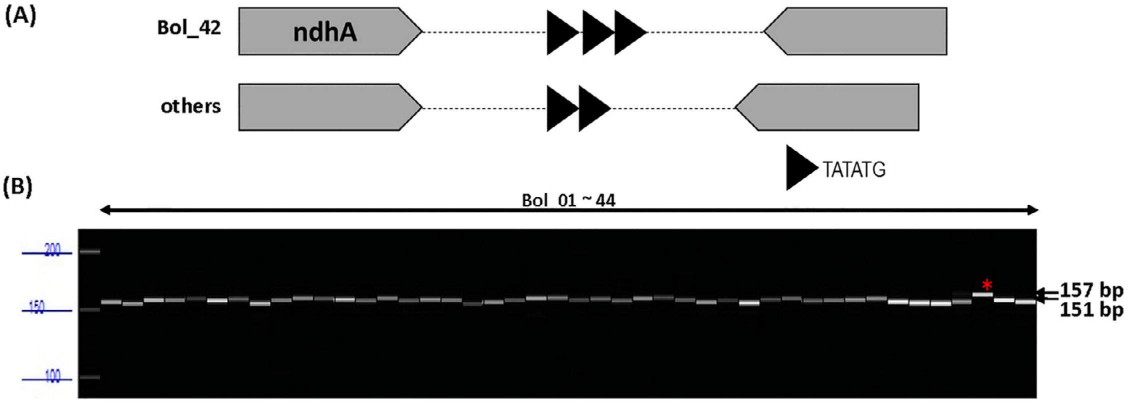

Fig. 4 Molecular validation of intergenic indel variation in *Brassica oleracea*
(A) Systematic diagram showing 6-bp insertions in the *ndhA* gene of the Bol_42 accession, as described in Table 2; (B) Validation of indel polymorphisms based on 44 *B. oleracea* accessions using a fragment analyser. The accession with the 6-bp insertion is denoted by a red asterisk.

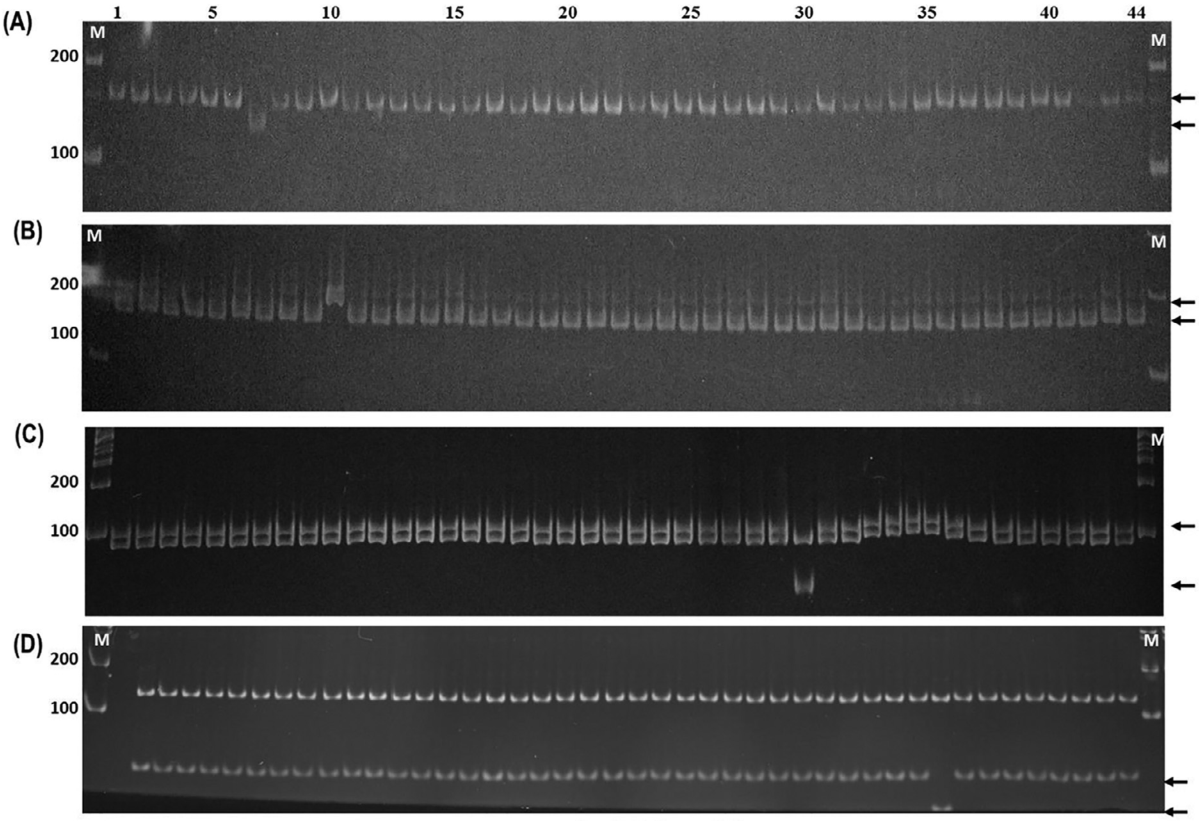

Fig. 5 Molecular validation of SNPs of chloroplast sequences in *Brassica oleracea*
Primer and locus information correspond to A–D in Table S6, and the 44 lanes correspond to the samples listed in Table 1. M: Marker.

copia (BoCop-1) and a CACTA (BoCACTA) element. Considerably more repeats were captured using the dnaLCW method compared with the repeats present in the published genome assembly. CentBo1, followed by CentBo2, was the most abundant repeat element (RE) in terms of copy number and total length (Tables S7 and S8). Although TRs were generally present as more copies than TEs, TEs tended to have considerably higher genome coverage due to their overall longer unit lengths than TRs. The Bo-CACTA element showed the highest abundance among the TEs included in this analysis.

Based on the reference mapping approach, 9.3% of the Brussels sprout genome and 11.6% of the Kohlrabi genome was estimated to consist of these major repeats. The mean abundance of each repeat varied between subspecies (Fig. S6). For example, broccoli and kohlrabi had low amounts of 45S nrDNA and 5S nrDNA, respectively (Tables S7 and S8). Furthermore, the to-

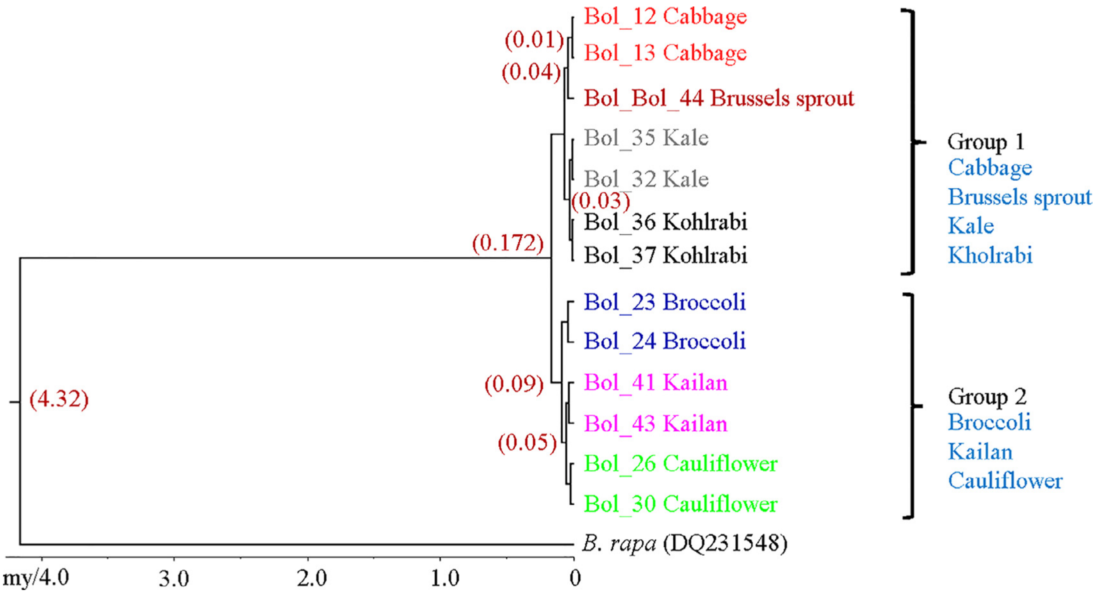

Fig. 6 Molecular divergence time based on complete chloroplast genome sequences of the representative accessions of seven *Brassica oleracea* subspecies

Divergence time of the species are on the nodes, represented as million years. The tree and dating were calculated based on the divergence time of 4.3 mya between *B. rapa* and *B. oleracea*.

tal amount of major repeat elements (REs) also varied between each accession, and some outliers were observed. Many accessions showed approximately 60 Mb of REs. However, accession No. 10 (cabbage) showed a considerable reduction in the amount of REs (lower than 30 Mb) and accession 38 (kohlrabi) showed high amounts of total REs (greater than 100 Mb) (Fig. 7, A).

The low-coverage WGS was also used to quantify the level of miniature TEs (mTEs) such as the 5 TRIMs, 15 SINEs, and 20 MITEs in *B. oleracea*. Similar to TRs and TEs, more mTEs were captured using low-coverage sequencing than those that were included in the reference genome assembly. We also observed that some repeat families were differentially present between subspecies, especially the MITE families BraSto-1, BrasSto-4, BraTo-1 and BraTo-9 (Fig. S7). Divergence in copy number was also observed among the 44 accessions. For instance, the BraSTo-4 MITE family was highly variable among the accessions and within and between the seven subspecies. Compared to SINE and TRIM families, the copy numbers of MITE family members were more variable between the seven subspecies (Fig. S7).

4. Discussion

4.1. *Genetic diversity among 44 B. oleracea breeding lines based on low-coverage WGS*

Given the robust *B. oleracea* reference genome (Liu et al., 2014), low-coverage WGS data can be utilized for GWAS (Li et al., 2011), the assembly of chloroplast and mitochondrial genomes (Kim et al., 2015b; Das et al., 2016), and the genome-wide identification and characterization of repetitive elements (Novák et al., 2013; Kelly et al., 2015; Macas et al., 2015). Here, we utilized the low-coverage WGS sequence of 44 *B. oleracea* accessions that are used as commercial breeding lines to identify potential markers from chloroplast and nrDNA assemblies, estimate the divergence time of each subspecies, and explore the organization, evolution, and dynamics of the *B. oleracea* genome. We have provided an online database and browser where nuclear genome diversity information can be searched for the interested regions. The database can be accessed at http://www.phyzen.com/GWASBrowser/index.html?name=Boleracea_FF.

We classified the seven morphotypes into two different groups based on nuclear SNP, which completely agreed with Cp-genome-based phylogenetic analysis. This two-group classification is well supported by the previous study of *B. oleracea* pan-genome analysis (Golicz et al., 2016). Furthermore, we have seven different morphotypes including kailan, which has not been included in any previous study that we are aware of (Cheng et al., 2016; Golicz et al., 2016). On the other hand, phylogenetic analysis based on 45S nrDNA did not agree with the two-group classification suggesting a high homozygous nature of 45S nrDNA in *B. oleracea* intraspecies level (Perumal et al., 2017; Kim et al., 2018).

High-quality SNP information is useful for molecular breeding programs. It has been suggested that whole genome triplication followed by sub-genome diversity are key features for the morphological diversification of *B. oleracea* morphotypes (Cheng et al., 2014, 2016). Based on our analyses, we also observed diversity specific to each sub-species/morphotype. High diversity in kailan among the other sub-species suggest that kailan might have been untapped and unique variations among other morphotypes. In addition, SNP diversity enriched to specific chromosomal region (in chromosome 3 for kailan) suggests potential hotspot regions for sub-species divergence. Functional annotation of genes revealed potential hotspot regions that are enriched with metabolism and transcription factor-related genes, supporting previous studies (Cheng et al., 2016).

Heading-type *B. oleracea* (kailan, broccoli, cauliflower) have clustered into a single clade by both nucleotide- and Cp-genome-

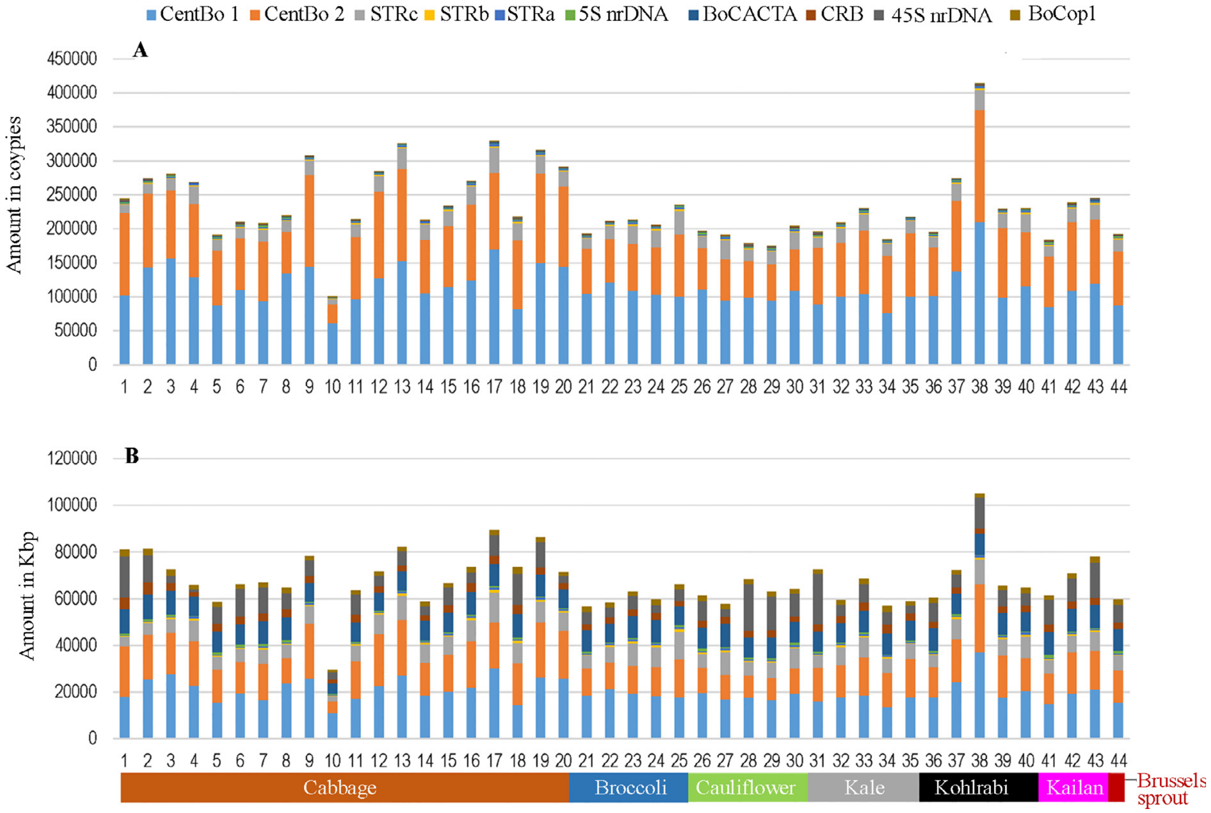

Fig. 7 Distribution of major repeat elements (A) and mTEs (B) among 44 *B. oleracea* accessions
Amounts in Kbp are shown for 13 repeats in each accession. The mTE families were selected based on Sampath et al. (2015).

based phylogeny suggesting a parallel divergence and evolution with other morphotypes which was also supported by the earlier and similar divergence period from other sub-species. There have been few candidate regions with potential impact to plant transcription factors such as protein kinase, but in-depth analysis is still needed to confirm the genes causing *B. oleracea* morphological diversity. Overall, this analysis provides insights into the diversity of different *B. oleracea* morphotypes based on low-depth WGS. Furthermore, data developed in this study will be useful in genotyping *B. oleracea* accessions and especially for GWAS for agronomically significant traits such as yield, resistance to pests and diseases, and for the introgression of target traits from wild *B. oleracea*. The SNP-based genotyping data facilitate the generation of a higher quality and density genetic map for *B. oleracea*, as well as contributing to QTL identification.

4.2. *Chloroplast genome diversity reveals inter-subspecies hybridization during the breeding process of the 44 accessions*

Chloroplast genome sequences are valuable tools for identifying genetic diversity and clarifying genome evolution. The chloroplast genomes of the 44 *B. oleracea* accessions were assembled and provided high-quality information that enabled the inter-subspecific relationships among the 44 *B. oleracea* accessions to be inferred. Several studies have reported that some intra-species Cp genome diversity identifies the lineage of different accessions in *Brassica* species and can be used as a classification tool (Nikiforova et al., 2013; Kim et al., 2015a, 2017; Joh et al., 2017). We also identified some cultivars that were bred via artificial inter-subspecies hybridization between *indica*- and *japonica*-type rice breeding lines, which induced mis-positioning, by following the maternally inherited chloroplast genome, although most nuclear genome markers derived from the paternal parents in *Oryza sativa* (Kim et al., 2015b). We identified 35 polymorphisms among 44 chloroplast genomes. Phylogenomic analysis based on these chloroplast genomes enabled subspecies to be clearly distinguished, with some exceptions. The occasional mis-positioning observed might reflect the maternal genome origin of chloroplast genomes in several breeding lines that were developed via complex breeding strategies involved in producing these commercial lines, including inter-subspecies hybridization.

Recently, we reported the chloroplast genome diversity and evolutionary relationship among 28 accessions belonging to six *Brassica* species (Perumal et al., 2017; Kim et al., 2018). Here, we also identified 35 intra-species polymorphic sites in chloroplast genomes, and evaluated five polymorphic markers among *B. oleracea* breeding lines: four SNPs and a single indel. These five markers were uniquely present in one accession out of 44 and can be used to identify specific breeding lines. A further 30 polymorphic sites can be used to develop DNA markers to classify cytoplasmic genotypes present in the chloroplast and mitochondrial genomes of *B. oleracea* breeding lines and resources.

4.3. *The distribution and evolution of repeats in B. oleracea*

REs comprise a considerable proportion of plant genomes (Michael and van Buren, 2015). In *B. oleracea*, approximately 39% of the genome consists of the REs (Liu et al., 2014). Sequencing data from a single accession can allow the genomic proportions of repeat families to be estimated, but data from multiple accessions, as generated in this study, can allow variations in repeat abundance within a species to be analyzed in more detail. The correlation between repeat abundance and appearance, and desirable traits can be further analyzed, considering the role of repetitive elements in environmental adaptation and intra-specific variation (Kalendar et al., 2000).

The characterization of highly abundant repeats can facilitate the design of FISH probes to understand their genome distribution, which can potentially reveal important information about a genome's history, as was the case for the *PgDel2* LTR element and Pg167TR satellite repeat in *Panax ginseng* (Choi et al., 2014; Waminal et al., 2017). The cytogenetic mapping of REs can ease the identification of homologous chromosome pairs for karyotyping. Moreover, RE-based marker development can also be applied to breeding, as occurred for miniature TEs from *B. rapa* and *B. oleracea* (Perumal et al., 2016).

We presented several applications (Fig. S8) and relevant information concerning genome-wide SNPs for the whole-genome sequences of 44 *B. oleracea* accessions. Then we characterized the genome diversity in chloroplast genomes and nrDNA, which are the most useful barcoding targets, and analyzed the variation in copy number for various major repeats and miniature transposable elements, such as MITEs, TRIMs, and SINEs. In addition, the same data can be used to explore the mitochondrial genome diversity among *B. oleracea* accessions (Park et al., 2013; Yang et al., 2016). Depending on the research objectives, the downstream analysis of low-coverage NGS data can be applied in diverse ways. Most studies focus on GWAS; however, it is clear that WGS data, even with low coverage, represents a valuable tool for breeding.

Acknowledgments

This research was carried out with the support by Golden Seed Project (Center for Horticultural Seed Development, Grant No. 213003-04-1-SB430), Ministry of Agriculture, Food and Rural Affairs (MAFRA), Ministry of Oceans and Fisheries (MOF), and "Cooperative Research Program for Agriculture Science & Technology Development (Grant No. PJ013238)" Rural Development Administration, Republic of Korea.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.hpj.2021.02.004.

REFERENCES

Allen, G., Flores-Vergara, M., Krasynanski, S., Kumar, S., Thompson, W., 2006. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc, 1: 2320–2325.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol, 10, e1003537.

Chalhoub, B., Denoeud, F., Liu, S.Y., Parkin, S.A.P., Tang, H.B., Wang, X.Y., Chiquet, J., Belcram, H., Tong, C.B., Samans, B., Corréa, M., Da Silva, C., Just, J., Falentin, C., Koh, C.S., Le Clainche, I., Bernard, M., Bento, P., Noel, B., Labadie, K., Alberti, A., Charles, M., Arnaud, D., Guo, H., Daviaud, C., Alamery, S., Jabbari, K., Zhao, M.X., Edger, P.P., Chelaifa, H., Tack, D., Lassalle, G., Mestiri, I., Schnel, N., Le Paslier, M.C., Fan Renault, V., Bayer, P.E., Golicz, A.A., Manoli, S., Lee, T.H., Thi, V.H.D., Chalabi, S., Hu, Q., Fan, C.C., Tollenaere, R., Lu, Y.H., Battail, C., Shen, J.X., Sidebottom, C.H.D., Wang, X.F., Canaguier, A., Chauveau, A., Bérard, A., Deniot, G., Guan, M., Liu, Z.S., Sun, F.M., Lim, Y.P., Lyons, E., Town, C.D., Bancroft, I., Wang, X.W., Meng, J.L., Ma, J.X., Pires, J.C., King, G.J., Brunel, D., Delourme, R., Renard, M., Aury, J.M., Adams, K.L., Batley, J., Snowdon, R.J., Tost, J., Edwards, D., Zhou, Y.M., Hua, W., Sharpe, A.G., Paterson, A.H., Guan, C.Y., Wincke, P., 2014. Early allopolyploid evolution in the post-neolithic *Brassica napus* oilseed genome. Science, 345: 950–953.

Cheng, F., Lysak, M.A., Mandáková, T., Wang, X., 2015a. The common ancestral genome of the Brassica Species, in: Wang, X., Kole, C. (Eds.), The *Brassica Rapa* Genome. Springer, Berlin, Heidlberg: 97–105 pp..

Cheng, F., Sun, R., Hou, X., Zheng, H., Zhang, F., Zhang, Y., Liu, B., Liang, J., Zhuang, M., Liu, Y., 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., Liang, J., Wang, X., 2015b. Genome triplication drove the diversification of *Brassica* plants, in: Wang, X., Kole, C. (Eds.), The *Brassica Rapa* Genome. Springer, Berlin, Heidlberg: 115–120.

Cheng, F., Wu, J., Wang, X., 2014. Genome triplication drove the diversification of *Brassica* plants. Hortic Res, 1: 14024.

Choi, H.I., Waminal, N.E., Park, H.M., Kim, N.H., Choi, B.S., Park, M., Choi, D., Lim, Y.P., Kwon, S.J., Park, B.S., Kim, H.H., Yang, T.J., 2014. Major repeat components covering one-third of the ginseng (*Panax ginseng* C.A. Meyer) genome and evidence for allotetraploidy. Plant J, 77: 906–916.

Das, S.P., Bit, A., Patnaik, S., Sahoo, L., Meher, P.K., Jayasankar, P., Saha, T.M., Patel, A.B., Patel, N., Koringa, P., Joshi, C.G., Agarwal, S., Pandey, M., Srivastava, S., Kushwaha, B., Kumar, R., Nagpure, N.S., Iquebal, M.A., Jaiswal, S., Kumar, D., Jena, J.K., Das, P., 2016. Low-depth shotgun sequencing resolves complete mitochondrial genome sequence of *Labeo rohita*. Mitochondrial DNA A DNA MappSeq Anal, 27: 3517–3518.

De Summa, S., Malerba, G., Pinto, R., Mori, A., Mijatovic, V., Tommasi, S., 2017. GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinform, 18: 119.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol, 29: 1969–1973.

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., Martinez, P.A., Chan, C.K.K., Severn-Ellis, A., McCombie, W.R., Parkin, I.A.P., Paterson, A.H., Pires, J.C., Sharpe, A.G., Tang, H., Teakle, G.R., Town, C.D., Batley, J., Edwards, D., 2016. The pangenome of an agronomically important crop plant Brassica oleracea. Nat Commun, 7: 13390.

Joh, H.J., Kim, N.H., Jayakodi, M., Jang, W., Park, J.Y., Kim, Y.C., In, J.G., Yang, T.J., 2017. Authentication of golden-berry P. ginseng cultivar 'Gumpoong' from a landrace 'Hwangsook' based on pooling method using chloroplast-derived markers. Plant Breed Biotechnol, 5: 16–24.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., Schulman, A.H., 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. In: Proceedings of the National Academy of Sciences of the United States of America, 97: 6603–6607.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform, 9: 286–298.

Kelly, L.J., Renny-Byfield, S., Pellicer, J., Macas, J., Novak, P., Neumann, P., Lysak, M.A., Day, P.D., Berger, M., Fay, M.F., Nichols, R.A., Leitch, A.R., Leitch, I.J., 2015. Analysis of the giant genomes of *Fritil-*

laria (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytologist, 208: 596–607.

Kim, C.K., Seol, Y.J., Perumal, S., Lee, J., Waminal, N.E., Jayakodi, M., Lee, S.C., Jin, S., Choi, B.S., Yu, Y.J., 2018. Re-exploration of U's triangle Brassica species based on chloroplast genomes and 45S nrDNA sequences. Sci Rep, 8: 7353.

Kim, K., Lee, S.C., Lee, J., Lee, H.O., Joh, H.J., Kim, N.H., Park, H.S., Yang, T.-J., 2015a. Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. PLoS One, 10, e0117159.

Kim, K., Lee, S.C., Lee, J., Yu, Y., Yang, K., Choi, B.S., Koh, H.J., Waminal, N.E., Choi, H.I., Kim, N.H., Jang, W., Park, H.S., Lee, J., Lee, H.O., Joh, H.J., Lee, H.J., Park, J.Y., Perumal, S., Jayakodi, M., Lee, Y.S., Kim, B., Copetti, D., Kim, S., Kim, S., Lim, K.B., Kim, Y.D., Lee, J., Cho, K.S., Park, B.S., Wing, R.A., Yang, T.J., 2015b. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species. Sci Rep, 5: 15655.

Kim, K., Nguyen, V.B., Dong, J., Wang, Y., Park, J.Y., Lee, S.C., Yang, T.J., 2017. Evolution of the Araliaceae family inferred from complete chloroplast genomes and 45S nrDNAs of 10 Panax-related species. Sci Rep, 7: 4917.

Le Nguyen, K., Grondin, A., Courtois, B., Gantet, P., 2019. Next-generation sequencing accelerates crop gene discovery. Trends Plant Sci, 24: 263–274.

Lee, J., Izzah, N.K., Jayakodi, M., Perumal, S., Joh, H.J., Lee, H.J., Lee, S.C., Park, J.Y., Yang, K.W., Nou, I.S., Seo, J., Yoo, J., Suh, Y., Ahn, K., Lee, J.H., Choi, G.J., Yu, Y., Kim, H., Yang, T.J., 2015. Genome-wide SNP identification and QTL mapping for black rot resistance in cabbage. BMC Plant Biol, 15: 1–11.

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., Abecasis, G.R., 2011. Low–coverage sequencing: implications for design of complex trait association studies. Genome Res, 21: 940–951.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., Zhang, Z., 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics, 28: 2397–2399.

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I.A.P., Zhao, M., Ma, J., Yu, J., Huang, S., Wang, X., Wang, J., Lu, K., Fang, Z., Bancroft, I., Yang, T.J., Hu, Q., Wang, X.F., Yue, Z., Li, H.J., Yang, L.F., Wu, J., Zhou, Q., Wang, W.X., King, G.J., Pires, J.C., Lu, C.X., Wu, Z.Y., Sampath, P., Wang, Z., Guo, H., Pan, S.K., Yang, L.M., Min, J.M., Zhang, D., Jin, D.C., Li, W.S., Belcram, H., Tu, J.X., Guan, M., Qi, C.K., Du, D.Z., Li, J.N., Jiang, L.C., Batley, J., Sharpe, A.G., Park, B.S., Ruperao, P., Cheng, F., Wangminal., N.E., Huang, Y., Dong, C.H., Wang, L., Li, J.P., Hu, Z.Y., Zhuang, M., Huang, Y., Huang, J.Y., Shi, J.Q., Mei, D.S., Liu, J., Lee, T.H., Wang, J.P., Jin, H.Z., Li, Z.Y., Li, X., Zhang, J.F., Xiao, L., Zhou, Y.M., Liu, Z.S., Liu, X.Q., Qin, R., Tang, X., Liu, W.B., Wang, Y.P., Zhang, Y.Y., Lee, J., Kim, H.H., Denoeud, F., Xu, X., Liang, X.M., Hua, W., Wang, X.W., Wang, J., Chalhoub, B., Paterson, A.H., 2014. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun, 5: 1–11.

Lysak, M.A., Koch, M.A., Pecinka, A., Schubert, I., 2015. Chromosome triplication found across the tribe Brassiceae. Genome Res, 15: 516–525.

Macas, J., Neumann, P., Navratilova, A., 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. BMC Genom, 8: 427.

Macas, J., Novak, P., Pellicer, J., Cizkova, J., Koblizkova, A., Neumann, P., Fukova, I., Dolezel, J., Kelly, L.J., Leitch, I.J., 2015. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. PLoS One, 10 e0143424.

Marhold, K., Lihová, J., 2006. Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. Plant Syst Evol, 259: 143–174.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20: 1297–1303.

Michael, T.P., van Buren, R., 2015. Progress, challenges and the future of crop genomes. Curr Opin Plant Biol, 24: 71–81.

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

Natali, L., Cossu, R.M., Barghini, E., Giordani, T., Buti, M., Mascagni, F., Morgante, M., Gill, N., Kane, N.C., Rieseberg, L., Cavallini, A., 2013. The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. BMC Genom, 14: 14.

Nikiforova, S.V., Cavalieri, D., Velasco, R., Goremykin, V.J.M., 2013. Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. Mol Biol Evol, 30: 1751–1760.

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J., 2013. RepeatExplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics, 29: 792–793.

Panjabi, P., Jagannath, A., Bisht, N.C., Lakshmi, K.L., Sharma, S., Gupta, V., Pradhan, A.K., Pental, D., 2008. Comparative mapping of *Brassica juncea* and *Arabidopsis thaliana* using Intron Polymorphism (IP) markers: homoeologous relationships, diversification and evolution of the A, B and C Brassica genomes. BMC Genom, 9: 1–19.

Park, J.Y., Lee, Y.P., Lee, J., Choi, B.S., Kim, S., Yang, T.J., 2013. Complete mitochondrial genome sequence and identification of a candidate gene responsible for cytoplasmic male sterility in radish (*Raphanus sativus* L.) containing DCGMS cytoplasm. Theor Appl Genet, 126: 1763–1774.

Parkin, I.A., Koh, C., Tang, H., Robinson, S.J., Kagale, S., Clarke, W.E., Town, C.D., Nixon, J., Krishnakumar, V., Bidwell, S.L., Denoeud, F., Belcram, H., Links, M.G., Just, J., Clarke, C., Bender, T., Huebert, T., Mason, A.S., Pires, J.C., Barker, G., Moore, J., Walley, P.G., Manoli, S., Batley, J., Edwards, D., Nelson, M.N., Wang, X.Y., Paterson, A.H., King, G., Bancroft, I., Chalhoub, B., Sharpe, A.G., 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biol, 15: R77.

Perumal, S., Waminal, N.E., Lee, J., Izzah, N.K., Jin, M., Choi, B.S., Yang, T.J., 2016. Next-generation sequencing based transposon display to detect high-throughput insertion polymorphism markers in Brassica. Plant Breed Biotechnol, 4: 285–296.

Perumal, S., Waminal, N.E., Lee, J., Lee, J., Choi, B.S., Kim, H.H., Grandbastien, M.A., Yang, T.J., 2017. Elucidating the major hidden genomic components of the A, C, and AC genomes and their influence on Brassica evolution. Sci Rep, 7: 17986.

Pop, M., 2019. Genome assembly reborn: recent computational challenges. Brief Bioinform, 10: 354–366.

Rakow, G., 2004. Species Origin and Economic Importance of Brassica, in: Pua, E.C., Douglas, C.J. (Eds.), Springer Berlin, Heidelberg, pp. 3–11.

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R.K., He, Z.J.M, 2017. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol Plant, 10: 1047–1064.

Sampath, P., Lee, J., Cheng, F., Wang, X., Yang, T.J., 2015. Miniature transposable elements (mTEs): impacts and uses in the Brassica genome, In The Brassica rapa Genome. Springer, Berlin, Heidelberg, Tokyo: 65–81.

Scheben, A., Batley, J., Edwards, D.J.P., 2017. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnol J, 15: 149–161.

Seol, Y.J., Kim, K., Kang, S.H., Perumal, S., Lee, J., Kim, C.K., 2015. The complete chloroplast genome of two Brassica species, *Brassica nigra* and *B. oleracea*. Mitochondrial DNA, 28: 167–168.

Seol, Y.J., Lee, T.H., Park, D.S., Kim, C.K., 2016. NABIC: a new access portal to search, visualize, and share agricultural genomics data. Evol Bioinform Online, 12: 51.

Snogerup, S., 1980. The wild forms of the Brassica oleracea group (2n = 18) and their possible relations to the cultiated ones, in: Tsunoda, S., Hinata, K., G6mez-Campo, C. (Eds.), Brassica Crop and Wild allies, Biology and Breeding. Japanese Scientific Society Press, Tokyo: 121–132.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol, 30: 2725–2729.

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A.E., 2016. GAPIT Version 2: an enhanced integrated tool for genomic association and prediction. Plant Genome, 9: 1–9.

Tank, D.C., Eastman, J.M., Pennell, M.W., Soltis, P.S., Soltis, D.E., Hinchliff, C.E., Brown, J.W., Sessa, E.B., Harmon, L.J., 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. New Phytologist, 207: 454–467.

Waminal, N.E., Choi, H.I., Kim, N.H., Jang, W., Lee, J., Park, J.Y., Kim, H.H., Yang, T.J., 2017. A refined Panax ginseng karyotype based on an ultra-high copy 167-bp tandem repeat and ribosomal DNAs. J Ginseng Res, 41: 469–476.

Waminal, N.E., Perumal, S., Lee, J., Kim, H.H., Yang, T.J., 2016. Repeat evolution in *Brassica rapa* (AA), *B. oleracea* (CC), and *B. napus* (AACC) genomes. Plant Breed Biotechnol, 4: 107–122.

Yang, J., Liu, G., Zhao, N., Chen, S., Liu, D., Ma, W., Hu, Z., Zhang, M., 2016. Comparative mitochondrial genome analysis reveals the evolutionary rearrangement mechanism in Brassica. Plant Biol, 18: 527–536.

Yang, T.J., Kim, J.S., Kwon, S.J., Lim, K.B., Choi, B.S., Kim, J.A., Jin, M., Park, J.Y., Lim, M.H., Kim, H.I., 2006. Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of *Brassica rapa*. Plant Cell, 18: 1339–1347.

Zhang, J., Chiodini, R., Badr, A., Zhang, G., 2011. The impact of next–generation sequencing on genomics. J Genet Genom, 38: 95–109.