SCIENTIFIC REPORTS

OPEN

Elucidating the major hidden genomic components of the A, C, and AC genomes and their influence on *Brassica* evolution

Received: 12 July 2017
Accepted: 4 December 2017
Published online: 21 December 2017

Sampath Perumal[1,2], Nomar Espinosa Waminal[2,3], Jonghoon Lee[4], Junki Lee[2], Beom-Soon Choi[5], Hyun Hee Kim[3], Marie-Angèle Grandbastien[6] & Tae-Jin Yang[2,7]

Decoding complete genome sequences is prerequisite for comprehensive genomics studies. However, the currently available reference genome sequences of *Brassica rapa* (A genome), *B. oleracea* (C) and *B. napus* (AC) cover 391, 540, and 850 Mbp and represent 80.6, 85.7, and 75.2% of the estimated genome size, respectively, while remained are hidden or unassembled due to highly repetitive nature of these genome components. Here, we performed the first comprehensive genome-wide analysis using low-coverage whole-genome sequences to explore the hidden genome components based on characterization of major repeat families in the *B. rapa* and *B. oleracea* genomes. Our analysis revealed 10 major repeats (MRs) including a new family comprising about 18.8, 10.8, and 11.5% of the A, C and AC genomes, respectively. Nevertheless, these 10 MRs represented less than 0.7% of each assembled reference genome. Genomic survey and molecular cytogenetic analyses validates our *insilico* analysis and also pointed to diversity, differential distribution, and evolutionary dynamics in the three *Brassica* species. Overall, our work elucidates hidden portions of three *Brassica* genomes, thus providing a resource for understanding the complete genome structures. Furthermore, we observed that asymmetrical accumulation of the major repeats might be a cause of diversification between the A and C genomes.

Members of the Brassicaceae represent one of the largest eudicot families, including about 338 genera and 3740 species, which have been highly diversified by complex whole genome duplication (WGD) and subsequent evolution. The *Brassica* genus includes many plants with agricultural importance as vegetables, oils, fodders, and condiments throughout the world[1]. The genetic relationship between commonly grown diploid and tetraploid *Brassica* species is described in U's triangle model[2]. Of the six *Brassica* species in this triangle, *B. rapa* (AA, $2n = 2x = 20$), *B. nigra* (BB, $2n = 2x = 20$) and *B. oleracea* (CC, $2n = 2x = 20$) are monogenomic diploids, whereas the remaining three, *B. juncea* (AABB, $2n = 4x = 20$) *B. napus* (AACC, $2n = 4x = 20$) and *B. carinata* (BBCC, $2n = 4x = 20$) are allopolyploids that derived from hybridization events between different AA, BB, CC diploid species.

WGD is common in flowering plants[3]. The *Brassica* genus experienced hexaploidization approximately 16 million years ago (MYA) after diverging from the *Arabidopsis* lineage[4,5]. This lineage-specific whole-genome triplication and selection promoted diversification of the *Brassica* genome[6,7]. Consequently, *Brassica* is rich in species, genetic, and morphological diversity, for example, in terms of leafy heads, stem enlargement, flower/inflorescence modification, and or elongated roots[6].

Repetitive elements (REs) are major players in genome reorganization and stabilization during and after WGD events that disrupt nuclear homeostasis[8]. This concept, and the high genome diversity in *Brassica*, provides a good

[1]Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S7N 0X2, Canada. [2]Department of Plant Science, Plant Genomics and Breeding Institute, Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul, 08826, Republic of Korea. [3]Department of Life Science, Plant Biotechnology Institute, Sahmyook University, Seoul, 01795, Republic of Korea. [4]Joeun Seed, Goesan-Gun, Chungcheongbuk-Do, 28051, Republic of Korea. [5]Phyzen Genomics Institute, Seongnam, 13558, Republic of Korea. [6]INRA AgroParisTech, IJPB, UMR 1318, INRA Centre de Versailles, Versailles, Cedex, France. [7]Crop Biotechnology Institute/GreenBio Science and Technology, Seoul National University, Pyeongchang, 232-916, Republic of Korea. Correspondence and requests for materials should be addressed to T.-J.Y. (email: tjyang@snu.ac.kr)

platform with which to study and explore the evolution of polyploid genomes in relation to RE dynamics[9]. REs, which include tandem repeats (TRs) and transposable elements (TEs), constitute a major genomic fraction (up to 85%) and are responsible for genome size increases in most organisms[10]. REs influence genome architecture, diversity and evolution via homologous recombination and chromosome rearrangements such as duplication, deletion, inversion, and translocation[11]. TRs are short elements (150–400 bp), present as an array of repeats up to 1 million copies, and are localized in heterochromatic regions such as the centromeres, peri-centromeres and sub-telomeres[12–14]. Though the size of TRs are similar between taxa, the sequences may diverge – even between closely related species – because of mutation and homogenization/fixation[15]. Housekeeping nuclear ribosomal DNA (nrDNA) sequences are one of the largest tandem array repeats[16]. They are localized in the peri-centromeric regions (5S nrDNA) and nucleolar organizer regions (45S nrDNA) of most plant species, including *Brassica*[17–19]. Growing evidence supports the importance of TRs in genome function and evolution[20–25].

TEs are also abundant and important for genome expansion, adaptation and evolution[26–28]. Based on their transposition mechanisms, TEs are classified into two major classes: I, retrotransposons, and II, DNA transposons[29]. Retrotransposons, especially those belonging to the *Gypsy* and *Copia* families, occupy a major fraction of most plant genomes. In some cases, a major proportion of the genome is made up of only a few retrotransposon families; for example, *Del* family retrotransposons occupy about 30% of the 3.5 Gb *Panax ginseng* genome[26]. Similarly, TRs can also make up a large fraction of a genome; for instance, ~50% of the olive genome was covered by TRs[30]. Although TEs are mostly conserved in structure, significant variations have been observed, even between close species[8,31,32]. In *Brassica*, asymmetric TE amplification may be important in genetic diversity, speciation, morphological differentiation and polyploidy adaptation[6,33].

Reference genomes are important for understanding genome structure and help to speed up functional genomics approaches to crop improvement[34]. Advances in sequencing technologies, such as next-generation sequencing (NGS), have provided insights into the structures and functions of plant genomes at an unprecedented pace[35,36]. However, achieving a pseudo-chromosome level of assembly is arduous, often because of REs. REs can, for example, hinder complete genome assembly and leave hidden gap regions, even in model organisms[37].

The availability of whole-genome pseudo-chromosome assembly for the major *Brassica* species such as *B. rapa* (330 Mb out of 485 Mb), *B. oleracea* (385 Mb/630 Mb) and *B. napus* (645 Mb/1130 Mb), has enabled better understanding of the genome architectures, compositions and evolution of these species[33,38–41]. Currently, 27 Mb (5%), 155 Mb (25%), and 205 Mb (18%) of the A, C and AC genomes, respectively, are unanchored scaffolds. The available reference genome assemblies cover 80.5%, 85.7%, 75.2% of the A, C and AC genomes, respectively, leaving 19% (94 Mb), 14% (90 Mb), and 25% (280 Mb) of the genomes unassembled, mostly because of RE assembly problems[10]. Several studies have characterized and localized REs in *Brassica* genomes, including nrDNAs[16], centromeric tandem repeats (CentB)[42], sub-telomeric tandem repeats (STR)[17], centromeric and peri-centromeric long terminal repeat (LTR) retrotransposons[19], terminal-repeat retrotransposons in miniature (TRIMs), and miniature inverted-repeat transposable elements (MITEs)[43–49].

Here, we explored the major repetitive elements of *B. rapa* (A genome), *B. oleracea* (C genome) and *B. napus* (AC genome) collectively using low-coverage, whole genome sequence (WGS) reads, termed the dnaLCW-RE approach. We characterized 10 major repeats including a new repeat – and inspected their genomic abundance, diversity, and distribution. This study provides insights into the interspecific and intraspecific diversity and evolution of the major *Brassica* repeats that form the previously hidden components of the *Brassica* genome.

Results

De novo assembly and mapping of low-coverage, WGS identifies high copy repeats in *B. rapa* and *B. oleracea*.

We previously demonstrated that *de novo* assembly using low-coverage, whole-genome sequences (the dnaLCW approach) can be used for complete and simultaneous assembly of high-copy genomes such as the chloroplast and nuclear ribosomal DNA[50]. Here, we used a similar approach, which we named dnaLCW-RE, to identify the sequences of the major high copy REs from *Brassica* plants of the A and C genomes.

Firstly, *B. rapa* (Br-1-1) low-coverage (2x), WGS Illumina pair-end reads were used for *de novo* assembly; 147 118 contigs were obtained with an average depth of 13x. Contigs were ordered based on read depth, and initially, the top 50 high-depth contigs were selected for further repeat analysis. Average depth of the top 50 contigs was 2 588x (1292–8064 copies on the haploid genome equivalent) with lengths ranging between 226 and 7453 bp (Supplementary Table 1). Among these 50 contigs, 47 showed similarity to known repeat families, with 33 CentB homologs, six nrDNAs, three STRs, and five transposons. The remaining three contigs were of unknown origin and were too small for further analysis. A total of eight repeat families were characterized, including centromeric tandem repeats of *B. rapa* (CentBr1 and CentBr2), sub-telomeric tandem repeats (STRa and STRb), nuclear ribosomal DNA units (5S nrDNA and 45S nrDNA), centromere-specific *Brassica* retrotransposons (CRB), and peri-centromeric *B. rapa* retrotransposons (PCRBr1a) (Supplementary Table 2).

De novo assembly of *B. oleracea* Bol-1-1 WGS (2x haploid genome equivalents) generated 260 198 contigs. The top 50 high-depth contig lengths ranged between 200 and 2103 bp, and read depth ranged between 139 and 13 366 copies. The average contig length of *B. oleracea* was much larger than that of *B. rapa*, and the contigs were annotated based on a sequence similarity searches against the Repbase and National Center for Biotechnology Information (NCBI) databases (Supplementary Table 3). Twelve contigs represented slightly different forms of *B. oleracea* centromeric tandem repeats (CentBo), 9 sub-telomeric repeats (BoSTR), 8 nrDNAs, and 14 known TEs. The remaining 7 contigs were unknown repeats. Deep investigation and grouping of these major contigs based on sequence similarity led to the identification of 10 major repeat families, including nine well-known repeats and a new *B. oleracea*-specific *Copia* retrotransposon (BoCop-1) (Supplementary Table 4).

In addition, we also applied RepeatExplorer method to characterize major repeats using the same WGS reads. RepeatExplorer based analysis revealed 90, 107, and 284 clusters with the genome occupancy of 46%, 39% and

		B. rapa						***B. oleracea***						***B. napus***						
		Reference genome (391 Mb)			**1x wgs (485 Mb)**			**Reference genome (540 Mb)**			**1x wgs (630 Mb)**			**Reference genome (850 Mb)**			**1x wgs (1130 Mb)**			
Element ID	**Element size (bp)**	**GR-R (n)**	**GR-R (Kb)**	**GP-R (%)**	**GR-W (n)**	**GR-W (Kb)**	**GP-W (%)**	**GR-R (n)**	**GR-R (Kb)**	**GP-R (%)**	**GR-W (n)**	**GR-W (Kb)**	**GP-W (%)**	**GR-R (n)**	**GR-R (Kb)**	**GP-R (%)**	**GR-W (n)**	**GR-W (Kb)**	**GP-W (%)**	**Homologous element**
CentB1	176	1,632	283	0.07	179,807	31,646	6.5	1,203	196	0.03	114,077	20,192	3.2	336	56	0	228,031	40,362	2.3	Lim *et al*. 2005
CentB2	176	1,182	204	0.05	36,765	6,470	1.3	1,924	317	0.05	89,827	15,899	2.5	518	85	0.001	51,093	9,043	0.5	Lim *et al*. 2005
5 S nrDNA	501	75	37	0.01	5,096	2,553	0.5	143	70	0.01	1,286	647	0.1	45	22	0	5,147	2,579	0.3	Waminal *et al*. 2015
45 S nrDNA	7,456	5	32	0.01	4,008	29,883	6.2	1	5	0	1,072	8,136	1.3	—	—	0	4,089	30,485	5	Waminal *et al*. 2015
STR-Br[a]	352	2,155	735	0.19	13,296	4,685	1	1,511	477	0.08	3,829	1,354	0.2	1,517	509	0.005	20,349	7,122	0.1	Koo *et al*. 2011
STR-Bo[b]	352	1,376	466	0.12	738	260	0.1	5,186	1,735	0.27	21,067	7,394	1.2	4,632	1,569	0.014	23,142	8,123	0.1	Koo *et al*. 2011
CRB	5,908	2	11.4	0.00	633	3,738	0.8	2	12	0	486	2,995	0.5	—	—	0	1,168	6,902	0.9	Lim *et al*. 2007
PCRBr1a	8,395	1	7.8	0.00	1,268	10,441	2.1	1	8	0	86	711	0.1	1	8	0	960.4	8,217	1.2	Lim *et al*. 2007
Cop-1	6,711	1	5.2	0.00	34	229	0.1	15	88	0.01	298	1,988	0.3	1	6	0	284.4	1,910	0.2	This study
CACTA	7,675	1	7.6	0.00	143	1101	0.2	1	9	0	956	8,987	1.4	1	8	0	1,266	9,713	0.9	Alix et al. 2008
Total		6,430	1,788.3	0.63	241,788	91,006.8	18.76	9,987	2,916.40	0.463	232,984	68,303	10.84	7,051	2,262.2	0.02	335,529	124,455	11.52	

Table 1. Composition of major repeats based on the reference genome assembly and 1x WGS of three *Brassica species*. GR-R (n): Number of repeat units represented in reference genome sequences; GR-R (kb): Total length of repeat units represented in reference genome sequences (kbp); GR-W (n): Number of repeat units represented in WGS; GR-W (kb): Total length of repeat units represented in WGS (kbp); GP-R: Proportion of the genome in reference genome sequence; GP-W: Proportion of the genome in WGS; Kb: amounts in kbp. [a] Based on two STRs, (STRa and STRb) of *B. rapa* (Br); [b]Based on two STRs, (STRa and STRb) of *B. oleracea* (Bo).

51% in A, C and AC genome, respectively (Supplementary Tables 5,6,7 and 8). Comparative analysis revealed that all of these RepeatExplorer clusters belonged to the 10 MRs identified through dnaLCW-RE. Moreover, those clusters containing 10 MRs occupied 21.8%, 14.6% and 12.4% of the A, C and AC genome, respectively, which is similar but slightly higher than dnaLCW-RE analysis (Supplementary Table 9). In addition, this analysis also provides information for repeats other than 10 MRs contributing significant fraction of the three *Brassica* genome.

Characterization of a new LTR-retrotransposon family in the *B. oleracea* genome. Nine of ten REs identified using the dnaLCW-RE approach were similar to those previously reported in the *Brassica* genome (Table 1). However, this approach also revealed a new, highly abundant, *B. oleracea*-specific LTR retrotransposon. Based on in-depth analysis of unclassified contigs from C (Fig. 1), this was characterized as a Ty1/*Copia* type (BoCop-1). Among the top 50 *B. oleracea* contigs, two unclassified contigs (numbers 12 and 29) were expected to represent 1423x per genome (Supplementary Table 3). A Repbase search revealed that both contigs were similar to *Copia*-type LTR retrotransposons. Contig lengths were extended by manual read walking to obtain the complete LTR retrotransposon structure. Annotation of the extended contig revealed signature structures of LTR retrotransposons, such as target site duplication, terminal repeats, and a functional coding domain. Likewise, unclassified *B. oleracea* contigs led to the identification of *Copia*-type LTR-retrotransposon in *B. oleracea* (Fig. 1A). Read mapping revealed 383 and 25 copies in the A and C genomes, respectively (Fig. 1B). Karyotype analysis using the tetraploid derivative AC genome showed C genome-specific amplification (Fig. 1C).

Estimation of copy numbers and genome proportions for the 10 major repeats in A, C and AC genomes. The relative genomic abundance of the 10 major repeat families was quantified in the reference genome, and also in the WGSs of 64 accessions belonging to the A, C and AC genomes (Supplementary Tables 1,2,4 and 10). Comparative analysis revealed that about 19%, 11% and 12% of the A, C and AC genomes, respectively, was occupied by these 10 REs, while <0.7% was found in the reference genomes (Fig. 2).

In the A genome, 18.8% of the genomes derived from 11 accessions, including six different subspecies, was occupied by these 10 major repeat families, while these repeats were present in less than 1% of the reference assembly (Supplementary Table 2). Total repeat length of each family varied, ranging between 0.2 Mb for BoCop-1 and 38 Mb for the two CentBr types. Of these, CentBr and 45 S nrDNA were estimated to occupy a larger fraction of the haploid genome; 7.9% and 6.2%, respectively. CentBr1 and CentBr2 were predominant components of the genome, representing 179 807 and 36 765 copies per haploid genome, respectively. Most *in silico* analyses showing the differential abundance of various repeat families were supported by analysis using fluorescence *in situ* hybridization (FISH) (Fig. 3). Different repeat families showed different chromosomal distribution patterns enabling easy identification of homologous pairs.

Analysis of 44 *B. oleracea* accessions – including eight different morphotypes – to interpret the repeat contribution of the genome, revealed that 10.8% of the C genome was made up of the 10 major repeat elements. However, the *B. oleracea* reference genome sequence contained only 2.9 Mbp of the 10 major repeats (Supplementary Table 4). Our read mapping-based calculation revealed that the major repeat proportion of the genome was 68.3 Mbp, ranging between 0.6 Mb and 20 Mb for each RE. Like *B. rapa*, CentBo1 and CentBo2 were present in high copy numbers, and occupied large proportions of the genome: 114 077 (3.2%) and 89 827 (2.5%), respectively. *De novo* analyses were supported by FISH using repeat-specific probes (Fig. 4). Furthermore, divergence time

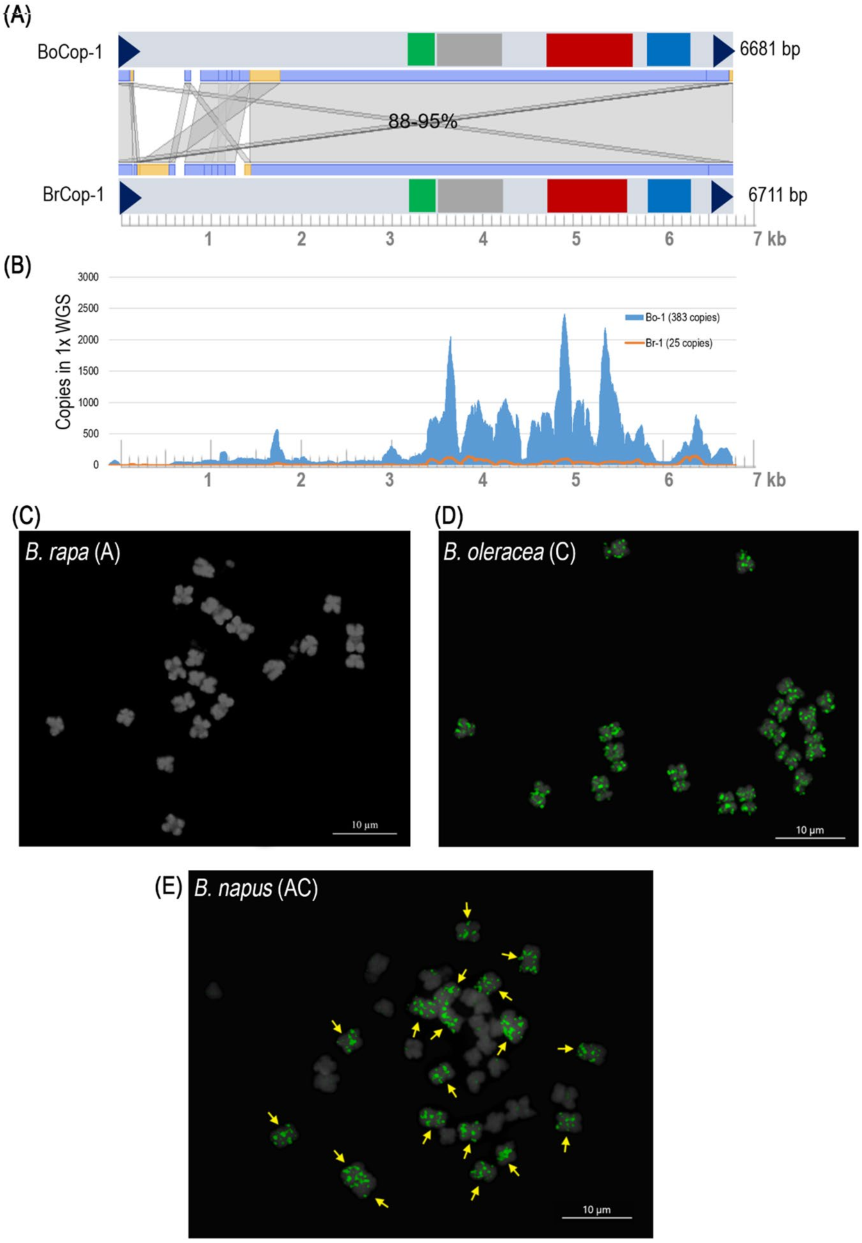

Figure 1. Structure and syntenic analysis of the new long terminal repeat (LTR) retrotranposon. (**A**) Micro-syntenic comparison of *B. oleracea*-specific LTR-RT (BoCop-1) with its ortholog (BrCop-1) from *B. rapa*. Arrows denote the LTR; Green, red and blue box denotes the functional domain, GAG, reverse transcriptase and integrase, respectively. Value for similarity with homologous regions is represented as a percentage. **(B)** Read mapping and copy number estimation in *B. oleracea* and *B. napus*. FISH analysis of BoCop1 in *B. rapa* **(C)** *B. oleracea* **(D)** and *B. napus* **(E)** genomes. Arrows indicate C sub-genomes in the *B. napus*-genome.

analysis of each high copy TR family, which revealed elements of recent and ancient origin, indicated that *B. rapa* had more recent amplification than *B. oleracea* in the span of up to 14 mya (Supplementary Figure 1).

Estimating the 10 major repeats for the allotetraploid *B. napus* showed that they made up a significant fraction (11.5%) of the genome, albeit a much lower percentage (0.02%) represented in the genome assembly (Supplementary Tables 2 and 10). Of the 10 repeats screened in the genome, 4 088 copies of 45 S nrDNA and 228 030 copies of CentBnp1 were found, representing 5% and 2.3% of the genome, respectively. Furthermore, 45 S nrDNA contributed the highest proportion, covering 30 Mb of the haploid genome. Compared with its parental genome, *B. napus* had relative low amounts of major repeats from the ancestral A and C genomes, although the overall composition of repeats was slightly reduced to around 2.6%. FISH analysis based on repeat-specific probes showed the relative abundance of each repeat family for rDNA, CentBnp, STR and CRB which is well agreement with quantification based on dnaLCW-RE (Fig. 5). Moreover FISH has also approved the C sub-genome specific distribution of CACTA in *B. napus* genome.

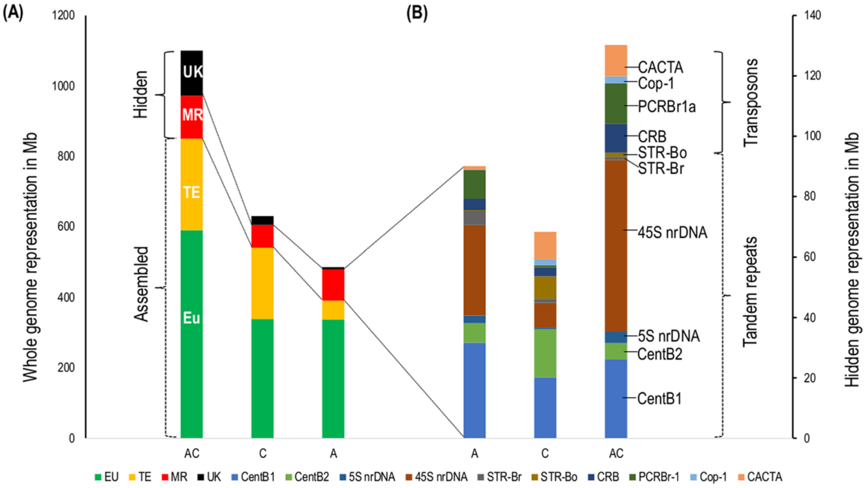

Figure 2. Classification of genome components based on current assembly, and elucidation of hidden genome components in three *Brassica* genomes. (**A**) The assembled reference genome was classified based on genome components such as non-repetitive euchromatic regions (EU) and repetitive TE portions (TE) from genome annotation and the hidden genome. The hidden genome is the proportion of unassembled genome. Here, red indicates the proportion of the genome occupied by 10 major repeats (MR), and gray indicates the unknown genome component (UK). (**B**) Representation of 10 MR families in the hidden genome. MR was subdivided into 10 repeat families: CentB1, CentB2, 5 S nrDNA, 45 S nrDNA, STR-Br, STR-Bo, PCRBr1a, Cop-1, CRB, and CACTA (ordered bottom to top).

Karyotype analysis-based genomic distribution and proportion of REs. Five-color FISH analysis revealed unique patterns of repeat distributions for most of the REs in three *Brassica* genomes (Figs 3,4 and 5). For example, centromere-specific localization was observed for CRB, and CentB1 and CentB2 family repeats. CRB signals were observed in all the chromosomes of the A and C genomes. However, different FISH signal distribution patterns were observed for CentBo/Br1 and CentBo/Br2 in both genomes. CentBr1 was distinctly localized to 8 out of 10 chromosomes of *B. rapa*, and the remaining chromosomes (A02 and A04) were occupied by CentBr2 (Fig. 4). Unlike those observed in *B. rapa*, CentBo1 and CentBo2 were intermingled to different degrees in all *B. oleracea* chromosomes. CRB remained in all centromeres of the *B. napus* A_nC_n chromosomes. In A_n chromosomes, CentBnp1 retained the pattern of CentBr1 distribution seen in A_r chromosomes, but CentBnp2 had a rearranged pattern. Chromosomes 2, 6, 7, and 9 had exclusively CentBnp1 signals, chromosomes 1, 3, and 8 had exclusively CentBnp2 signals, and chromosomes 4 and 5 had colocalized CentBnp signals (Fig. 5).

Unlike the centromeric tandem repeats, STRs preferentially accumulated into the sub-telomeric regions of some chromosomes. *B. rapa* STRs were observed in only a few chromosomes: BrSTRa was in three chromosomes, and BrSTRb was in seven chromosomes. STR-Bo repeats were present in the sub-telomeric regions of most chromosomes, although with different intensities. Patterns from the A_r and C_o genomes were retained in the A_nC_n genome. In addition to chromosome-specific distribution, genome-specific amplification was observed for BoCop-1. Like BoCACTA, BoCop-1 was specific to the C genome and was widely distributed in *B. oleracea* chromosomes (Fig. 1).

Discussion

dnaLCW-RE is a useful tool to characterize the hidden components of *Brassica* and other genomes.

High-throughput NGS technologies enable assembly of the genomes of many important crops at unprecedented pace and accuracy; consequently, this advance makes comparative studies and many downstream analyses possible[34,51–56]. Complete assembly of plant genomes is hampered by the complex genome structure caused by different REs[37]. Genome-wide exploration of the repetitive parts of the genome will us help to understand complete genome structures and compositions[18,57–60].

Here, we performed the first comprehensive genome-wide analysis to identify major repeat families in the *B. rapa*, and *B. oleracea*, genomes using the dnaLCW-RE approach, which makes use of low-coverage (2x coverage) WGS. We found 10 major repeats in both the A and C genomes, including three new repeat families. Among the 10 MRs, six were common to both the *B. rapa* and *B. oleracea* genomes and four elements were specific to one or the other of the genomes, e.g., STR-Br, and PCRBr were abundant in *B rapa* whereas Cop-1 and CACTA were abundant in *B. oleracea*. *In silico* analysis estimated these 10 major repeat families to occupy about 19%, 11% and 12% of the A, C, and AC genomes, respectively, reflecting 48 and 76% of the proportions of the hidden genomes of A and C genomes, respectively. Tandem repeats (TRs), such as CenB, STRs, and 45 S rDNA, occupied greater portions of the *B. rapa* genome than the *B. oleracea* genome. TRs present in highly condensed arrays are difficult to assemble, thus explaining why large fractions of the *B. rapa* reference genome sequence have remained hidden. By contrast, TEs were amplified in the hidden genome of *B. oleracea* compared to that of *B. rapa*, although many more TEs were included in the assembled *B. oleracea* reference genome sequence than that of *B. rapa* (Fig. 2).

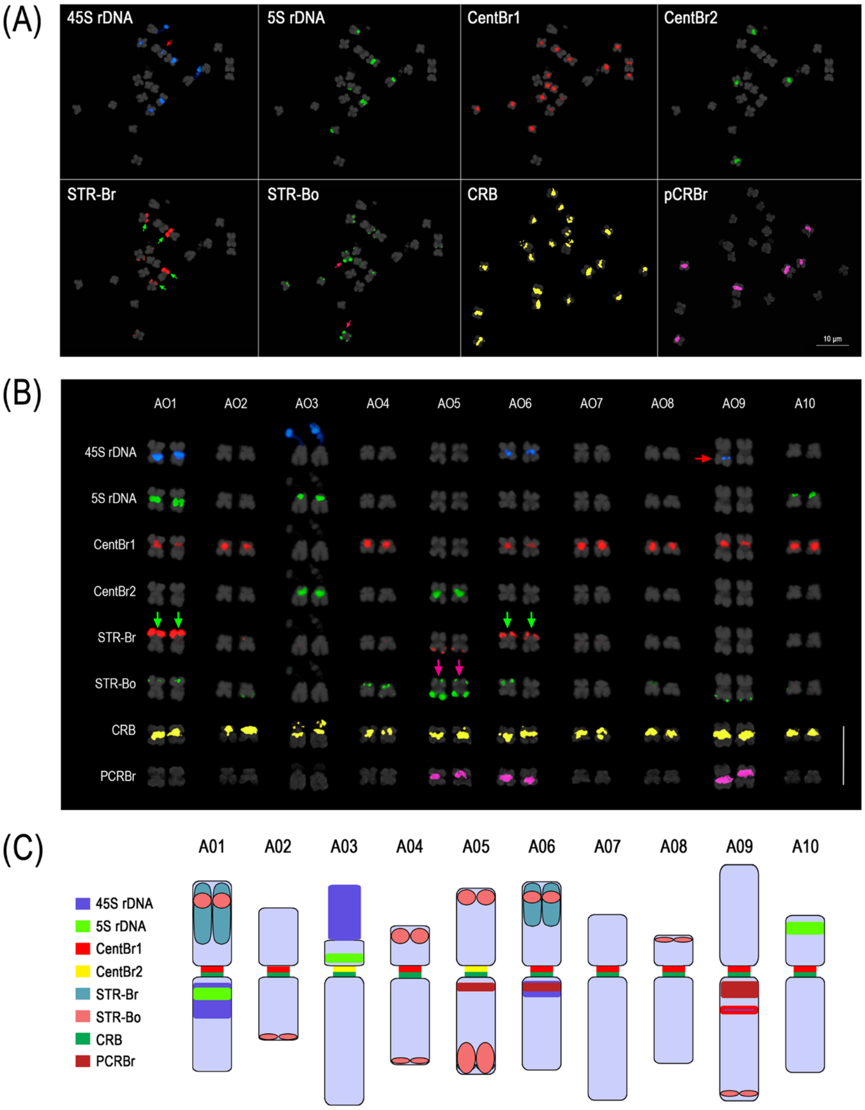

Figure 3. Physical mapping of major repeats in *Brassica rapa* through FISH analysis. **(A)** FISH analysis of major *Brassica* repeats. Red, green and pink arrows indicate a minor hemizygous 45 S rDNA locus, major STR-Br and STR-Bo signals, respectively. **(B)** Karyogram of *B. rapa* based on the distribution of major DNA tandem repeats. Green and pink arrows indicate chromosomes with major STR-Br and STR-Bo signals, respectively, which amounted to three major signals, and red arrow indicates a hemizygous 45 S rDNA. Bar = 10 µm. **(C)** Karyotype idiogram of *B. rapa*. 45 S rDNA with red border indicates hemizygous locus.

RepeatExplorer based repeat characterization provided wealth of support for dnaLCW-RE approach based major repeat characterization. All the 10 MRs were able to find in the RepeatExplorer output, though there was a slight difference in the estimation of genome proportions. Both approaches were significantly good at capturing complete unit of the short tandem repeats especially CentB and STR. However, RepeatExplorer produced more number of repeats in less number of contigs which shows the advantage over dnaLCW-RE approach for characterizing other repeats. Since our analysis is based on a low-coverage and *de novo* assembly approach, it is possible that some major REs have been missed, and more repeats may be identified with increased depth of contig analysis. Combining both tool will provide good opportunity to understand the complex heterochromatin structure in plant genome.

Cytogenetic analysis of these MRs revealed the genomic distribution, diversity and abundance of each repeat family in the three *Brassica* genomes, which supported our *in silico* survey. Notably, we conclude that about 40% of the A, C, and AC genomes are occupied by REs, indicating that other uncharacterized REs with moderate copy numbers (e.g., DNA transposons and retrotransposons) may be found in the hidden genome and are not represented in this survey (Fig. 2). However, these uncharacterized REs may be explored with the dnaLCW-RE method if analysis is extended to unannotated *de novo* assembled contigs or combination with RepeatExplorer approach. The hidden portion of each genome is expected to be larger than what was captured in this survey.

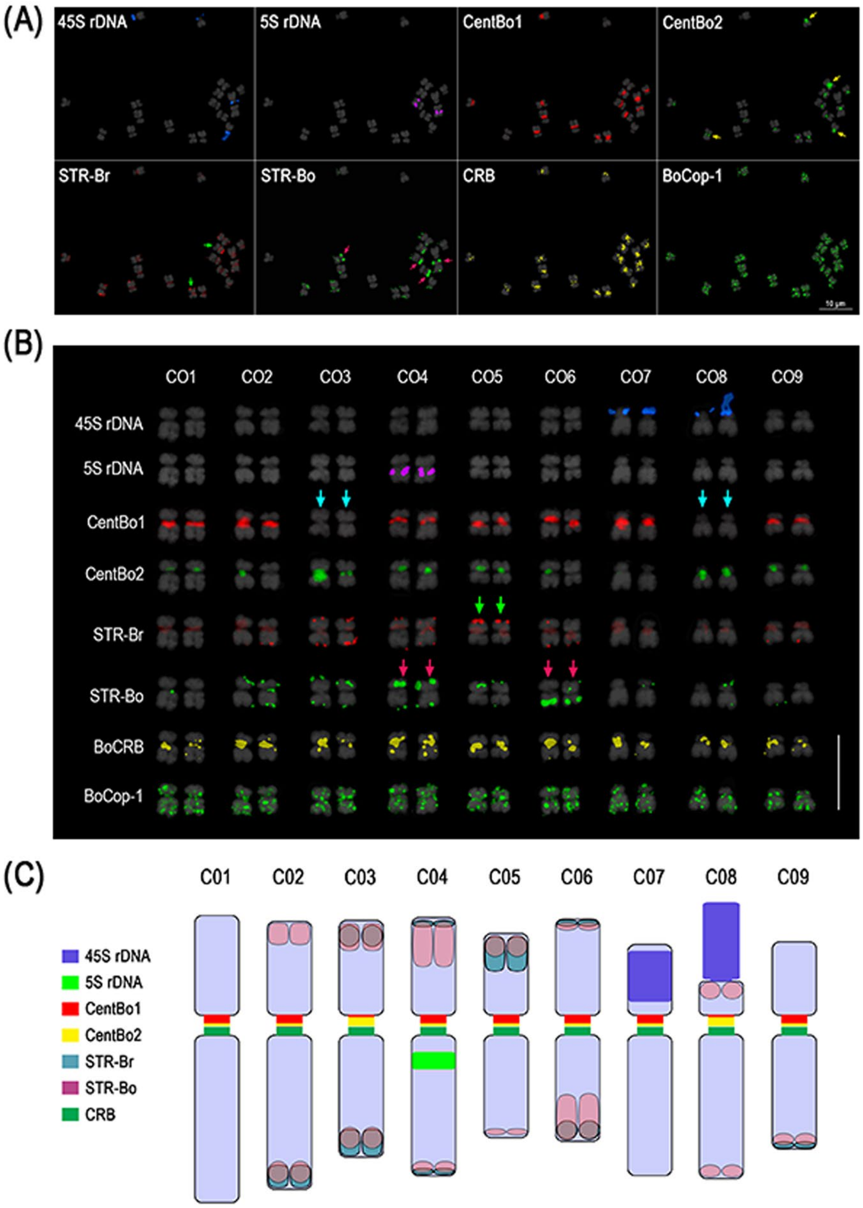

Figure 4. Genomic distribution of major repeats in *B. oleracea* through FISH analysis. **(A)** FISH analysis of major *Brassica* repeats. Green and pink arrows indicate major STR-Br and STR-Bo signals, respectively, amounting to three pairs of BoSTR signals (as observed by Koo *et al*., 2011.) **(B)** Karyogram of *B. oleracea* based on the distribution of major DNA tandem repeats. Blue-green arrows indicate chromosomes with very weak CentBo1 but strong CentBo2 signals. CentBo1 and CentBo2 are colocalized on most other chromosomes. Green and pink arrows indicate major signals of STR-Br and STR-Bo, respectively, which amounted to three major signals. Bar = 10 μm. **(C)** Karyotype idiogram of *B. oleracea*. Centromeric bars represent arbitrary signal hybridization intensities.

Most repeat families are similar to those previously reported in the *Brassica* genome, which was characterized by multiple, independent research groups (Table 1)[19,42]. Nevertheless, our approach was able to identify those MRs in a single study. Furthermore, our approach provided information about new RE in *B. oleracea* (BoCop-1), which will fuel future studies on genome diversification and evolution in the *Brassicaceae*.

BoCop-1 is a C-genome specific LTR-retrotransposon. Over 370 copies of BoCop-1 were predicted in the C genome, but few were found in the A genome (Fig. 1). This indicates that the LTR retrotransposon has undergone C genome-specific amplification over the last 4.6 million years. FISH data showed BoCop-1 signals widely distributed in the *B. oleracea* genome, but absent in the *B. rapa* genome. A similar scenario was seen with the DNA transposon, BoCACTA. C genome-specific BoCACTA was exclusively amplified in the C genome after diversification with the A genome, and is expected to be important in C genome evolution and gene proliferation[61]. Like BoCACTA, BoCop-1 was also widely distributed throughout all C-genome chromosomes. The

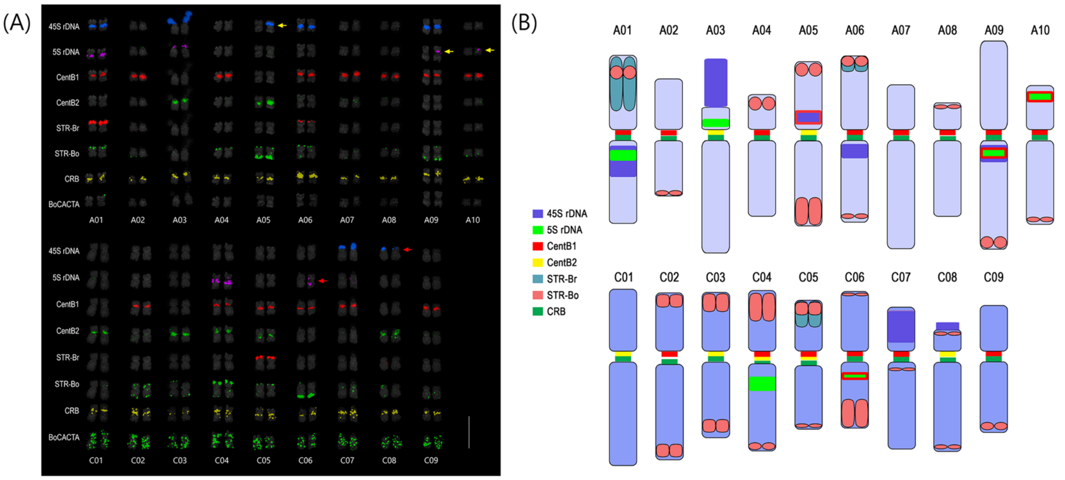
(A)
45S rDNA
5S rDNA
CentB1
CentB2
STR-Br
STR-Bo
CRB
BoCACTA
A01 A02 A03 A04 A05 A06 A07 A08 A09 A10
C01 C02 C03 C04 C05 C06 C07 C08 C09
(B)
A01 A02 A03 A04 A05 A06 A07 A08 A09 A10
C01 C02 C03 C04 C05 C06 C07 C08 C09
45S rDNA
5S rDNA
CentB1
CentB2
STR-Br
STR-Bo
CRB

Figure 5. Genomic distribution of major repeats in *Brassica napus*. (**A**) Karyogram of *B. napus* based on the distribution of major DNA tandem repeats. Yellow and red arrows indicate major chromosomal rearrangements within the A_n and C_n genomes, respectively. Note that the C genome chromosomes, although fewer in number, are generally larger than those of the A genome, reflecting the genomic differences between the two diploid species. BoCRB is seen in all chromosomes, while BoCACTA elements are specific to C. Bar = 10 µm. **(B)** Karyotype idiogram of *B. napus*. rDNA with red border represents hemizygous loci, most likely from homeologous unequal crossover. Darker chromosomes of the C genome indicate the preferential hybridization of the BoCACTA transposon.

abundance and widespread nature of BoCop-1 makes it an excellent cytogenetic marker for identifying the C genome in tetraploids or unknown species. BoCop-1 may also be involved in *B. oleracea* evolution, diversification and speciation, like other TEs in other plants[61,62].

Comparative analysis of major repeats in 64 *Brassica* accessions reveals repeat dynamics in the interspecies and intraspceies *Brassica* genomes. Comparative analysis of REs will aid our understanding of repeat-mediated genome diversity, and genome evolution[63]. Abundance and diversity of the major repeat families were comparatively analyzed for between and within the three *Brassica* genomes based on 64 *Brassica* accessions (Fig. 6). Overall, significant copy number divergence was observed between A, C and AC genomes (Figs 2 and 6A,B). Of the three *Brassica* species surveyed, the highest proportion of the 10 MRs was found in the A genome (18.8%) compared with AC (11.5%) and C (10.8%) genomes. Large variations were observed between accessions of the different genomes, e.g., Br-8 in the A genome, and Bo-9 in the C genome represented the highest (22.56%) and lowest (4.6%) proportion of their respective genomes, demonstrating RE dynamics in these three *Brassica* genomes. Furthermore, repeat families such as CentBr1, 45 s nrDNA, STR-Br, and PCRBr were more abundant in the A genome than in the C or AC genomes, suggesting that evolution and amplification occurred after divergence from *B. oleracea* around 4.6 MYA (Supplementary Figure 1a).

Overall, relatively low diversity was observed within each of the three *Brassica* genomes (Fig. 6C–F). The *B. oleracea* genome had low divergence of the repeat fraction (3%) compared to *B. rapa* (>5%) and *B. napus* (>5%), suggesting that in the *B. oleracea* genome the repeat families were highly consistent and stably amplified even between different subspecies or morphotypes. This supports the idea that triplication and selection leads to genome diversification in *B. oleracea*[7]. However, few families showed diversity in the A genome, especially 5S nrDNA, CentBr1 and PCRBr1a, which deviated greatly between the different cultivars analyzed. This suggests that these repeats might be involved in evolution of the C genome subspecies.

Molecular cytogenetic analysis reveals asymmetrical evolution of major repeats of *B. rapa* and *B. oleracea*. Quantifying repeats based on FISH signals revealed that about 33%, 21%, and 30% of these respective genomes were occupied by 10 major genomic repeats, estimated to represent about 20%, 10% and 12% of the A, C, and AC genomes, respectively, based on *in silico* analysis (Fig. 2). The discrepancy between FISH and *in silico* analyses may be explained by the fact that FISH detects only two-dimensional hybridization signals from a three-dimensional chromosome structure. A wider area may also be covered by fluorescence than the actual hybridization loci. Altogether, FISH analysis provided an enhanced view of the genomic distribution and abundance of each RE. FISH-based quantification of the MRs thus enabled more accurate estimation of interspecies diversity in *Brassica* genomes, and insights into their evolution. Rapidly evolving asymmetrical amplification of MRs may promote speciation via chromosome reorganization and interruption of chromosome pairing, which can result in a reproduction barrier between organisms.

Materials and Methods

Plant materials. Leaf samples from 44 *B. oleracea* accessions were collected from two seed companies, Asia seed Co. and Joeun Seed Co., in South Korea used for resequencing. These accessions belong to eight phenotypic groups in seven subspecies (Supplementary Table 11). Total genomic DNA was extracted and purified using the

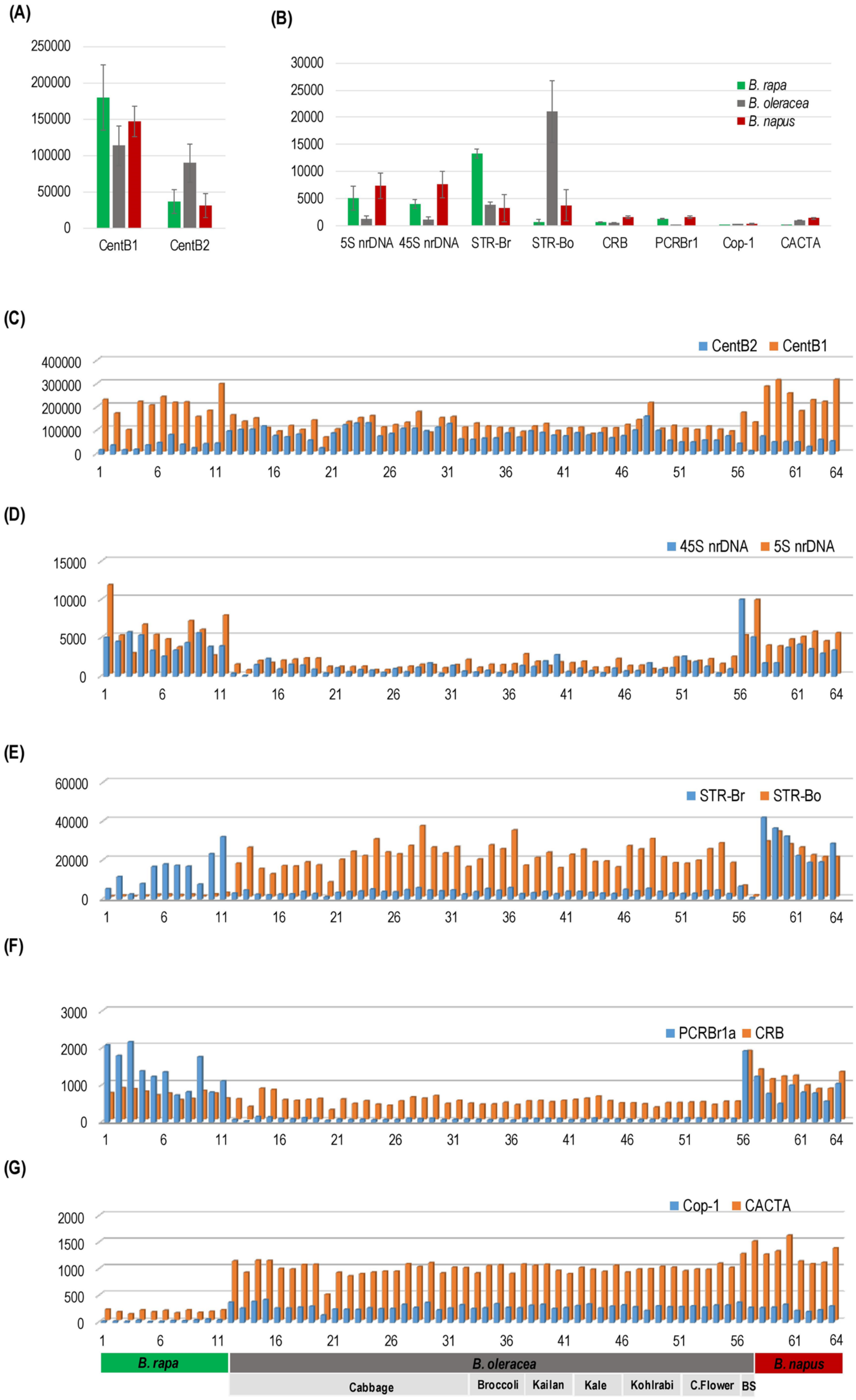

Figure 6. Survey of the composition of major repeats in *Brassica* genomes based on 1 × read mapping. Overall composition of repeats in these three *Brassica* genomes include centromeric tandem repeats (CentB-1 and CentB-2) (**A**), and 8 other repetitive elements **(B)**. Repeat abundance in 64 accessions of *B. oleracea* (Bo-1-44), *B. rapa* (Br-1-11) and *B. napus* (Bnp-1-9), including **(C)** CentB-1 and 2 **(D)** 5 S and 45 S ribosomal DNA (nrDNA), **(E)** STRs **(F)** centromere-specific retrotransposon of *Brassica* (CRB) and peri-centromeric retrotransposon of *B. rapa* (PCRBr1a), (**G**) Ty1/*Copia* retrotransposon of *Brassica* (BoCop-1), and CACTA DNA transposon of *B. oleracea* (BoCACTA).

modified cetyl-trimethyl ammonium bromide (CTAB) method[64]. The quantity and quality of the genomic DNA were examined using a nanodrop spectrometer.

Genomic datasets. Approximately 5 ng total genomic DNA from each *B. oleracea* accession was utilized to decode genomic information using an Illumina genome analyzer (Hiseq 2000) at Macrogen, (Seoul, South Korea).

Randomly sheared genomic libraries were prepared, with an insert size of 300 bp, following the 101 bp paired-end approach recommended by the manufacturer. Genome sequences of 11 *B. rapa* accessions belonging to eight phenotypic groups, and nine *B. napus*, were obtained from previous studies[40,65]. Raw reads were preprocessed using the CLC-quality trim tool to remove any remaining linker, adapter or low quality sequences. Sequences of 0.8x to 4.4x genome coverage from 64 *Brassica* accessions were utilized for further analyses (Supplementary Table 11).

De novo assembly and identification of highly abundant genomic components. We previously demonstrated the use of genome-skimming approach called *de novo* assembly of low-coverage WGS (dnaLCW) to obtain complete and simultaneous assemblies of chloroplast and nuclear ribosomal DNA genomes[50]. Low-coverage WGS, approximately 2x haploid genome-equivalent of NGS reads from *B. rapa* and *B. oleracea*, were used to independently retrieve the major and most highly abundant repetitive regions using a bioinformatics pipeline called dnaLCW-RE (Supplementary Figure 2). *De novo* assemblies of 2x haploid genome-equivalent WGS *B. rapa* (Br-1-1) from the quality reads filtered by the CLC-quality trim tool, were then assembled using CLC genome assembler (ver. 4.06, CLC Inc, Rarhus, Denmark) with 200–500 bp autonomously controlled overlap size. Genomic abundance in terms of average read depth (RD), along with the length of the contig (LC), were identified using a CLC reference assembly approach. The top 50 contigs of greatest depth were retrieved according to highest genomic representation (RD x LC). These were then annotated by BLASTn (best hit) against the *Repbase* database[66] and *Brassica*TED internal database, using previously reported *Brassica* REs, and Genbank. REs were classified as known repeats if contigs shared 80% similarity and 80% sequence alignment according the 80:80 rule[67]. Partial or truncated repeats containing contigs were manually analyzed and characterized with a complete structure based on reference sequence information or manual reads. Likewise, the *B. oleracea* (Bo-1-1) genome sequence was independently analyzed to identify the major REs using the abovementioned approach. Sequences of the individual repeat families for each species were stored in the *Brassica*TED[48]. In addition, graph-based clustering and characterization of repetitive elements by RepeatExplorer was performed using the 0.1x WGS reads from each *B. rapa*, *B. oleracea*, and *B. napus* genome[68]. Clustering was performed with the criteria of more than 90% sequences similarity and at least 55% sequence length cover were likely to be grouped into a single cluster. Clusters were characterized against *Repbase* database.

Quantification of repeat proportion in the three _Brassica_ genome assemblies. Whole-genome pseudo-chromosome assemblies with unanchored scaffold sequences were obtained for *B. rapa* (v2.1), *B. oleracea* (v1.0), and *B. napus* (v4.1) from Genbank and turned into an in-house database. Total copies of the MRs were identified based on BLASTn searches against the corresponding *Brassica* reference assembly. We followed the universal 80:80 rule for identification of members including intact and diverse members from the three reference sequences[67]. In this approach, a repeat element should have at least 80% sequence similarity and 80% sequence coverage to be considered a full-length element[44]. Members were then classified based on maximum identity, i.e., if hits were produced at the same position with more than 80% sequence similarity between them (especially for TRs)[33]. Copy numbers of the each repeats were estimated by read depth (RD) approach quantified using WGS from 64 *Brassica* accessions. RD approach has been one of the major approach for copy number estimation. The basis of RD approach is that to calculate the depth of the coverage of a genomic region is corresponds with the copy number of the region which are expected to provide relatively accurate estimation[69]. Assuming that the WGS reads used in this experiment were randomly sampled without bias, abundance was then quantified using WGS from 64 *Brassica* accessions based on the CLC-reference assembly. Paired-reads were mapped to the MRs with high threshold set at greater than 80% identity and over 50% of the read length. And the overall mean of the read depth were calculated according to the numbers of reads were mapped on to the MRs. Outputs were normalized to 1x genome coverage for *B. rapa*, *B. oleracea*, and *B. napus* genomes based on corresponding genome sizes. And copies of MRs were multiplied by its size to calculate the MR abundance in total genome (GA) and the genomic proportion of each MR representing total genome was calculated.

Fluorescence _in situ_ hybridization analysis. Mitotic metaphase chromosome spreads were obtained from root samples from commercial hybrid seeds *B. oleraceae* ssp. *capitata* 'Sun Power', *B. rapa* ssp. *pekinensis* 'Saeronam Spring' (Asia Seed Company, Korea) and *B. napus* ssp. *napus* 'Tapidor' (*Brassica* seedbank, Chungnam National University, South Korea) according to a previous study[47].

Repeat-specific probes were developed based on multiple sequence alignment, and primers were designed using the NCBI primer BLAST tool (Supplementary Table 12). Repeat-specific probes were then confirmed via PCR amplification of *B. oleracea* and *B. rapa* genomic DNA. Probes were labeled by direct nick translation using the fluors mentioned in Supplementary Table 12. The hybridization solution contained 50% formamide, 2x saline-sodium citrate buffer, with or without 5 ng/μl salmon sperm DNA, 10% dextran sulfate, and 25 ng/μl of each DNA probe, adjusted to a total volume of 40 μl/slide with DNase-free and RNase-free water (Sigma, USA, #W4502). FISH experiments, including slide pre-treatment, probe hybridization and signal detection, were carried out as reported by Waminal *et al.* (2012).

References

1. Kiefer, M. *et al.* BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution. *Plant and Cell Physiology*, pct158 (2013).
2. U, N. Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization. *Jap J Bot* **7**, 389–452 (1935).
3. Tank, D. C. *et al.* Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist* **207**, 454–467 (2015).
4. Mun, J. H. *et al.* Sequence and structure of Brassica rapa chromosome A3. *Genome biology* **11**, R94, https://doi.org/10.1186/gb-2010-11-9-r94 (2010).

5. Yang, T. J. *et al*. Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of Brassica rapa. *The Plant cell* **18**, 1339–1347, https://doi.org/10.1105/tpc.105.040535 (2006).
6. Cheng, F., Wu, J. & Wang, X. Genome triplication drove the diversification of Brassica plants. *Horticulture Research* **1**, 14024 (2014).
7. Cheng, F. *et al*. Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. *Nature genetics* **48**, 1218–1224 (2016).
8. Fedoroff, N. V. Transposable elements, epigenetics, and genome evolution. *Science* **338**, 758–767 (2012).
9. Fang, L., Cheng, F., Wu, J. & Wang, X. The Impact of Genome Triplication on Tandem Gene Evolution in Brassica rapa. *Frontiers in plant science* **3**, 261, https://doi.org/10.3389/fpls.2012.00261 (2012).
10. Michael, T. P. & VanBuren, R. Progress, challenges and the future of crop genomes. *Current opinion in plant biology* **24**, 71–81 (2015).
11. Sigman, M. J. & Slotkin, R. K. The first rule of plant transposable element silencing: location, location, location. *The Plant cell* **28**, 304–313 (2016).
12. Mehrotra, S. & Goyal, V. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. *Genomics, proteomics & bioinformatics* **12**, 164–171, https://doi.org/10.1016/j.gpb.2014.07.003 (2014).
13. Melters, D. P. *et al*. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology* **14**, R10, https://doi.org/10.1186/gb-2013-14-1-r10 (2013).
14. Hardman, N. Structure and function of repetitive DNA in eukaryotes. *The Biochemical journal* **234**, 1–11 (1986).
15. Heslop-Harrison, J. & Schmidt, T. Plant nuclear genome composition. *eLS* (2012).
16. Yang, K. *et al*. Diversity and Inheritance of Intergenic Spacer Sequences of 45S Ribosomal DNA among Accessions of Brassica oleracea L. var. capitata. *International journal of molecular sciences* **16**, 28783–28799 (2015).
17. Koo, D. H. *et al*. Rapid divergence of repetitive DNAs in Brassica relatives. *Genomics* **97**, 173–185, https://doi.org/10.1016/j.ygeno.2010.12.002 (2011).
18. Macas, J., Neumann, P. & Navratilova, A. Repetitive DNA in the pea (Pisum sativum L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. *BMC genomics* **8**, 427, https://doi.org/10.1186/1471-2164-8-427 (2007).
19. Lim, K. B. *et al*. Characterization of the centromere and peri-centromere retrotransposons in Brassica rapa and their distribution in related Brassica species. *The Plant journal: for cell and molecular biology* **49**, 173–183, https://doi.org/10.1111/j.1365-313X.2006.02952.x (2007).
20. Wei, L. *et al*. New insights into nested long terminal repeat retrotransposons in Brassica species. *Molecular plant* **6**, 470–482, https://doi.org/10.1093/mp/sss081 (2013).
21. Nowak, R. Mining treasures from'junk DNA'. *Science* **263**, 608–610 (1994).
22. Shapiro, J. A. & von Sternberg, R. Why repetitive DNA is essential to genome function. *Biological reviews of the Cambridge Philosophical Society* **80**, 227–250 (2005).
23. Pardue, M. L. & DeBaryshe, P. G. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annual review of genetics* **37**, 485–511, https://doi.org/10.1146/annurev.genet.38.072902.093115 (2003).
24. Hall, S. E., Luo, S., Hall, A. E. & Preuss, D. Differential Rates of Local and Global Homogenization in Centromere Satellites From Arabidopsis Relatives. *Genetics* **170**, 1913–1927, https://doi.org/10.1534/genetics.104.038208 (2005).
25. Biémont, C. & Vieira, C. Genetics: junk DNA as an evolutionary force. *Nature* **443**, 521–524 (2006).
26. Choi, H. I. *et al*. Major repeat components covering one-third of the ginseng (Panax ginseng C.A. Meyer) genome and evidence for allotetraploidy. *The Plant journal: for cell and molecular biology* **77**, 906–916, https://doi.org/10.1111/tpj.12441 (2014).
27. Parisod, C. *et al*. Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytologist* **186**, 37–45 (2010).
28. Lisch, D. How important are transposons for plant evolution? *Nature reviews. Genetics* **14**, 49–61 (2013).
29. Sampath, P., Lee, J., Cheng, F., Wang, X. & Yang, T.-J. in *The Brassica rapa Genome* 65–81 (Springer, 2015).
30. Kelly, L. J. *et al*. Analysis of the giant genomes of Fritillaria (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytologist* **208**, 596–607 (2015).
31. Huang, X., Lu, G., Zhao, Q., Liu, X. & Han, B. Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant physiology* **148**, 25–40 (2008).
32. Naito, K. *et al*. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461**, 1130–1134, https://doi.org/10.1038/nature08479 (2009).
33. Liu, S. *et al*. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature communications* **5**, 3930, https://doi.org/10.1038/ncomms4930 (2014).
34. Edwards, D. & Batley, J. Plant genome sequencing: applications for crop improvement. *Plant biotechnology journal* **8**, 2–9 (2010).
35. Renny-Byfield, S. *et al*. Next generation sequencing reveals genome downsizing in allotetraploid Nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNAs. *Molecular biology and evolution*, msr112 (2011).
36. Varshney, R. K. & May, G. D. Next-generation sequencing technologies: opportunities and obligations in plant genomics. *Brief Funct Genomics* **11**, 1–2, https://doi.org/10.1093/bfgp/els001 (2012).
37. Gao, D., Jiang, N., Wing, R. A., Jiang, J. & Jackson, S. A. Transposons play an important role in the evolution and diversification of centromeres among closely related species. *Frontiers in plant science* **6** (2015).
38. Wang, X. *et al*. The genome of the mesopolyploid crop species Brassica rapa. *Nature genetics* **43**, 1035–1039, https://doi.org/10.1038/ng.919 (2011).
39. Parkin, I. A. *et al*. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome biology* **15**, R77, https://doi.org/10.1186/gb-2014-15-6-r77 (2014).
40. Chalhoub, B. *et al*. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345**, 950–953 (2014).
41. Cai, C. *et al*. Brassica rapa Genome 2.0: A Reference Upgrade through Sequence Re-assembly and Gene Re-annotation. Mo*lecular plant* **10**, 649–651, https://doi.org/10.1016/j.molp.2016.11.008.
42. Lim, K. B. *et al*. Characterization of rDNAs and tandem repeats in the heterochromatin of Brassica rapa. *Molecules and cells* **19**, 436–444 (2005).
43. Sampath, P. *et al*. Characterization of a new high copy Stowaway family MITE, BRAMI-1 in Brassica genome. *BMC Plant Biol* **13**, 56 (2013).
44. Sampath, P. *et al*. Genome-Wide Comparative Analysis of 20 Miniature Inverted-Repeat Transposable Element Families in Brassica rapa and B. oleracea. *PloS one* **9**, e94499, https://doi.org/10.1371/journal.pone.0094499 (2014).
45. Sampath, P. & Yang, T.-J. Comparative analysis of Cassandra TRIMs in three Brassicaceae genomes. *Plant Genetic Resources* **12**, S146–S150, https://doi.org/10.1017/S1479262114000446 (2014).
46. Waminal, N. E., Perumal, S., Lee, J., Kim, H. H. & Yang, T.-J. Repeat Evolution in Brassica rapa (AA), B. oleracea (CC), and B. napus (AACC) Genomes. *Plant Breeding and Biotechnology* **4**, 107–122 (2016).
47. Waminal, N. E. *et al*. In *The Brassica rapa Genome* 83–96 (Springer, 2015).
48. Murukarthick, J. *et al*. BrassicaTED - a public database for utilization of miniature transposable elements in Brassica species. *BMC Research Notes* **7**, 379 (2014).
49. Yang, T. J. *et al*. Characterization of terminal-repeat retrotransposon in miniature (TRIM) in Brassica relatives. *TAG. Theoretical and applied genetics. Theoretische und angewandte. Genetik* **114**, 627–636, https://doi.org/10.1007/s00122-006-0463-3 (2007).

50. Kim, K. *et al*. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species. *Scientific reports* **5**, 15655 (2015).
51. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Human Molecular Genetics* **19**, R227–R240, https://doi.org/10.1093/hmg/ddq416 (2010).
52. Michael, T. P. & Jackson, S. The First 50 Plant Genomes. *Plant Genome-Us* **6**, https://doi.org/10.3835/plantgenome2013.03.0001in (2013).
53. Metzker, M. L. Next Generation Technologies: Basics and Applications. *Environ Mol Mutagen* **51**, 691–691 (2010).
54. Schatz, M. C., Witkowski, J. & McCombie, W. R. Current challenges in de novo plant genome sequencing and assembly. *Genome biology* **13**, 243, https://doi.org/10.1186/gb4015 (2012).
55. Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human genome sequencing in health and disease. *Annual review of medicine* **63**, 35–61, https://doi.org/10.1146/annurev-med-051010-162644 (2012).
56. Golicz, A. A. *et al*. The pangenome of an agronomically important crop plant Brassica oleracea. *Nature communications* **7**, 13390, https://doi.org/10.1038/ncomms13390 (2016).
57. Swaminathan, K., Varala, K. & Hudson, M. E. Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC genomics* **8**, 132, https://doi.org/10.1186/1471-2164-8-132 (2007).
58. Hawkins, J. S., Kim, H., Nason, J. D., Wing, R. A. & Wendel, J. F. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res* **16**, 1252–1261, https://doi.org/10.1101/gr.5282906 (2006).
59. Barghini, E. *et al*. The peculiar landscape of repetitive sequences in the olive (Olea europaea L.) genome. *Genome biology and evolution* **6**, 776–791 (2014).
60. Chalopin, D., Naville, M., Plard, F., Galiana, D. & Volff, J.-N. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome biology and evolution* **7**, 567–580, https://doi.org/10.1093/gbe/evv005 (2015).
61. Alix, K. *et al*. The CACTA transposon Bot1 played a major role in Brassica genome divergence and gene proliferation. *The Plant journal: for cell and molecular biology* **56**, 1030–1044, https://doi.org/10.1111/j.1365-313X.2008.03660.x (2008).
62. Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. & Schulman, A. H. Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences* **97**, 6603–6607 (2000).
63. Zhang, X. & Wessler, S. R. Genome-wide comparative analysis of the transposable elements in the related species Arabidopsis thaliana and Brassica oleracea. *Proc Natl Acad Sci USA* **101**, 5589–5594, https://doi.org/10.1073/pnas.0401243101 (2004).
64. Allen, G., Flores-Vergara, M., Krasynanski, S., Kumar, S. & Thompson, W. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nature protocols* **1**, 2320–2325 (2006).
65. Schmutzer, T. *et al*. Species-wide genome sequence and nucleotide polymorphisms from the model allopolyploid plant Brassica napus. *Scientific data* **2**, 150072, https://doi.org/10.1038/sdata.2015.72 (2015).
66. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462–467 (2005).
67. Wicker, T. *et al*. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973–982, https://doi.org/10.1038/nrg2165 (2007).
68. Novak, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793, https://doi.org/10.1093/bioinformatics/btt054 (2013).
69. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics* **14**(Suppl 11), S1, https://doi.org/10.1186/1471-2105-14-S11-S1 (2013).

Acknowledgements

This work was supported by Golden Seed Project (Centre for Horticultural Seed Development, No. 213003-04-4-SB430), Ministry of Agriculture, Food and Rural Affairs (MAFRA) and the Next Biogreen 21 program (PJ01101101) of the Rural Development Administration in the Republic of Korea. We also grateful to Minkyu Park, University of Georgia, and Yasuhito Sakuraba, University of Tokyo, for their critical comments.

Author Contributions

S.P. and T.J.Y. planned and designed the research. Jonghoon L., J.L. and B.C. contributed materials and *in silico* analysis. S.P. performed experiments and analyzed data. N.E.W. and H.H.K., performed FISH analysis. S.P. and T.J.Y. wrote the manuscript with the help of N.E.W. and M.A.G.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-18048-9.

Competing Interests: The authors declare that they have no competing interests.

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