Genomic Survey of the Hidden Components of the *B. rapa* Genome

7

Nomar Espinosa Waminal, Sampath Perumal, Ki-Byung Lim, Beom-Seok Park, Hyun Hee Kim and Tae-Jin Yang

Abstract

The sequencing of the *Brassica rapa* genome has enabled better understanding of its structure and evolution, and created numerous opportunities for exploration of genome function and breeding applications. Nevertheless, the currently available completed genome sequences are estimated to cover only about 60 % of the genome, while the remaining 40 % is unassembled mainly due to the highly repetitive nature of this portion of the genome. Elucidation of the nature and distribution of repeat elements in the context of the entire genome would enhance our understanding of their role in genome structure, function, and evolution. In this chapter, we review the genomic distribution, characterization and evolutionary implications of currently identified repeat elements comprising the 'hidden' portion of the *B. rapa* genome. Low-coverage whole-genome sequence (WGS) was used to survey the major genomic repeats and their proportion in the *B. rapa* genome. Coupling this with molecular cytogenetics, we characterized the abundance and genomic distribution of seven major repeats, namely centromeric tandem repeats 1 and 2, centromeric retrotransposons, pericentromeric retrotransposons, 5S rDNA, 45S rDNA, and subtelomeric tandem repeats. These repeats accounted for approximately 20 % of the *B. rapa* genome, which is much more than the <1 % covered by repeats in the currently available genome

N.E. Waminal · S. Perumal · T.-J. Yang (✉)
Department of Plant Sciences, Plant Genomics and Breeding Institute and Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Republic of Korea
e-mail: tjyang@snu.ac.kr

K.-B. Lim
Department of Horticultural Science, Kyungpook National University, Daegu 702-701, Korea

B.-S. Park
National Academy of Agricultural Science, Rural Development Administration, 150 Suinro, Suwon 441-707, Republic of Korea

H.H. Kim
Department of Life Science, Plant Biotechnology Institute, Sahmyook University, Seoul 139-742, Republic of Korea

X. Wang and C. Kole (eds.), *The Brassica rapa Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-47901-8_7

assembly. We also compared their distributions among different *B. rapa* accessions and in the close relative *Brassica oleracea,* for better understanding of the plasticity of the *Brassica* genomes.

7.1 Introduction

Knowledge of genome sequences has a huge impact in plant biology (Schadt et al. 2010). The number of plant genomes being sequenced is rising (Michael and Jackson 2013) due to the rapid advancement of genome sequencing technologies, including those that allow high-throughput sequencing of longer reads and high-resolution assembly algorithms (Edwards and Batley 2010; Metzker 2010; Schatz et al. 2012). However, a common hurdle is assembly accuracy, especially considering the highly repetitive nature of plant genomes (Macas et al. 2007; Schatz et al. 2012). For example, bread wheat, which has one of the largest genomes among those sequenced from plants (17,000 Mbp; Brenchley et al. 2012), has an estimated repeat content of 80 % and the sequences assembled into scaffolds covered only 22 % of the genome (Brenchley et al. 2012; Michael and Jackson 2013). Even for Chinese cabbage (*Brassica rapa*), which has a relatively small genome of 529 Mbp, only about 60 % of the genome was assembled into pseudo-chromosome sequences, with the remaining 40 % made up mainly of repeat elements (Johnston et al. 2005; Wang et al. 2011; Michael and Jackson 2013).

Repetitive components of genomes are responsible for the extensive genome size variation in higher plants (Hardman 1986; Pagel and Johnstone 1992; Macas et al. 2007) and used to be considered 'junk' (Doolittle and Sapienza 1980; Nowak 1994; Shapiro and von Sternberg 2005). However, many recent studies have shown that repetitive elements have diverse functions within cells (Biémont and Vieira 2006; Biémont 2010), from involvement in maintaining chromosome integrity (Nowak 1994), and gene expression (Biémont and Vieira 2006), to changing phenotypes (Biémont and Vieira 2006). Therefore, characterization of these components in relation to genome assemblies is fundamental to understanding the holistic landscape and deciphering the complexity of plant genomes (Biémont 2010).

Despite their importance, repetitive sequences have hindered genome assembly and increased costs in terms of both time and money (Schatz et al. 2012). They remain largely unexplored and unassembled in many sequenced plant genomes (Wang et al. 2011; Michael and Jackson 2013; Liu et al. 2014), because most assembly algorithms are designed for less complex sequences (Schatz et al. 2012). However, the large amount of information that could be gathered from these repeats would be useful for understanding genome structure and evolution (Biémont 2010).

In the assembled genome sequences, most of the repetitive elements that occupy ~40 % of the *B. rapa* genome are transposon related (Wang et al. 2011; Michael and Jackson 2013). However, more redundant repeats such as centromeric and pericentromeric LTR retrotransposons (CRBs and PCRBrs, respectively; Lim et al. 2007), centromeric tandem repeats (including CentBr1 and CentBr2; Lim et al. 2005), and subtelomeric tandem repeats (STRs; Koo et al. 2011), in addition to the rDNA arrays were not included in the assembled genome sequence. Less than 1 % of these repeats are included in the currently available 283 Mbp assembled sequences (Table 7.1) despite coverage of >98 % of the euchromatic regions (Wang et al. 2011). This discrepancy demonstrates the difficulty of anchoring repeats in the assembly. Characterizing, quantifying and cytogenetically mapping these elements should aid in the final refinement of the genome structure.

Table 7.1 Comparison of major repeat composition identified in the reference genome assembly of *B. rapa* 'Chiifu' (Wang et al. 2011) with that found in 1x WGS sequence of 11 *B. rapa* accessions

Repeat element	Unit length (bp)	Reference genome (283 Mbp) GR[a]						1x WGS (529 Mbp)[b]		GP by FISH (%)	Genome appearance (%)[d]
		256 Mbp pseudo-molecule		Unanchored scaffold		Total (a + b) (kb) (A)	GP[c] (%)	GR (Kbp) (B)	GP (%)		
		Copy	(kb) (a)	Copy	(kb) (b)						
CENTBr1	176	48	8.1	51	8.8	16.9	0.006	34,700 (±8568)	6.56	11.4	0.0
CENTBr2	176	147	25.3	93	16.1	41.4	0.015	7095 (±3177)	1.34	2.3	0.5
STR	351	831	272.6	135	43.3	315.9	0.112	5908 (±2942)	1.12	2.4	2.6
45S rDNA	7764	0	4.2	0	3.2	7.4	0.003	42,534 (±13,265)	8.04	5.9	0.0
5S rDNA	501	12	5.9	5	2.5	8.4	0.003	2631 (±1160)	0.50	1.7	0.2
CRB	5908	1	5.9	0	0.0	5.9	0.002	4098 (±613)	0.77	2.5	0.2
PCRBr	8395	0	0.0	0	0.0	0.0	0.000	11,221 (±4087)	2.12	3.3	0.0
Total		1039	322.4	284	73.9	395.9	0.140	108,186 (±15,415)	20.45	29.5	0.3

[a]Genome representation: average read depth × contig length
[b]Average value from 11 *Brassica rapa* accessions
[c]Genome proportion: (total GR/reference genome size in kbp) × 100
[d]Appearance in genome sequence based on the GR value of Chiifu WGS (%) = (A/B) × 100

In this chapter, we describe a genomic survey for major repeats of *B. rapa* using 1x whole-genome sequence (WGS) that captured a substantial portion of previously reported repeats and allowed us to characterize others. We also review the possible evolutionary roles of the identified repetitive elements in shaping the *B. rapa* genome. We further demonstrate the utility of combining in silico mapping of low-coverage WGS and fluorescence in situ hybridization (FISH) techniques to localize and estimate the genomic distribution and abundance of each repeat family. Finally, we discuss exciting applications and future prospects for this approach, especially for large and repeat-replete genomes and resource-deficient plant species.

7.2 The Hidden Genome: Characterization of Major Repeats

Knowing the distribution of repetitive elements within a genome is important in understanding genome organization, evolution, and function (Harrison and Heslop-Harrison 1995). In *B. rapa*, analysis using mitotic chromosome spreads demonstrated that heterochromatin is mostly concentrated in the centromeric and pericentromeric regions (Lim et al. 2005). These regions were later shown to contain major repetitive elements including the centromeric tandem repeats CentBr1 and CentBr2, centromeric retrotransposon of *Brassica* (CRB) and peri-centromeric retrotransposon of *B. rapa* (PCRBr; Harrison and Heslop-Harrison 1995; Lim et al. 2000, 2005; Koo et al. 2004). Repeats that are not concentrated in the centromeric regions have also been characterized (Wang et al. 2011; Liu et al. 2014). In addition to the tandemly repeated housekeeping 5S and 45S rRNA genes, a tandem repeat named STR based on its localization in the subtelomeric regions of several *Brassica* species was recently discovered (Koo et al. 2011). Collectively, these elements constitute the major repeat components of the hidden portion of the *B. rapa* genome (Table 7.1).

Most of these repeats have been identified by capture and characterization of single or a few elements via various efforts by independent groups; thus, global and comparative analyses of repetitive elements among related genomes has been limited (Macas et al. 2007). Oftentimes, considerable time and resources were spent to characterize these elements. For example, CentBr1, CentBr2, CRB, and PCRBr were isolated after identification of patterns in restriction enzyme digestion, screening several thousand BAC clones, downstream cloning of isolated sequences, sequencing, and cytogenetic mapping (Harrison and Heslop-Harrison 1995; Koo et al. 2004; Lim et al. 2005, 2007). With the current availability of NGS technology, a huge amount of information now awaits capture and utilization without the tedium and expense of more traditional approaches.

7.2.1 Reconstruction of Nuclear rDNA Units

Owing to the vital function they play in protein biosynthesis and cellular function, ribosomal RNA genes are highly conserved across plant species (Hershkovitz and Zimmer 1996; Martins and Wasko 2004; Waminal et al. 2014). However, the spacers between each rDNA repeat unit are more divergent among species, making them an excellent tool for phylogenetic studies (Martins and Wasko 2004). Additionally, they have been exploited as cytogenetic FISH markers for studies related to genome dynamics and evolution (Roa and Guerra 2012; Waminal et al. 2012). However, complete sequences of *B. rapa* nuclear rDNAs have not yet been reported. Using de novo assembly of low-coverage WGSs (dnaLCW; Kim et al. 2015), we obtained the complete 5S unit without gaps and 45S rDNA unit sequences with small gaps in the intergenic spacer (IGS) for *B. rapa*.

The complete 5S rDNA unit was 501 bp, comprising a 120-bp 5S rRNA gene and 381-bp IGS (Fig. 7.1a). Based on mapping of raw reads

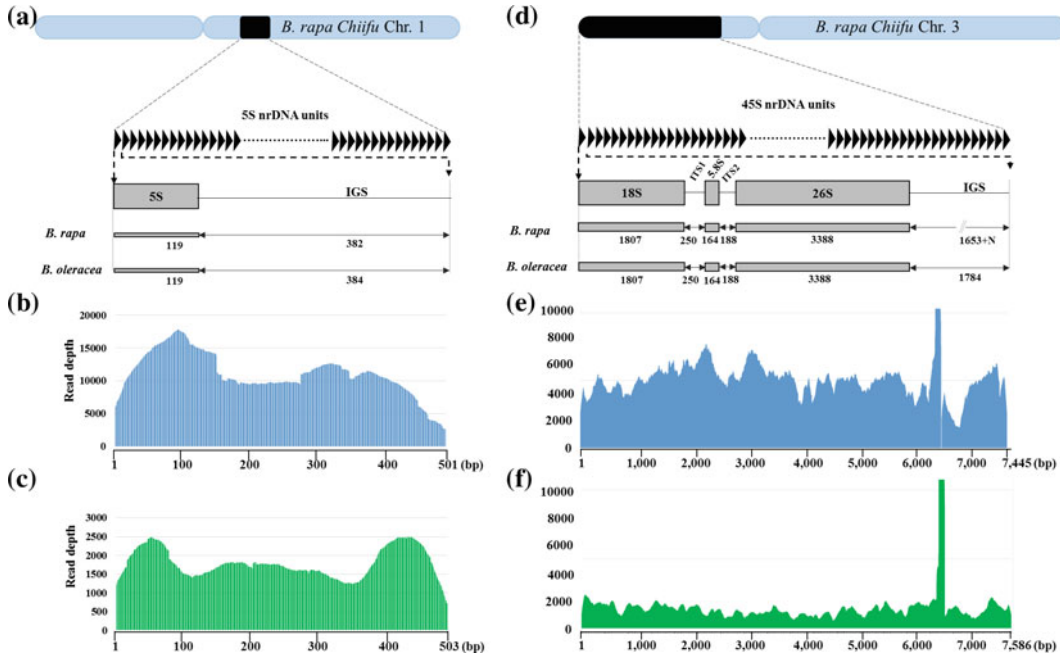

Fig. 7.1 Structure of 5S and 45S rDNAs of *B. rapa* 'Chiifu' and *B. oleracea* C1234 and raw read mapping. **a** Structure of the complete 5S rDNA unit of *B. rapa* and *B. oleracea* assembled based on the dnaLCW method (Kim et al. 2015). **b**, **c** Coverage of the 5S rDNA unit based on raw read mapping against the 1x genomes of *B. rapa* (genbank no: KM538957) and *B. oleracea* (genbank no: KM538957), respectively. **d** Structure of the 45S rDNA unit of *B. rapa* (partial) (genbank no: KM538957) and *B. oleracea* (complete) (genbank no: KM538957) assembled based on the dnaLCW method. **e**, **f** Coverage of 45S rDNA unit based on raw read mapping against the 1x genome of *B. rapa* and *B. oleracea*, respectively

to the complete 5S rDNA contig (Fig. 7.1b), it was estimated that there were 16,756 copies of the 5 rDNA unit in the haploid 'Chiifu' genome (Fig. 7.2c). Likewise, the complete 5S rDNA unit for *Brassica oleracea* 'C1234' totaled 503 bp with 119-bp genic and 384-bp IGS regions. However, only 1743 copies were estimated to be present in the *B. oleracea* genome based on raw read mapping (Fig. 7.1c); a value much lower than that in the *B. rapa* 'Chiifu' genome, and supported by FISH analysis (Fig. 7.3a, b, e, f). Obtaining the complete unit of the 45S rDNA sequence for *B. rapa* 'Chiifu' was hindered by GC-rich repeats in the IGS region. Due to the abundant subrepeat regions and possible heterogeneous sequences in the IGS, gap-filling methods were ineffective, leaving a small gap in the 45S rDNA unit of 7764 bp for *B. rapa* 'Chiifu' (Fig. 7.1d). Nevertheless, using the same methods we successfully obtained a complete 7586-bp 45S rDNA unit for *B. oleracea*. Mapping 1x NGS reads to 45S rDNA sequences of *B. rapa* 'Chiifu' and *B. oleracea* C1234 (Fig. 7.1e, f) revealed 8709 and 1339 copies, respectively (Fig. 7.2c).

7.2.2 Exploring the Hidden Portion of the Genome

A few studies have reconstructed and estimated the genomic content of major repeats using low-coverage NGS sequences (Hawkins et al. 2006; Macas et al. 2007; Swaminathan et al. 2007). This approach allowed the identification of up to 48 % of the 75–97 % repeats in the 4300 Mbp *Pisum sativum* genome (Macas et al. 2007). Even though not all of the repeats were captured in silico, enough information was available to carry out comparative studies among closely-related species. Coupled with FISH, this

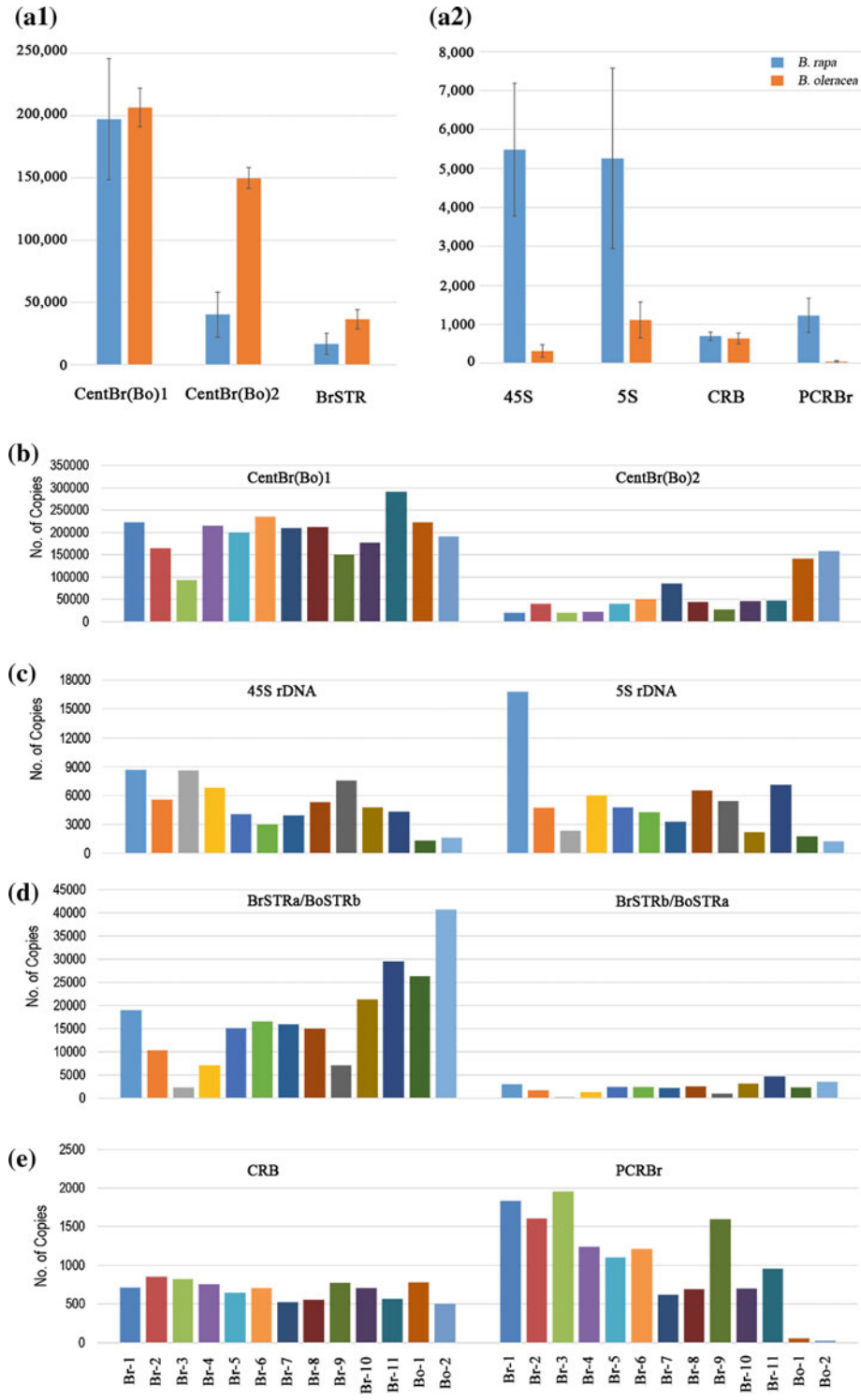

Fig. 7.2 Sequences identified via genomic survey of major repeats among 11 different *B. rapa* and two *B. oleracea* accessions for comparison. **a1** Comparison of centromeric and subtelomeric tandem repeat copy numbers and **a2** rDNA and centromeric retrotransposons between *B. rapa* and *B. oleracea*. Error bars represent standard deviation. Copy numbers of **b** centromeric tandem repeats of *B. rapa* (CentBr), **c** ribosomal DNA (rDNA), **d** *B. rapa* and *B. oleracea* subtelomeric satellite repeats (BrSTR and BoSTR, respectively), and **e** centromere-specific retrotransposon of *Brassica* (CRB) and peri-centromeric retrotransposon of *B. rapa* (PCRBr)

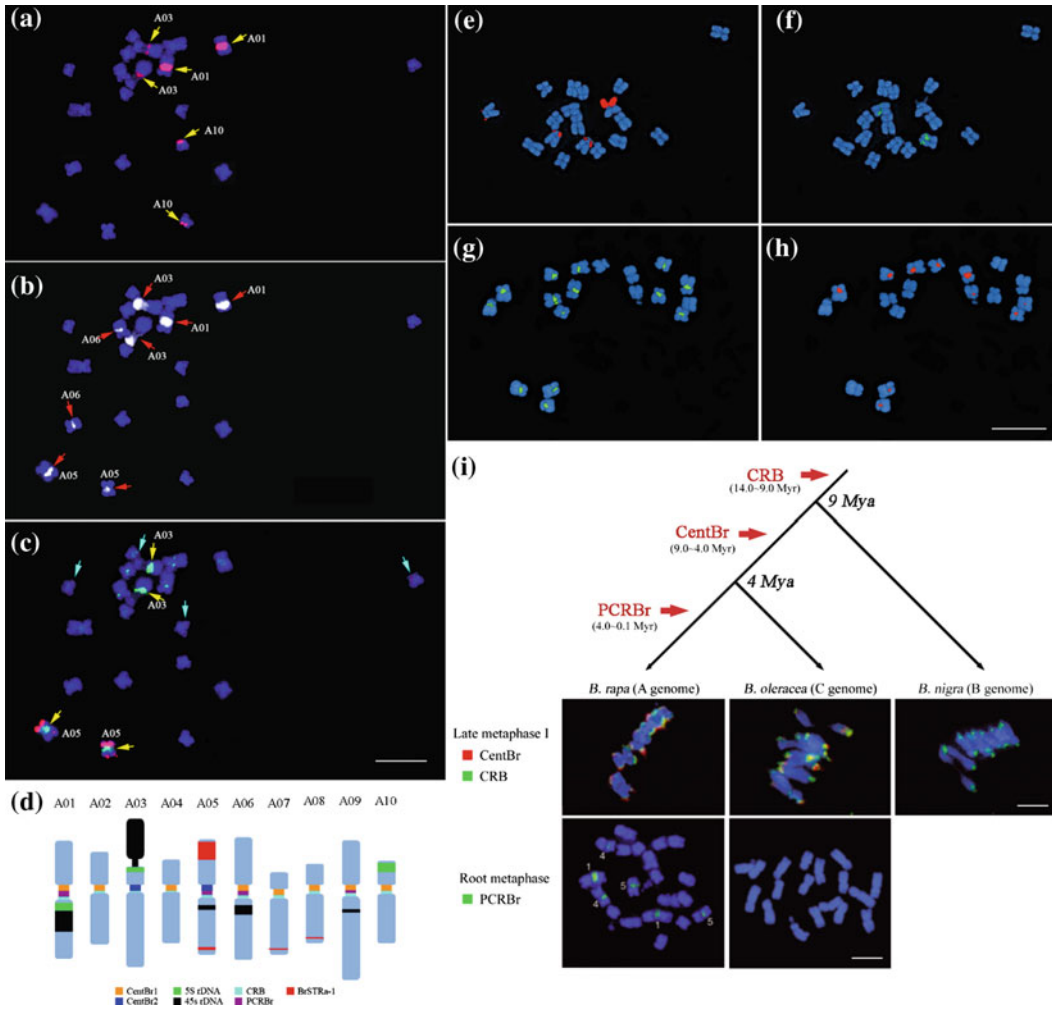

Fig. 7.3 Cytogenetic mapping and evolution of *B. rapa* and *B. oleracea* major repeats. *B. rapa* **a** FISH signals of 5S rDNA (*yellow arrows*), **b** 45S rDNA (*red arrows*), **c** CentBr2 (*green signals, yellow arrows* indicate 4 major signals) and BrSTR (*red*, on both arms of chromosome A05, *blue arrows* indicate weak BrSTR signals) in *B. rapa* root metaphase chromosomes. **d** Karyotype idiogram showing the cytogenetic distribution of major repeats. Chromosome numbering is according to Xiong and Pires (2011). **e**–**h** *B. oleracea*. **e** 45S rDNA **f** 5S rDNA **g** CentBo1, and **h** CentBo2. **i** Genome-specific evolution of *Brassica* centromeric repeats showing lineage divergence (Mya) at nodes and repeats with corresponding estimated insertion and amplification time (Myr). Bars in **a**–**h** = 10 μm, **i** = 5 μm

approach was able to reveal the distribution of the newly identified tandem repeats, providing a better picture of their actual location and abundance in the genome.

In *B. rapa*, several types of major repeats have been characterized, including the centromeric and pericentromeric LTR retrotransposons, CRB and PCRBr, respectively (Lim et al. 2007), centromeric tandem repeats CentBr1 and CentBr2 (Lim et al. 2005), and subtelomeric tandem repeat STR (Koo et al. 2011). We used these publicly available sequences along with the *B. rapa* rDNA sequences we assembled herein to survey the abundance of each element using 1x Illumina WGS data with at least 80 % sequence similarity as a criterion. As stated above,

repetitive elements currently identified in the *B. rapa* pseudo-chromosome sequences covered less than 1 % of the total assembled sequence (Wang et al. 2011). Here, we identified repetitive elements representing more than 20 % of the genome. Accordingly, only 0.3 % of these sequences are represented in the current genome assembly (Table 7.1).The most abundant repeats in the *B. rapa* genome were 45S rDNA (8 %), followed by CentBr1 (7 %) and PCRBr (2 %).

In *B. rapa* (A genome), CentBr1 is more abundant than CentBr2 (Fig. 7.2a, b), unlike their orthologous sequences in *B. oleracea* (C genome), CentBo1 and CentBo2, which are present in similar copy numbers (Fig. 7.2a, b; Lim et al. 2007; Koo et al. 2011). This was supported by our 1x WGS survey of 11 *B. rapa* and two *B. oleracea* accessions that revealed large copy number differences between CentBr1 and CentBr2, but not much difference between CentBo1 and CentBo2 (Fig. 7.2b).

The 1x WGS survey also identified >5000 and >300 times more 45S and 5S rDNA, respectively, than what was included in the assembled pseudo-chromosome sequences (Table 7.1). When compared to *B. oleracea*, *B. rapa* had 5 and 17 times more copies of 5S and 45S rDNA, respectively (Fig. 7.2a), which was consistent with FISH results (Fig. 7.3a, b, e, f; Xiong and Pires 2011).

Previous reports have identified two classes of subtelomeric tandem repeats in *Brassica*, STRa and STRb which share 89 % sequence identity (Koo et al. 2011). More sequences were identified from the 1x WGS reads when searching with BrSTRa compared to BrSTRb, suggesting that BrSTRa type TR sequences are more abundant than BrSTRb type sequences in both the *B. rapa* and *B. oleracea* genomes (Fig. 7.2d). In addition, different accessions of *B. rapa* and *B. oleracea* showed orders of magnitude difference in abundance for other repeat elements, indicative of genome plasticity which may reflects phenotypic polymorphism among accessions (Fig. 7.2b–e).

There was not much copy number variation for CRB elements among different *B. rapa* and *B. oleracea* accessions (Fig. 7.2e), supporting their common existence in the genus *Brassica* (Lim et al. 2007). By contrast, PCRBr was significantly more abundant in *B. rapa* compared with the negligible amount found in *B. oleracea* (Fig. 7.2e), supporting the observation of Lim et al. (2007) that PCRBr is specific to the A genome.

7.2.3 Cytogenetic Mapping of Repetitive Elements

FISH is an invaluable tool in genetic and genomic studies. It has allowed confirmation of chromosomal segment inversions (van der Knaap et al. 2004; Huang et al. 2009; Cabo et al. 2014), localization of centromeric repeats (Lee et al. 2005; Wolfgruber et al. 2009), visualization of transposons (Yu et al. 2007; Neumann et al. 2011) and repetitive elements (Lamb et al. 2007a; Macas et al. 2007; Suzuki et al. 2012), and even detection of single genes (Khrustaleva and Kik 2001; Lamb et al. 2007b) and transgenes (Santos et al. 2006; Park et al. 2010). Macas et al. (2007) demonstrated the utility of FISH to cytogenetically map the major repeats identified in the pea genome in a survey of 454 NGS sequence data. Additionally, there are some limitations in identifying these repetitive elements through computational analysis, which may not always accurately report the proportion of repeats that resides in that genome (Macas et al. 2007; Schatz et al. 2012).

With our analysis of the *B. rapa* genome, FISH data afforded us a better view of the genomic proportion of each repetitive element. Whereas about 20 % of the total repetitive elements were captured using in silico analysis, FISH generally revealed about 29 % of all the repetitive elements in the genome (Table 7.1). We consider the FISH signal likely to represent an overestimate because it only detects two-dimensional hybridization signals from the three-dimensional chromosome structure.

In *B. rapa*, CentBr1 and CentBr2 show about 85 % sequence similarity and are separately distributed to eight and two chromosome pairs, respectively (Lim et al. 2007). However, in *B. oleracea*, there is less distinct separation between

the chromosomal locations of CentBo1 and CentBo2, which show co-localization in several centromeres (Fig. 7.3g, h; Lim et al. 2007; Koo et al. 2011; Liu et al. 2014). This is consistent with there being little copy number difference between CentBo1 and CentBo2 compared to CentBr1 and CentBr2 in the 1x WGS survey (Fig. 7.2a). This also suggests that there was a different rate of homogenization of centromeric tandem repeats between *B. rapa* and *B. oleracea* genomes as well as among centromeres within each genome, as observed in some Brassicaceae species (Hall et al. 2005).

CentBr arrays are intermingled with a major centromeric LTR retrotransposon, CRB. Although CRB is common to the three basic *Brassica* lineage A, B, and C genomes, CentBr is present only in the A and C genomes (Lim et al. 2007). Additionally, the A genome-specific retrotransposon PCRBr hybridized to *B. rapa* chromosomes, but not to those of *B. oleracea* and *B. nigra* (Fig. 7.3i; Lim et al. 2007). It localized to four chromosomes with major heterochromatin blocks in *B. rapa*, which could explain the relatively high genomic proportion of PCRBr identified based on the 1x WGS survey (Fig. 7.2a, e). In addition, although Koo et al. (2011) reported three loci on three separate chromosomes for BrSTR, our data showed two loci on both arms of chromosome A05, with a major locus on the short arm, and two other very weak loci on two short chromosomes (Fig. 7.3c). This may be explained by the different sensitivity of FISH experiments, or different cytotypes used in the experiments, noting that *Brassica* genomes are highly dynamic and polymorphic (Koo et al. 2011). This was also demonstrated by Xiong and Pires (2011), who showed different numbers of 5S rDNA loci between different *B. rapa* accessions 'Chiifu' and the double haploid *B. rapa* IMB218. Taken together, the satellite repeat distribution in *B. rapa* further supports the general observation that centromeric and subtelomeric regions are havens for satellite repeats (Charlesworth et al. 1994).

Although in silico analysis identified more 45S rDNA than CentBr1, FISH showed that 45S rDNA was second to CentBr1 in terms of genomic abundance (Table 7.1). This suggests that some CentBr1 may not have been thoroughly captured despite their relative abundance; this is likely true for the other types of sequence as well considering that our analysis identified only half of the 40 % unassembled sequences.

There are more 5S and 45S rDNA loci in *B. rapa*, three and five, respectively, (Lim et al. 2005; Koo et al. 2011; Xiong and Pires 2011) compared with *B. oleracea*, which has only one and two (Liu et al. 2014). This underlies the higher genomic proportion of rDNA in *B. rapa* relative to that in *B. oleracea* (Fig. 7.2a, c).

A summary of the cytogenetic distribution of *B. rapa* repeats is presented in Fig. 7.3d. Genome composition of the eight major repeats studied in this study account for about half of the unassembled sequence based on mapping of 1x WGS reads, indicating that more repeats such as DNA transposons still remain hidden in the genome and could be further identified through a refined dnaLCW method (Table 7.2).

7.3 Functions and Evolutionary Implications of Repetitive Elements

The differential accumulation of repetitive elements, rather than gene sequences, is mainly responsible for the differences in C-value in plant genomes (Wei et al. 2013), a phenomenon commonly known as the C-value paradox (Hardman 1986; Pagel and Johnstone 1992; Macas et al. 2007). A growing amount of evidence supports the importance of these repeats in genome functions and evolution (Nowak 1994; Pardue and DeBaryshe 2003; Hall et al. 2005; Shapiro and von Sternberg 2005; Biémont and Vieira 2006; Wei et al. 2013).

Transposable elements (TE) are now known to possess characteristics that help shape the structure and evolution of genomes. They help regulate genes, defend genomes from retrotransposon proliferation and retrovirus invasion, cause

Table 7.2 Summary of different *B. rapa* and *B. oleracea* accessions used in this survey

	ID	Morphotype	Species	Sub species	Accession (cultivar)	Genome	WGS reads for repeat analysis	
							Amounts (Mbp)	Coverage (x)
1	Br-1	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	Chiifu	AA	2321.4	4.4
2	Br-2	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	Kenshin	AA	1498.9	2.8
3	Br-3	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	DF10C062	AA	1410.9	2.7
4	Br-4	Chinese cabbage	*B. rapa*	ssp. *pekinensis*	Z16	AA	1496.2	2.8
5	Br-5	Turnip Asian	*B. rapa*	ssp. *rapifera*	Yoya	AA	1495.9	2.8
6	Br-6	Rapini-Caixin	*B. rapa*	ssp. *parachinensis*	L58	AA	1492.5	2.8
7	Br-7	Pak Choi	*B. rapa*	ssp. *chinensis*	Suzhouqing	AA	1495.3	2.8
8	Br-8	Canola	*B. rapa*	ssp. *oleifera*	R-o-18	AA	1497.8	2.8
9	Br-9	Mizuna	*B. rapa*	ssp. *nipposinica*	Mizuna	AA	1497.5	2.8
10	Br-10	Turnip Europe	*B. rapa*	ssp. *rapifera*	Manchester	AA	1484.5	2.8
11	Br-11	Canola-rapid cycling	*B. rapa*	ssp. *oleifera*	L144	AA	1496.6	2.8
12	Bo-1	Cabbage	*B. oleracea*	ssp. *capitata*	C1176	CC	1541	2.2
13	Bo-2	Cabbage	*B. oleracea*	ssp. *capitata*	C1220	CC	1606.8	2.3

1–11 WGS of *B. rapa* accessions were kindly provided by Xiaowu Wang (Key Laboratory of Horticultural Crops Genetic Improvement of Ministry of Agriculture, Sino-Dutch Joint Lab of Horticultural Genomics Technology, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China). *12–13* WGS of *B. oleracea* was generated with support of a grant from the *Golden Seed Project* (*Center for Horticultural Seed Developmenet,* ***No. 213003-04-3-SB430***), *Ministry of Agriculture, Food and Rural Affairs* (*MAFRA*)

mutations, influence recombination rates, protect chromosomes through telomerase-independent fashion, and maintain centromeres, which play a significant role in chromosome segregation (Pardue and DeBaryshe 2003; Wolfgruber et al. 2009; Biémont 2010; Sarilar et al. 2011; Goodier et al. 2012; Sampath et al. 2013). In *Brassica*, MITE transposons preferentially accumulate near or inside of genic regions indicating these likely play roles in gene evolution (Sarilar et al. 2011; Sampath et al. 2013, 2014).

Most plant centromeric DNA is composed of 150–180 bp tandem repeats and centromere-specific retrotransposons (CR; Jiang et al. 2003; Lim et al. 2007; Talbert and Henikoff 2010; Neumann et al. 2011; Jiang 2013). The centromeric tandem repeat arrays can extend to several megabases and are often interrupted by CRs, which can also insert into other CRs, forming a complex nested pattern, and play a significant role in centromere function and evolution (Jiang et al. 2003; Lim et al. 2007; Wei et al. 2013). Association of these tandem repeats and CRs with modified histone H3 (CENH3), the hallmark of active centromeres, further indicates their active role in centromere function (Neumann et al. 2011; Jiang 2013).

Some evidence has been presented to help explain the rapid evolution of centromeric tandem arrays across different centromeres within a species. Unequal crossover, gene conversion, and repeat transposition have been invoked as key players in the homogenization and spread of repeats intra-chromosomally, between sister chromatids, between homologous chromosomes, and between non-homologous chromosomes (Walsh 1987; Charlesworth et al. 1994; Cohen et al. 2003; Hall et al. 2005). Unequal crossovers

usually result in higher-order repeat units consisting of more than one type of element and variation in lengths of arrays (Hall et al. 2005; Talbert and Henikoff 2010). Other mechanisms such as gene conversion and repeat transposition may amplify satellite arrays and cause their spread into nonhomologous chromosomes (Hall et al. 2005).

In *Brassica*, CentBr and CRB are major components of the centromere (Lim et al. 2007). The CRB is a common centromeric component of the A, B, and C genomes. However, the absence of CentBr hybridization in *B. nigra* (B genome) indicates that the B genome diverged from the A and C genomes earlier, supporting the 9 MYA divergence time for the B genome (Fig. 7.3i; Lim et al. 2007; Koo et al. 2011). This was further supported by the FISH results with the subtelomeric repeat STR, which also showed genome-specific evolution. The BnSTR tandem repeat from *B. nigra* (B genome) did not hybridize to either the A or C genome, and BrSTR from the A genome did not hybridize to either the B or C genome, although BoSTR from the C genome hybridized to both the A and C genomes (Koo et al. 2011). However, those tandem repeats (CentB and STR) show high sequence similarity between species (Lim et al. 2005, 2007; Koo et al. 2011), suggesting that the tandem repeats subsequently diverged in the A, B, and C genomes after speciation even though they shared a single origin in the ancient genome.

The pericentromeric retrotransposon PCRBr showed A-genome specificity (Fig. 7.3i). PCRBr is a gypsy type retrotransposon and is accumulated in several chromosomes of *B. rapa* suggesting that these retrotransposons were rapidly amplified in the A genome after divergence from the C genome during the last 4.6 MYA (Wang et al. 2011; Liu et al. 2014). Additionally, CentBr1 and CentBr2 have diverged in sequence and chromosomal distribution in *B. rapa* and *B. oleracea*. CentBr2 has both *Hin*dIII (AAGCTT) and *Sau*3AI (GATC) restriction sites while CentBr1 has lost the *Sau*3AI site (Koo et al. 2011). This phenomenon was also observed for maize CentC and Cen4 (Kato et al. 2004). Collectively, these results highlight the dynamic nature of the genomes in the genus *Brassica* and present examples of lineage- and genome-specific rapid evolution of centromeric components (Koo et al. 2011).

7.4 Conclusion and Perspectives

As exemplified by Macas et al. (2007) in *Pisum sativum*, survey of plant genomes using low-coverage NGS data proved to be an excellent tool for capturing the highly repetitive genomic sequences that are mostly left out during assembly. Our application of this technique to *Brassica* species further corroborated the usefulness of this approach. Characterizing the genomic abundance and distribution of these repetitive sequences is further facilitated when 1x WGS genomic survey is coupled with molecular cytogenetic techniques such as FISH.

Using this approach, independent analysis of repetitive elements from genome assembly data can provide huge amount of information regarding genome structure and evolution when comparative analyses are performed with closely and distantly related species. This approach may also promote our knowledge of plants with huge genomes such as *Allium* (Jakse et al. 2008). Repetitive sequences can be analyzed using low-coverage WGS before completion of genome sequencing and can provide guidance for complete elucidation of the genome structure of the target plant. This combined genome survey and cytogenetic approach will also be useful for evolutionary genomics analysis of plant families lacking available genome sequences by allowing comparison of the repetitive yet highly informative portions of their genomes, as exemplified by our work in ginseng (*Panax ginseng*; Choi et al. 2014).

Acknowledgments This research was carried out with the support by Golden Seed Project (Center for Horticultural Seed Development, No. 213003-04-3-SB430), Ministry of Agriculture, Food and Rural Affairs (MAFRA), Ministry of Oceans and Fisheries (MOF), Rural Development Administration (RDA) and Korea Forest Service (KFS), Republic of Korea.

References

Biémont C (2010) A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics 186(4):1085–1093

Biémont C, Vieira C (2006) Genetics: junk DNA as an evolutionary force. Nature 443(7111):521–524

Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491(7426):705–710

Cabo S, Carvalho A, Martin A, Lima-Brito J (2014) Structural rearrangements detected in newly-formed hexaploid tritordeum after three sequential FISH experiments with repetitive DNA sequences. J Genet 93(1):183–188

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371(6494):215–220

Choi HI, Waminal NE, Park HM, Kim NH, Choi BS et al (2014) Major repeat components covering one-third of the ginseng (*Panax ginseng* C.A. Meyer) genome and evidence for allotetraploidy. Plant J 77(6):906–916

Cohen S, Yacobi K, Segal D (2003) Extrachromosomal circular DNA of tandemly repeated genomic sequences in *Drosophila*. Genome Res 13(6A):1133–1145

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284(5757):601–603

Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 8(1):2–9

Goodier JL, Cheung LE, Kazazian HH Jr (2012) MOV10 RNA Helicase is a potent inhibitor of retrotransposition in cells. PLoS Genet 8(10): e1002941

Hall SE, Luo S, Hall AE, Preuss D (2005) Differential rates of local and global homogenization in centromere satellites from *Arabidopsis* relatives. Genetics 170(4):1913–1927

Hardman N (1986) Structure and function of repetitive DNA in eukaryotes. Biochem J 234(1):1–11

Harrison GE, Heslop-Harrison JS (1995) Centromeric repetitive DNA sequences in the genus *Brassica*. Theor Appl Genet 90(2):157–165

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16(10):1252–1261

Hershkovitz MA, Zimmer EA (1996) Conservation patterns in angiosperm rDNA ITS2 sequences. Nucl Acids Res 24(15):2857–2867

Huang S, Li R, Zhang Z, Li L, Gu X et al (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41(12):1275–1281

Jakse J, Meyer JD, Suzuki G, McCallum J, Cheung F et al (2008) Pilot sequencing of onion genomic DNA reveals fragments of transposable elements, low gene densities, and significant gene enrichment after methyl filtration. Mol Genet Genome 280(4):287–292

Jiang J (2013) Centromere evolution. In: Jiang J, Birchler JA (eds) Plant centromere biology. Wiley, Oxford, pp 159–168

Jiang J, Birchler JA, Parrott WA, Kelly Dawe R (2003) A molecular view of plant centromeres. Trends Plant Sci 8(12):570–575

Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G et al (2005) Evolution of genome size in Brassicaceae. Ann Bot 95(1):229–235

Kato A, Lamb JC, Birchler JA (2004) Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc Natl Acad Sci USA 101(37):13554–13559

Khrustaleva LI, Kik C (2001) Localization of single-copy T-DNA insertion in transgenic shallots (*Allium cepa*) by using ultra-sensitive FISH with tyramide signal amplification. Plant J 25(6):699–707

Kim K, Lee SC, Lee J, Lee HO, Choi BS, Joh JH, Kim NH, Park HS, Yang TJ (2015) Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within Panax ginseng species. Plos One (in press)

Koo DH, Plaha P, Lim YP, Hur Y, Bang JW (2004) A high-resolution karyotype of *Brassica rapa* ssp. *pekinensis* revealed by pachytene analysis and multicolor fluorescence in situ hybridization. Theor Appl Genet 109(7):1346–1352

Koo DH, Hong CP, Batley J, Chung YS, Edwards D et al (2011) Rapid divergence of repetitive DNAs in *Brassica* relatives. Genomics 97(3):173–185

Lamb JC, Danilova T, Bauer MJ, Meyer JM, Holland JJ et al (2007a) Single-gene detection and karyotyping using small-target fluorescence in situ hybridization on maize somatic chromosomes. Genetics 175(3):1047–1058

Lamb JC, Meyer JM, Corcoran B, Kato A, Han F et al (2007b) Distinct chromosomal distributions of highly repetitive sequences in maize. Chrom Res 15(1):33–49

Lee HR, Zhang W, Langdon T, Jin W, Yan H et al (2005) Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species. Proc Natl Acad Sci USA 102(33):11793–11798

Lim KY, Kovarik A, Matyasek R, Bezdek M, Lichtenstein CP et al (2000) Gene conversion of ribosomal DNA in *Nicotiana tabacum* is associated with undermethylated, decondensed and probably active gene units. Chromosoma 109(3):161–172

Lim KB, de Jong H, Yang TJ, Park JY, Kwon SJ et al (2005) Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassicarapa*. Mol Cells 19 (3):436–444

Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY et al (2007) Characterization of the centromere and peri-centromere retrotransposons in *Brassicarapa* and their distribution in related *Brassica* species. Plant J 49(2):173–183

Liu S, Liu Y, Yang X, Tong C, Edwards D et al (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun 5. doi:10.1038/ncomms4930

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 Genome 8:427

Martins C, Wasko AP (2004) Organization and evolution of 5S ribosomal DNA in the fish genome. In: Williams CR (ed) Focus in genome research. Nova Science Publishers, Hauppauge, pp 335–363

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46

Michael TP, Jackson S (2013) The first 50 plant genomes. Plant Genome 6(2)

Neumann P, Navratilova A, Koblizkova A, Kejnovsky E, Hribova E et al (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA 2(1):4

Nowak R (1994) Mining treasures from 'junk DNA'. Science 263(5147):608–610

Pagel M, Johnstone RA (1992) Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. Proc Roy Soc Lond Sr B: Biol Sci 249(1325):119–124

Pardue ML, DeBaryshe PG (2003) Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu Rev Genet 37:485–511

Park HM, JeonEJ WaminalNE, ShinKS KweonSJ et al (2010) Detectioin of transgenes in three genetically modified rice lines by fluorescence in situ hybridization. Genes Genomics 32:527–531

Roa F, Guerra M (2012) Distribution of 45S rDNA sites in chromosomes of plants: structural and evolutionary implications. BMC Evol Biol 12(1):225

Sampath P, Lee SC, Lee J, Izzah NK, Choi BS et al (2013) Characterization of a new high copy Stowaway family MITE, BRAMI-1 in *Brassica* genome. BMC Plant Biol 13:56

Sampath P, Murukarthick J, Izzah NK, Lee J, Choi HI et al (2014) Genome-wide comparative analysis of 20 miniature inverted-repeat transposable element families in *Brassica rapa* and *B. oleracea*. PLoS ONE 9 (4):e94499

Santos AP, Wegel E, Allen GC, Thompson WF, Stoger E et al (2006) In situ methods to localize transgenes and transcripts in interphase nuclei: a tool for transgenic plant research. Plant Methods 2:18

Sarilar V, Marmagne A, Brabant P, Joets J, Alix K (2011) BraSto, a Stowaway MITE from *Brassica*: recently active copies preferentially accumulate in the gene space. Plant Mol Biol 77(1–2):59–75

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19(R2): R227–R240

Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in de novo plant genome sequencing and assembly. Genome Biol 13(4):243

Shapiro JA, von Sternberg R (2005) Why repetitive DNA is essential to genome function. Biol Rev Camb Philos Soc 80(2):227–250

Suzuki G, Ogaki Y, Hokimoto N, Xiao L, Kikuchi-Taura A et al (2012) Random BAC FISH of monocot plants reveals differential distribution of repetitive DNA elements in small and large chromosome species. Plant Cell Rep 31(4):621–628

Swaminathan K, Varala K, Hudson ME (2007) 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

Talbert PB, Henikoff S (2010) Centromeres convert but don't cross. PLoS Biol 8(3):e1000326

van der Knaap E, Sanyal A, Jackson SA, Tanksley SD (2004) High-resolution fine mapping and fluorescence in situ hybridization analysis of sun, a locus controlling tomato fruit shape, reveals a region of the tomato genome prone to DNA rearrangements. Genetics 168 (4):2127–2140

Walsh JB (1987) Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics 115(3):553–567

Waminal N, Park HM, Ryu KB, Kim JH, Yang TJ et al (2012) Karyotype analysis of *Panax ginseng* C.A. Meyer, 1843 (Araliaceae) based on rDNA loci and DAPI band distribution. Comp Cytogenet 6(4):425–441

Waminal NE, Ryu KB, Park BR, Kim HH (2014) Phylogeny of cucurbitaceae species in korea based on 5S rDNA non-transcribed spacer. Genes Genomics 36(1):57–64

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li B, Liu B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, Lin C, Edwards D, Mu D, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema G, Tang H, Wang H, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Wang H, Jin H, Parkin IA, Batley J, Kim JS, Just J, Li J, Xu J, Deng J, Kim JA, Li J, Yu J, Meng J, Wang J, Min J, Poulain J, Wang J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao M, Jin M, Ramchiary N, Drou N, Berkman PJ, Cai Q, Huang Q, Li R, Tabata S, Cheng S, Zhang S, Zhang S, Huang S, Sato S, Sun S, Kwon SJ, Choi SR, Lee TH, Fan W, Zhao X, Tan X, Xu X, Wang Y,

Qiu Y, Yin Y, Li Y, Du Y, Liao Y, Lim Y, Narusaka Y, Wang Y, Wang Z, Li Z, Wang Z, Xiong Z, Zhang Z, Brassica rapa C, Genome Sequencing Project (2011) The genome of the mesopolyploid crop species Brassica rapa. Nat Genet 43 (10):1035–1039

Wei L, Xiao M, An Z, Ma B, Mason AS, Qian W, Li J, Fu D (2013) New insights into nested long terminal repeat retrotransposons in Brassica species. Mol Plant 6(2):470–482

Wolfgruber T K, Sharma A, Schneider KL, Albert PS, Koo D-H, Shi J, Gao Z, Han F, Lee H, Xu R, Allison J, Birchler JA, Jiang J, Dawe RK, Presting GG (2009) Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. PLoS Genet 5(11):e1000743

Xiong ZY, Pires JC (2011) Karyotype and identification of all homoeologous chromosomes of allopolyploid Brassica napus and Its diploid progenitors. Genetics 187(1):37–49

Yu W, Lamb JC, Han F, Birchler JA (2007) Cytological visualization of DNA transposons and their transposition pattern in somatic cells of maize. Genetics 175 (1):31–39