Development and genetic mapping of microsatellite markers from whole genome shotgun sequences in *Brassica oleracea*

Haitao Li · Xun Chen · Yuan Yang · Jinsong Xu · Jianxun Gu · Jie Fu · Xiaoju Qian · Shunchang Zhang · Jiangsheng Wu · Kede Liu

Received: 14 June 2010 / Accepted: 10 September 2010 / Published online: 26 September 2010

Abstract The availability of whole genome shotgun sequences (WGSs) in *Brassica oleracea* provides an unprecedented opportunity for development of microsatellite or simple sequence repeat (SSR) markers for genome analysis and genetic improvement in *Brassica* species. In this study, a total of 56,465 non-redundant SSRs were identified from the WGSs in *B. oleracea*, with dinucleotide repeats being the most abundant, followed by tri-, tetra- and pentanucleotide repeats. From these, 1,398 new SSR markers (designated as BoGMS) with repeat length ≥25 bp were developed and used to survey polymorphisms with a panel of six rapeseed varieties, which is the largest number of SSR markers developed for the C genome in a single study. Of these SSR markers, 752 (69.5%) showed polymorphism among the six varieties. Of these, 266 markers that showed clear scorable polymorphisms between *B. napus* varieties No. 2127 and ZY821 were integrated into an existing *B. napus* genetic linkage map. These new markers are preferentially distributed on the linkage groups in the C genome, and significantly increased the number of SSR markers in the C genome. These SSR markers will be very useful for gene mapping and marker-assisted selection of important agronomic traits in *Brassica* species.

Keywords Brassica oleracea · Whole genome shotgun sequences · Microsatellites · Simple sequence repeats · Brassica napus · Linkage map

Introduction

The genus *Brassica* includes a diverse range of species of global economic importance. Rapeseed (mainly *B. napus*) has become an important agricultural product during the past 30 years and is now the world's third leading source of both vegetable oil and oil meal (Snowdon et al. 2006). The *Brassica* crops not only provide oil, vegetables, fodder and condiment, but also are a valuable source for dietary fiber, vitamin C and other beneficial factors such as anticancer compounds (Fahey and Florens 1995). *Brassica* crop species are also popular for producing high-quality biodiesel due to their relatively low polyunsaturated and saturated fatty acids. The major crop types include three diploid species, *B. rapa* (AA, $2n = 20$), *B. nigra* (BB, $2n = 16$) and *B. oleracea* (CC, $2n = 18$), and also three amphidiploid species, *B. juncea* (AABB, $2n = 36$), *B. napus* (AACC, $2n = 38$) and *B. carinata* (BBCC, $2n = 34$). The genetic

Electronic supplementary material The online version of this article (doi:10.1007/s11032-010-9509-y) contains supplementary material, which is available to authorized users.

H. Li · X. Chen · Y. Yang · J. Xu · J. Gu ·
J. Fu · X. Qian · S. Zhang · J. Wu · K. Liu (✉)
National Key Laboratory of Crop Genetic Improvement
and National Center of Plant Gene Research (Wuhan),
Huazhong Agricultural University, Wuhan 430070, China
e-mail: kdliu@mail.hzau.edu.cn

relationship of the crop *Brassica* species has been well studied and is referred to as U's triangle (U N 1935). The long history of cultivation and artificial selection of the *Brassica* crop species has resulted in various crop types with an extraordinary level of intraspecific morphological phenotypic variation in their adaptation for cultivation under varied agroclimatic conditions (Liu 1983). For instance, well-established garden vegetables in *B. oleracea* comprise a number of morphologically diverse crops, including cabbage, broccoli and cauliflower.

Microsatellites, or simple sequence repeats (SSRs), are tandem repeats of 1–6 nucleotides present in all eukaryotic genomes. Because SSRs are easily detectable by PCR, amenable to high-throughput analysis, codominantly inherited, multi-allelic, highly polymorphic, abundant and evenly distributed in genomes, and also require only a small amount of DNA for analysis (Gupta and Varshney 2000), they have become valuable tools for genetic mapping (Padmaja et al. 2005; Saito et al. 2006), association mapping (Breseghello and Sorrells 2006), comparative mapping (Iniguez-Luy et al. 2009; Suwabe et al. 2006), diversity analysis (Chen et al. 2008, 2010; Hasan et al. 2006), quantitative trait locus analysis (Chen et al. 2007b; Delourme et al. 2006; Lombard and Delourme 2001; Long et al. 2007; Qiu et al. 2006; Zhao et al. 2006) and marker-assisted selection (Zhang et al. 2003; Zhou et al. 2003).

Extensive efforts have been made to develop SSR markers in *B. napus* and its two diploid progenitors, *B. rapa* and *B. oleracea*, through genomic library screening using probes containing repeated motifs and followed by DNA sequencing (Kresovich et al. 1995; Lowe et al. 2004; Plieske and Struss 2001; Suwabe et al. 2002; Szewc-McFadden et al. 1996; Uzunova and Ecke 1999; Varghese et al. 2000). However, as indicated at the *Brassica* Microsatellite Information Exchange (http//www.brassica.info/ssrinfo.htm), the total number of currently mapped SSR markers in *Brassica* is still too limited. Experimental methods to develop SSR markers are laborious, time-consuming and expensive. With the ever-increasing number of sequences in public databases, development of SSR markers using in silico approaches has become a practicable and inexpensive alternative for many crop species (Chen et al. 2007a; McCouch et al. 2002; Shoemaker et al. 2008; Shultz et al. 2007; Song et al. 2005).

Recently, several genome sequencing projects for *Brassica* species, especially for *B. rapa* and *B. oleracea*, have been conducted for various purposes (Ayele et al. 2005; Katari et al. 2005; Lim et al. 2006; Yang et al. 2005). *B. rapa* ssp. *pekinensis*, which has the smallest genome among *Brassica* species (Johnston et al. 2005), was selected as the representative for *Brassica* A-genome sequencing in the Multinational *Brassica* Genome Project (MBGP) (Yang et al. 2005; see also http://www.brassica.info) with the aim of identifying the complete sequence of this genome using a BAC-by-BAC strategy. A total of 200,017 BAC-end sequences (BESs) and 128,582 expressed sequence tags (ESTs) were generated from the *B. rapa* genome. At the same time, the Institute for Genomic Research, Cold Spring Harbor Laboratories and Washington University carried out the *B. oleracea* whole genome shotgun (WGS) sequencing project and obtained 595,321 random shotgun sequences. The sequences were aligned against the *Arabidopsis* genome sequence using BLAST, and the results were compared with the annotation to improve the *Arabidopsis* annotation by identifying genes that were missed in the previous annotation (Ayele et al. 2005; Katari et al. 2005). The availability of *Brassica* genomic sequence data offers an unprecedented opportunity to conduct a detailed comparative analysis of the relationships between the *Brassica* A and C genomes, and also between these two *Brassica* genomes and the complete genome of the model plant *Arabidopsis thaliana*. Using the WGSs in *B. oleracea* and BESs in *B. rapa*, the transposable elements (TEs) in the *B. oleracea* and *B. rapa* genomes were analyzed and compared with the model plant, *A. thaliana*, to understand the dynamics of TE-mediated genome expansion (Hong et al. 2006; Zhang and Wessler 2004). The availability of large quantity of *Brassica* DNA sequences also provides a vast resource with which to survey SSR motifs and marker development. Hong et al. (2007) estimated the abundance and distribution of SSRs in the *B. rapa* genome using the BES data, and compared it to that in *Arabidopsis*. Several preliminary studies had been conducted to develop SSR markers from ESTs and genome survey sequences (GSSs), including BESs and WGSs, for *B. rapa* (Choi et al. 2007; Kim et al. 2009; Ling et al. 2007), *B. oleracea* (Iniguez-Luy et al. 2008), *B. juncea* (Hopkins et al. 2007) and *B. napus* (Batley et al. 2007). However, the number of SSR markers with good quality available to the *Brassica* community was still too limited to carry out gene tagging and molecular marker-assisted breeding in *Brassica* species.

In this paper, we identified SSRs from the WGSs in *B. oleracea*. Our objectives were to assess the distribution of SSRs in the *B. oleracea* genome, develop a large number of SSR markers for the *Brassica* community, and integrate the newly developed markers into an existing *B. napus* genetic linkage map.

Materials and methods

Plant materials

A panel of six rapeseed varieties (S1, S2, M201 M202, No. 2127 and ZY821) that had been used as parents for three established mapping populations was used for SSR polymorphism screening. A double-haploid (DH) population with 88 individuals derived from the cross between No. 2127 and ZY821 (Cheng et al. 2009; Xiao et al. 2007) was used for linkage mapping. In addition, six cabbage cultivars, Jingfeng 1, Chunmian, Qiangli55, Xinfeng, Shuanghuan E and Xinglv Chunfeng, were used to test the amplification of those markers that did not produce any PCR products in *B. napus*.

Total DNA was isolated from young leaves of the six *B. napus* varieties, the six cabbage cultivars and the 88 DH lines using the cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle 1987). Total DNA was adjusted to a concentration of 25 ng/μl and used as templates for PCR amplification.

Source of sequences and SSR identification

The WGSs of *B. oleracea* were downloaded in FASTA format from GenBank using the Entrez query "*Brassica oleracea* and GSS (genome survey sequence)", and used for SSR mining and marker development. These WGSs contained mitochondrial and chloroplast genome sequences and were redundant, so a protocol was designed to remove mitochondrial, chloroplast and redundant sequences, as indicated in Fig. 1.

To reduce the task of removing the chloroplast and mitochondrial genome sequences and redundant sequences, we first identified and selected the SSR-containing sequences from the WGSs using SPUTNIK software (http://espressosoftware.com/pages/sputnik.jsp). The criteria for SSR selection were set at six repeats for dinucleotides and four repeats for tri-, tetra- and pentanucleotides. The resulting SSR-containing sequences were selected for subsequent analysis.

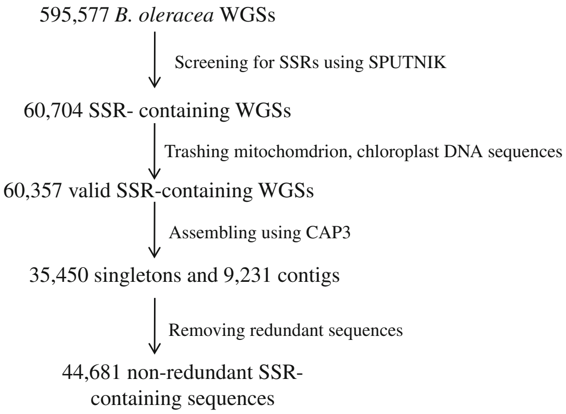

595,577 *B. oleracea* WGSs
Screening for SSRs using SPUTNIK
60,704 SSR- containing WGSs
Trashing mitochomdrion, chloroplast DNA sequences
60,357 valid SSR-containing WGSs
Assembling using CAP3
35,450 singletons and 9,231 contigs
Removing redundant sequences
44,681 non-redundant SSR-containing sequences

Fig. 1 The protocol for removing mitochondrial, chloroplast and redundant DNA sequences from the WGSs in *B. oleracea*

The second step was to remove the contaminated mitochondrial and chloroplast genome sequences. A total of 124 published complete genome sequences of chloroplast and 1,945 complete genome sequences of mitochondrion with accession numbers were obtained from GOBASE (http://gobase.bcm.umontreal.ca/). The SSR-containing sequences were then searched against the chloroplast and mitochondrial DNA sequences using Seqclean (http://compbio.dfci.harvard.edu/tgi/software/) to remove the matching sequences.

The third step was to remove the redundant SSR-containing sequences. The repetitive sequences were masked using RepeatMasker (http://www.repeatmasker.org/). The masked sequences were then assembled using CAP3 software (http://pbil.univ-lyon1.fr/cap3.php) with overlap length cutoff of 40 nucleotides and overlap percent identity cutoff of 95. After removing redundancy, the resulting consensus sequences of contigs and singletons were again mined for GSS-SSRs for further analysis.

Marker development and polymorphism detection

Primer3 software (Rozen and Skaletsky 2000) was used to design PCR primers flanking SSR sequences from the non-redundant SSR-containing sequences obtained previously. The primer length was between

18 and 23 nucleotides, with an optimum size of 20 nucleotides. The melting temperatures ranged from 50 to 70°C, with an optimum temperature of 55°C. The optimum GC content was set to 50%, with a minimum of 30% and a maximum of 70%. The predicted PCR products ranged from 100 to 400 bp. A total of 1,398 SSRs with repeat length of 25 bp or longer were selected for marker development. The newly developed GSS-SSR markers were designated as "BoGMS", representing *Brassica oleracea* genomic microsatellites. Primers were synthesized by GeneRay Biotech Company (Shanghai, China).

All SSR markers were subjected to polymorphism detection using a panel of six rapeseed varieties as described above. PCR amplification, product separation and staining were performed as described previously (Cheng et al. 2009).

Linkage analysis and map construction

The mapping population comprised 88 DH individuals derived from the cross between No. 2127 and ZY821 (Cheng et al. 2009; Xiao et al. 2007). The newly developed BoGMS markers were integrated into the existing *B. napus* genetic linkage map consisting of 244 SSR markers (Cheng et al. 2009) using JoinMap3.0 (Van Ooijen and Voorrips 2001). The threshold for goodness-of-fit was set to $\leq$5.0, with a recombination frequency of <0.4 and minimum logarithm of odds (LOD) scores of 2.0. All genetic distances were expressed in centimorgans (cM) as derived by the Kosambi function (Kosambi 1944). The segregation of each marker in the DH population was analyzed by a chi-square test for "goodness-of-fit" to an expected 1:1 ratio.

Results

Identification and characterization of GSS-SSRs in *B. oleracea*

A total of 595,577 *B. oleracea* shotgun reads with an average length of 677 bp (Ayele et al. 2005; Katari et al. 2005) were downloaded from GenBank. The total length of these sequences was 403 Mb, which was estimated to cover 0.58 of the genome, assuming that the size of the *B. oleracea* genome is approximate 696 Mb (Johnston et al. 2005). After downloading, searches were made using the SPUTNIK procedures for SSR motifs containing di-, tri-, tetra-, and pentanucleotide repeats in each sequence. A total of 60,704 SSR-containing sequences were identified, suggesting an average frequency of one SSR in every 6.6 kb in the WGSs.

However, it should be noted that the WGSs are redundant and were contaminated by the chloroplast and mitochondrial DNA sequences. Seqclean searches identified 347 chloroplast and mitochondrial DNA sequences. After removing the matching sequences, a total of 60,357 SSR-containing sequences were obtained. Cluster analysis with CAP3 revealed 35,450 singletons and 9,231 contigs. The contigs were composed of 24,907 sequences with an average of 2.70 sequences per contig. Most of the contigs (6,528) only contained two sequences, 2,616 contigs each contained three to ten sequences, and 84 contigs each contained 11–100 sequences. Three contigs each had more than 100 sequences, and one had as many as 322 sequences. After removing these redundant sequences, 44,681 non-redundant SSR-containing sequences were obtained. Analysis of SSR motifs in the non-redundant SSR-containing sequences identified 56,465 GSS-SSRs, with 5,563 contigs and 30,263 singletons each having one SSR. The remaining contigs and singletons each contained two or more SSRs. The SSR repeat lengths ranged from 12 to 432 bp, with dinucleotide SSRs showing the greatest range and highest average repeat length (Table 1). Among the repeats longer than 40 nucleotides, the dinucleotide repeats were much more frequent than the other repeat types.

Of the total GSS-SSRs, dinucleotide repeats (DNRs) (28,755, 50.9%) were the most abundant, followed by trinucleotide repeats (TNRs) (23,137, 41.0%), tetranucleotide repeats (2,874, 5.1%) and pentanucleotide repeats (1,699, 3.0%). Table 1 lists the numbers of the major SSR types identified from the *B. oleracea* WGSs. In DNRs, $(AT)_n$ was the most abundant repeat motif (28.8%), followed by $(AG)_n$ (19.3%) and $(AC)_n$ (2.8%). $(GC)_n$ repeats were very rare. All ten possible combinations of TNRs were observed in the GSS-SSRs. Among the TNRs, the $(AAG)_n$ motif was the most common (12.7%), followed by the $(AAT)_n$ (7.8%), $(ATC)_n$ (5.0%), $(AGG)_n$ (4.2%), $(AAC)_n$ (4.1%) and $(ACC)_n$ (2.9%) motifs. The $(ACT)_n$, $(ACG)_n$, $(CCG)_n$ and $(AGC)_n$ motifs were the least abundant (Table 1). Thirty-three combinations of tetranucleotide repeats and 78 combinations of pentanucleotide repeats were

Table 1 Distribution of major SSR types identified from the WGSs in *B. oleracea*

Motifs	Number[a] (%)	Range (bp)	Total length (bp)	Average length (bp)
Dinucleotide	28,755 (50.9)	12–432	556,052	19.34
AT	16,264 (28.8)	12–122	316,998	19.49
AG	10,877 (19.3)	12–432	214,864	19.75
AC	1,587 (2.8)	12–60	23,832	15.02
CG	27 (0.1)	12–20	358	13.26
Trinucleotide	23,137 (41.0)	12–198	327,147	14.14
AAG	7,153 (12.7)	12–93	101,631	14.21
AAT	4,398 (7.8)	12–198	64,185	14.59
ATC	2,813 (5.0)	12–66	40,044	14.24
AGG	2,376 (4.2)	12–183	34,482	14.51
AAC	2,300 (4.1)	12–39	31,254	13.59
ACC	1,624 (2.9)	12–78	22,233	13.69
AGC	900 (1.6)	12–33	12,351	13.72
CCG	687 (1.2)	12–30	8,892	12.94
ACG	496 (0.9)	12–33	6,771	13.65
ACT	390 (0.7)	12–51	5,304	13.60
Tetranucleotide	2,874 (5.1)	16–116	57,072	19.86
AAAT	1,092 (1.9)	16–44	20,160	18.46
AATT	366 (0.7)	16–68	10,584	28.92
AAAG	331 (0.6)	16–68	6,528	19.72
AAAC	262 (0.5)	16–28	4,640	17.71
Others	823 (1.5)	16–116	15,160	18.42
Pentanucleotide	1,699 (3.0)	20–410	38,840	22.86
AACCG	567 (1.0)	20–50	12,315	21.72
AAAAT	211 (0.4)	20–50	4,675	22.16
AAAAG	103 (0.2)	20–40	2,310	22.43
Others	818 (1.5)	20–410	19,540	23.89
Total	56,465 (100.0)	12–432	979,111	17.34

[a] Numbers in parentheses are percentage of motifs per total SSR

observed in the GSS-SSRs. The distributions of tetranucleotide and pentanucleotide repeats were uneven, with $(AAAT)_n$ being the most abundant tetranucleotide repeat (1.9%) and $(AACCG)_n$ being the most frequent pentanucleotide repeat (1.0%). All the other repeat types were very rare in the *B. oleracea* genome.

Microsatellite marker development and polymorphism detection

Microsatellites are usually classified into two categories based on the length of SSR tracts. Empirical data had indicated that Class I microsatellites (SSRs $\geq$ 20 nucleotides) are hypervariable (Temnykh et al. 2001; Yi et al. 2006) and Class II microsatellites (SSRs $\geq$ 12 but <20 nucleotides) are less variable. In order to increase the efficiency of marker development, 1,398 Class I microsatellites with repeat length $\geq$25 nucleotides were selected for primer design and are designated as "BoGMS" markers hereafter. They included 871 (62.3%) di-, 232 (16.6%) tri-, 59 (4.2%) tetra- and 84 (6%) pentanucleotide repeats and 152 (10.9%) composite microsatellites that consisted of more than two repeat motifs (Table 2).

All these 1,398 BoGMS markers were evaluated for successful PCR amplification and useful polymorphism by testing the genomic DNA of a panel of six oilseed rape cultivars for three established permanent mapping populations. Of these, 1,040 (74.4%) successfully amplified at least one fragment from the *B. napus* genome, 42 (3.0%) amplified smear PCR products, and 316 (22.6%) did not give

Table 2 Characteristics of *B. oleracea* SSR markers and efficiency of marker development

Motifs	No. of designed primers[a]	No. of amplified primers[b]	No. of polymorphic primers[c]
Dinucleotide	871(62.3)	671(77.0)	482(71.8)
AC	14 (1.0)	11 (78.6)	8 (72.7)
AG	469 (33.5)	371 (79.1)	276 (74.4)
AT	388 (27.8)	289 (74.5)	198 (68.5)
Trinucleotide	232 (16.6)	203 (87.5)	142 (70.0)
AAC	13 (0.9)	11 (84.6)	5 (45.5)
AAG	107 (7.7)	96 (89.7)	70 (72.9)
AAT	35 (2.5)	27 (77.1)	17 (63.0)
ACT	3 (0.2)	2 (66.7)	1 (50.0)
ATC	40 (2.9)	38 (95.0)	29 (76.3)
GC-rich TNRs[d]	34 (2.4)	29 (85.3)	20 (69.0)
Tetranucleotide	59 (4.2)	43 (72.9)	26 (60.5)
Pentanucleotide	84 (6.0)	61 (72.6)	37 (60.7)
Composite	152 (10.9)	104 (68.4)	65 (62.5)
Total	1,398 (100.0)	1,082 (77.4)	752 (69.5)

[a] Numbers in parentheses are percentage of designed primers per total primers

[b] Numbers in parentheses are percentage of successfully amplified primers per designed primers

[c] Numbers in parentheses are percentage of polymorphic markers per amplified primers

[d] GC-rich TNRs: the SSRs contain repeat motifs with two or three G/C

any amplification. Among the successful markers, 310 detected a single locus, 394 detected two loci in each cultivar, and 336 detected three or more loci, which is consistent with the allotetraploid nature of *B. napus* and genome triplication of its two progenitor species, *B. oleracea* and *B. rapa*. A subset of 752 (69.5%) BoGMS markers showed polymorphism among the six varieties. Information on the new markers is listed in Electronic Supplementary Material Table 1, which includes the GSS accession ID in GenBank, microsatellite ID, type of repeat motifs, number of repeats, expected amplicon size, forward and reverse primers, scorability, and polymorphism among the six *B. napus* varieties.

We evaluated the efficiency of marker development based on the success rate of PCR amplification and the level of polymorphism for each SSR motif (Table 2). The overall success rate of PCR amplification was 77.4%. DNR (71.8%) and TNR (70.0%) markers had the highest levels of polymorphism, followed by composite repeats (62.5%), pentanucleotide repeats (60.7%), and tetranucleotide repeats (60.5%). The $(ATC)_n$, $(AAG)_n$ and $(AG)_n$ motifs had the highest level of polymorphisms (76.3, 72.9 and 74.4%, respectively), while the $(AAC)_n$ repeat motif had the lowest level of polymorphism (45.5%). The other motifs had polymorphism levels in excess of 50%. There was no obvious correlation between polymorphism level and repeat length or repeat number of SSR markers in *B. napus*.

For the 316 markers that did not produce any amplification in the *B. napus* genome, we further tested their amplification in *B. oleracea*, the source genome of these markers, using six cabbage cultivars. Fifty markers successfully amplified least one fragment from the *B. oleracea* genome, suggesting that the targets of these markers might be deleted or the binding sites mutated in the *B. napus* genome due to extensive rearrangement including deletion or insertions or mutations after polyploidization.

Integration of newly developed SSR markers into existing linkage map

A total of 282 BoGMS markers that detected clear and scorable polymorphisms between No. 2127 and ZY821 were selected to survey the DH population. Among these, 266 markers, corresponding to 312 BoGMS loci, could be integrated into the existing *B. napus* genetic linkage map constructed previously using the same population (Cheng et al. 2009), and 16 BoGMS markers failed to be integrated into any linkage groups. These unmapped markers might be distributed to the extreme ends of chromosomes and could not be grouped with other markers. The majority of markers (247) detected a single locus, while 25, 9 and 1 markers detected 2, 3 and 4 loci, respectively. The integrated linkage map contained 312 BoGMS loci and 286 anchor loci from previous linkage maps (Chen et al. 2007b; Cheng et al. 2009; Piquemal et al. 2005; Suwabe et al. 2008), and covered a total length of 1,842.9 cM with an average interval of 3.08 cM between adjacent loci (Fig. 2). The 19 linkage groups were designated as A1–A10 and C1–C9 based on the nomenclature proposed at the MBGP Steering Committee meeting (http://www.brassica.info/resource/maps/lg-assignments.php).

Of the mapped BoGMS loci, 233 loci were located on the C genomes, while the other 79 loci were located on the A genome, indicating that these GSS-SSRs obviously skewed to the C genome. These BoGMS loci were randomly distributed all over the A and C chromosomes. The number of BoGMS loci varied dramatically from chromosome to chromosome. In the A genome, A5 and A9 had as many as 12 and 18 BoGMS loci, respectively, while A2 and A8 each had only two BoGMS loci. In the C genome, C3 had as many as 48 BoGMS loci, while C4 and C5 only had 11 and 9 loci, respectively. The other chromosomes in the C genome had 17–35 BoGMS loci. Most of the markers segregated with the expected 1:1 Mendelian ratio in the DH population. However, 192 (32.1%) loci including newly developed SSR markers and anchored markers deviated significantly ($P < 0.01$) from this ratio (Fig. 2). The biased loci were distributed unevenly: most of them were on A2, A3, A4, A6, C3, C4 and C9. The loci that biased to the same parent were clustered together, with 79 loci skewed towards ZY821 and 113 loci skewed towards No. 2127.

Discussion

The WGSs in *B. oleracea* represent a random sampling of the genome, which permitted an estimation of the abundance and distribution of SSRs and comparison with related species. In this study, SSRs were identified from a total of 403 Mb of whole genome shotgun sequences, which covers 0.58 of the genome of *B. oleracea*. In *B. oleracea*, $(AT)_n$ was the most abundant repeat motif, followed by $(AG)_n$ and $(AC)_n$, which is similar to *B. napus* and *A. thaliana* (Cheng et al. 2009; Lawson and Zhang 2006), while in the *B. rapa* and *Oryza sativa* genomes, $(AG)_n$ is the most abundant dinucleotide repeat, followed by $(AT)_n$ and $(AC)_n$ repeats (Hong et al. 2007; Lawson and Zhang 2006). In human and *Drosophila*, $(AC)_n$ is the most frequent DNR, followed by $(AT)_n$ and $(AG)_n$ (Katti et al. 2001). $(GC)_n$ is extremely rare in all eukaryotic genomes studied so far and this is also the case for *B. oleracea*. The frequency of SSR occurrence is one every 6.6 kb in the *B. oleracea* genome, which is much lower than that reported in the *B. rapa* (one every 4.7 kb, Hong et al. 2007) and *B. napus* genomes (one every 4.0 kb, Cheng et al. 2009). The genome size of *B. oleracea* (~696 Mb) is significantly larger than that of *B. rapa*, which is ~529 Mb (Johnston et al. 2005). Hence, the number of SSRs was estimated to be about 1.05×10^5 in *B. oleracea*, which is similar to that estimated in the *B. rapa* genome (1.12×10^5) (Hong et al. 2007).

In this study, the majority of the successful markers detected two or more loci, one or two derived from the A genome, the other one or two from the C genome, which is consistent with the allotetraploid nature of *B. napus* and genome triplication of its two progenitor species, *B. oleracea* and *B. rapa*. However, there were also a large number of markers (310) which detected a single locus in the *B. napus* genome. Most of the single-locus SSR markers (247) had been mapped on the *B. napus* genetic linkage map and were preferentially randomly distributed on the C genome in *B. napus*, suggesting that these single-locus markers were C genome-specific. Previous studies also reported that the majority of the primer pairs were able to produce specific products in *B. oleracea* whereas almost half failed to amplify in most *B. rapa* accessions (Mitchell et al. 1997). Earlier investigations have shown that SSR markers developed from *B. rapa* tend to amplify a product only in the A genome, rather than in the C genome in *B. napus* (Lowe et al. 2004; Suwabe et al. 2008), which has been proved to be an efficient strategy for construction of a high-density genetic linkage map of the A genome in *B. rapa*, *B. napus* and *B. juncea* (Kim et al. 2009; Lowe et al. 2004; Suwabe et al. 2008).

Extensive efforts had been taken to develop SSR markers for *Brassica* species (Batley et al. 2007; Cheng et al. 2009; Iniguez-Luy et al. 2008; Ling et al. 2007; Lowe et al. 2004; Suwabe et al. 2002), and several SSR-based genetic linkage maps have been

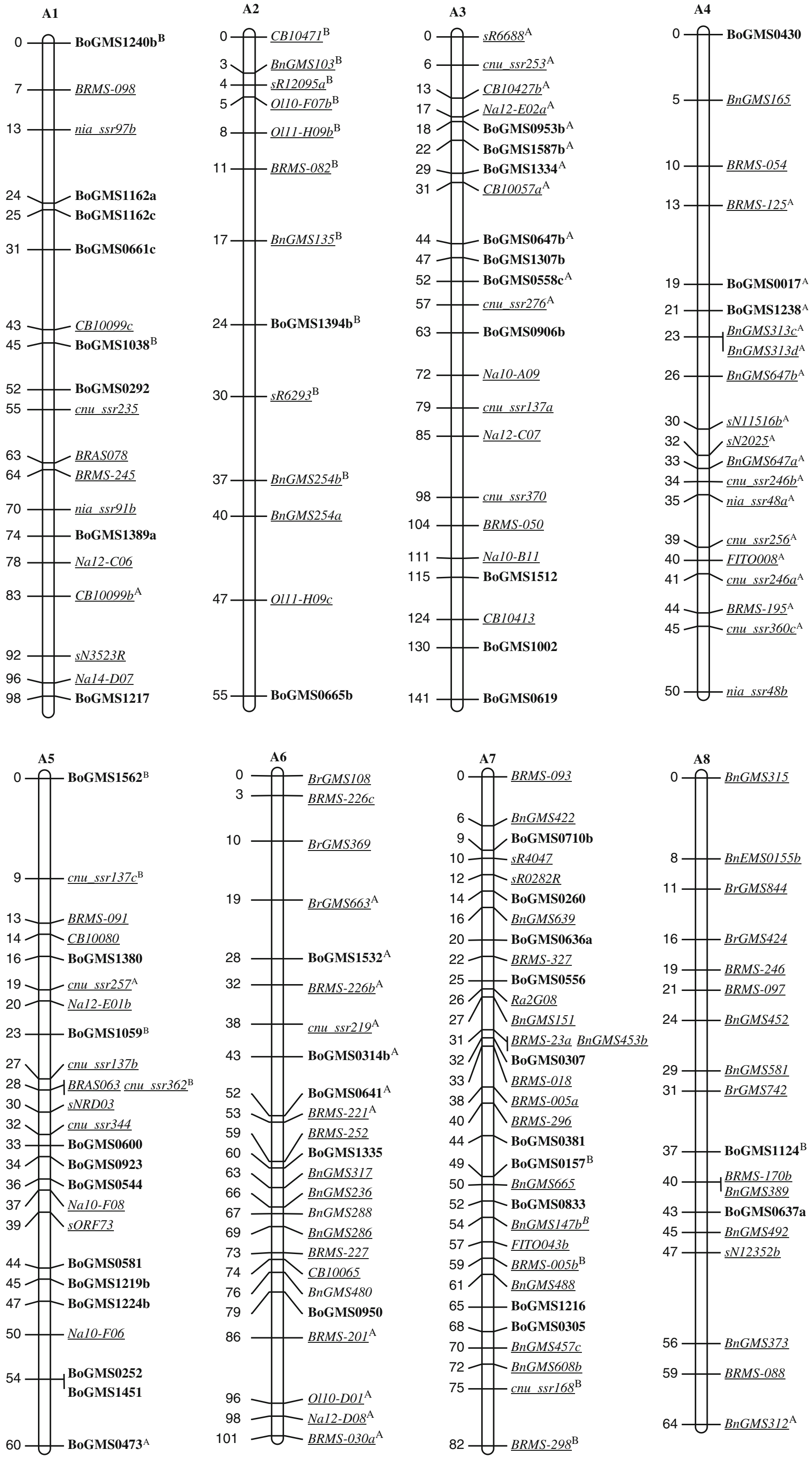


Fig. 2 The integrated genetic linkage map of *B. napus* constructed using 88 DH lines derived from the F1 hybrid between No. 2127 and ZY821. Markers indicated in *italics with underline* represent the anchor SSR markers from previous linkage maps (Chen et al. 2007b; Piquemal et al. 2005; Suwabe et al. 2008; Cheng et al. 2009). Markers indicated in **bold** represent new SSR markers developed in this study. Markers in *superscript letters* indicate segregation distortion; *superscript letter A* represents markers deviated toward No. 2127, and *superscript letter B* represents markers deviated toward ZY821

Fig. 2 continued

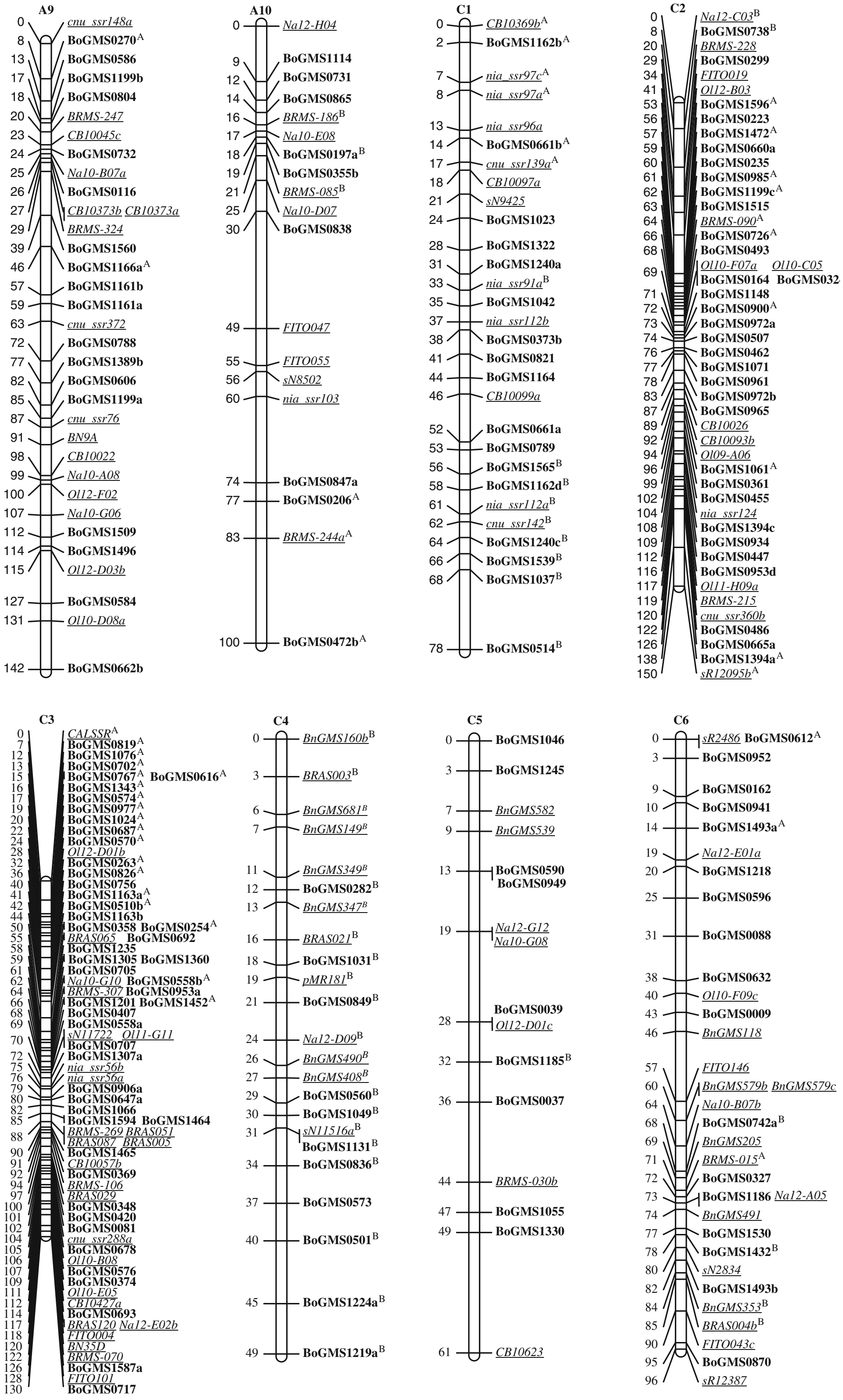
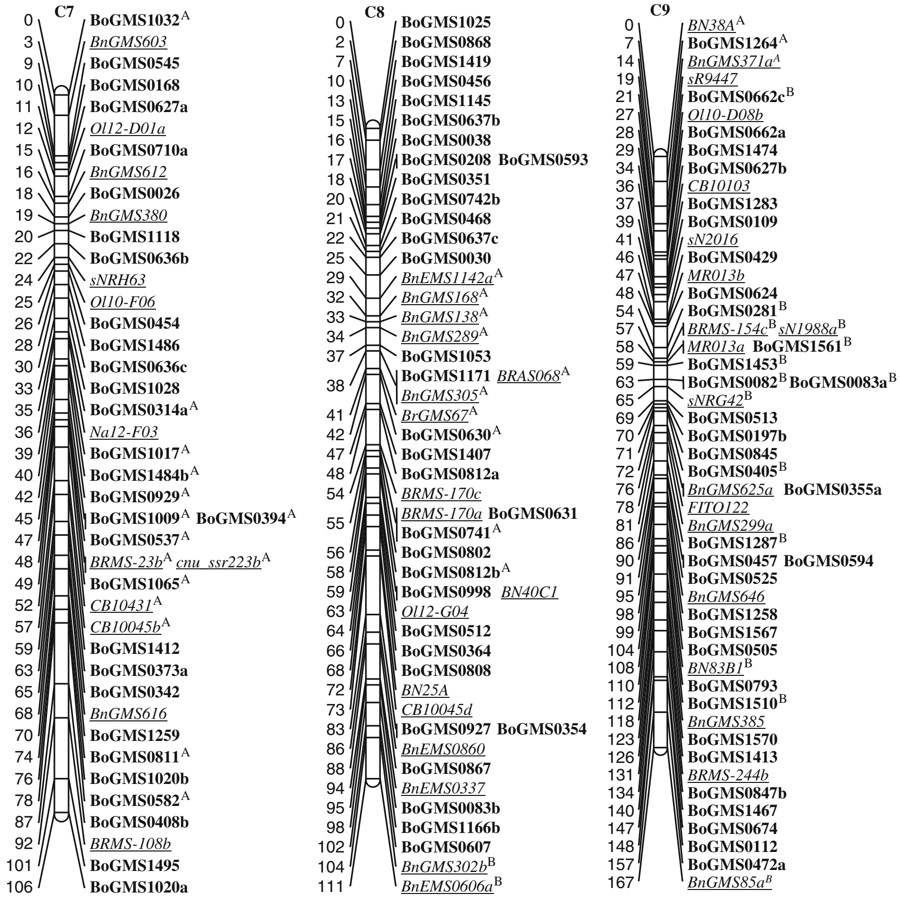


Fig. 2 continued



constructed for gene mapping and molecular marker-assisted breeding (Chen et al. 2007b; Cheng et al. 2009; Choi et al. 2007; Iniguez-Luy et al. 2009). In previous studies, more attention was paid to the *B. rapa* genome, especially after the initiation of the *B. rapa* whole genome sequencing project. Recently, a large number of SSR markers have been developed for *B. rapa* from SSR-enriched genomic libraries, genome survey sequences (GSSs) and sequenced BACs (Choi et al. 2007; Kim et al. 2009; Ling et al. 2007; Suwabe et al. 2002). Although 587 SSR markers had also been developed from the WGSs in *B. oleracea* (Iniguez-Luy et al. 2008), the number of SSR markers was still much less than that in *B. rapa*. Thus, the distribution of SSR markers was uneven between the A and C genomes. The lack of enough markers in the C genome will slow down the pace of gene mapping, cloning and marker-assisted selection for agronomically important traits in the C genome. In this study, 1,398 BoGMS markers were developed from the WGSs of *B. oleracea*, and 266 have been integrated into the existing *B. napus* genetic linkage map. To date, this is the largest number of SSR markers developed for the C genome in a single study. These SSR markers will be very useful for constructing a high-density genetic linkage map and facilitating gene mapping and marker-assisted selection of agronomic traits in *Brassica* species.

Acknowledgments This research was supported by the National Key Basic Research and Development Plan of China (2006CB101600) and the National Science Foundation of China (30623012).

References

Ayele M, Haas B, Kumar N, Wu H, Xiao Y, Van Aken S, Utterback T, Wortman J, White O, Town C (2005) Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in *Arabidopsis*. Genome Res 15:487–495

Batley J, Hopkins C, Cogan N, Hand M, Jewell E, Kaur J, Kaur S, Li X, Ling A, Love C (2007) Identification and characterization of simple sequence repeat markers from

Brassica napus expressed sequences. Mol Ecol Notes 7:886–889

Breseghello F, Sorrells M (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165–1177

Chen C, Yu Q, Hou S, Li Y, Eustice M, Skelton RL, Veatch O, Herdes RE, Diebold L, Saw J, Feng Y, Qian W, Bynum L, Wang L, Moore PH, Paull RE, Alam M, Ming R (2007a) Construction of a sequence-tagged high-density genetic map of papaya for comparative structural and evolutionary genomics in *Brassicales*. Genetics 177:2481–2491

Chen W, Zhang Y, Liu X, Chen B, Tu J, Tingdong F (2007b) Detection of QTL for six yield-related traits in oilseed rape (*Brassica napus*) using DH and immortalized F_2 populations. Theor Appl Genet 115:849–858

Chen S, Nelson M, Ghamkhar K, Fu T, Cowling W (2008) Divergent patterns of allelic diversity from similar origins: the case of oilseed rape (*Brassica napus* L.) in China and Australia. Genome 51:1–10

Chen S, Zou J, Cowling W, Meng J (2010) Allelic diversity in a novel gene pool of canola-quality *Brassica napus* enriched with alleles from *B. rapa* and *B. carinata*. Crop Pasture Sci 61:483–492

Cheng X, Xu J, Xia S, Gu J, Yang Y, Fu J, Qian X, Zhang S, Wu J, Liu K (2009) Development and genetic mapping of microsatellite markers from genome survey sequences in *Brassica napus*. Theor Appl Genet 118:1121–1131

Choi S, Teakle G, Plaha P, Kim J, Allender C, Beynon E, Piao Z, Soengas P, Han T, King G (2007) The reference genetic linkage map for the multinational *Brassica rapa* genome sequencing project. Theor Appl Genet 115:777–792

Delourme R, Falentin C, Huteau V, Clouet V, Horvais R, Gandon B, Specel S, Hanneton L, Dheu JE, Deschamps M, Margale E, Vincourt P, Renard M (2006) Genetic control of oil content in oilseed rape (*Brassica napus* L.). Theor Appl Genet 113:1331–1345

Doyle J, Doyle J (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19: 11–15

Fahey J, Florens HE (1995) The role of crucifers in cancer chemoprotection. In: Talalay P, Gustine DL (eds) Phytochemicals and health. American Society of Plant Physiologists, Rockville, MD, pp 87–93

Gupta PK, Varshney RK (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica 113:163–185

Hasan M, Seyis F, Badani A, Pons-Kühnemann J, Friedt W, Lühs W, Snowdon R (2006) Analysis of genetic diversity in the *Brassica napus* L. gene pool using SSR markers. Genet Resour Crop Evol 53:793–802

Hong C, Plaha P, Koo D, Yang T, Choi S, Lee Y, Uhm T, Bang J, Edwards D, Bancroft I (2006) A survey of the *Brassica rapa* genome by BAC-end sequence analysis and comparison with *Arabidopsis thaliana*. Mol Cells 22:300–307

Hong C, Piao Z, Kang T, Batley J, Yang T, Hur Y, Bhak J, Park B, Edwards D, Lim Y (2007) Genomic distribution of simple sequence repeats in *Brassica rapa*. Mol Cells 23:349–356

Hopkins CJ, Cogan NOI, Hand M, Jewell E, Kaur J, Li X, Lim GAC, Ling AE, Love C, Mountford H, Todorovic M, Vardy M, Spangenberg GC, Edwards D, Batley J (2007) Sixteen new simple sequence repeat markers from *Brassica juncea* expressed sequences and their cross-species amplification. Mol Ecol Notes 7:697–700

Iniguez-Luy F, Voort A, Osborn T (2008) Development of a set of public SSR markers derived from genomic sequence of a rapid cycling *Brassica oleracea* L. genotype. Theor Appl Genet 117:977–985

Iniguez-Luy F, Lukens L, Farnham M, Amasino R, Osborn T (2009) Development of public immortal mapping populations, molecular markers and linkage maps for rapid cycling *Brassica rapa* and *B. oleracea*. Theor Appl Genet 120:31–43

Johnston J, Pepper A, Hall A, Chen Z, Hodnett G, Drabek J, Lopez R, Price H (2005) Evolution of genome size in *Brassicaceae*. Ann Bot 95:229–235

Katari M, Balija V, Wilson R, Martienssen R, McCombie W (2005) Comparing low coverage random shotgun sequence data from *Brassica oleracea* and *Oryza sativa* genome sequence for their ability to add to the annotation of *Arabidopsis thaliana*. Genome Res 15:496–504

Katti M, Ranjekar P, Gupta V (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167

Kim H, Choi S, Bae J, Hong C, Lee S, Hossain M, Van Nguyen D, Jin M, Park B, Bang J (2009) Sequenced BAC anchored reference genetic map that reconciles the ten individual chromosomes of *Brassica rapa*. BMC Genomics 10:432

Kosambi DD (1944) The estimation of map distance from recombination values. Ann Eugen 12:172–175

Kresovich S, Szewc-McFadden A, Bliek S, McFerson J (1995) Abundance and characterization of simple-sequence repeats (SSRs) isolated from a size-fractionated genomic library of *Brassica napus* L. (rapeseed). Theor Appl Genet 91:206–211

Lawson M, Zhang L (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. Genome Biol 7:R14

Lim Y, Plaha P, Choi S, Uhm T, Hong C, Bang J, Hur Y (2006) Toward unraveling the structure of *Brassica rapa* genome. Physiol Plant 126:585–591

Ling A, Kaur J, Burgess B, Hand M, Hopkins C, Li X, Love C, Vardy M, Walkiewicz M, Spangenberg G (2007) Characterization of simple sequence repeat markers derived in silico from *Brassica rapa* bacterial artificial chromosome sequences and their application in *Brassica napus*. Mol Ecol Notes 7:273–277

Liu HL (1983) Studies on the breeding of yellow-seeded *Brassica napus*. Proceedings of the 6th International Rapeseed Congress, Paris, France, pp 637–641

Lombard V, Delourme R (2001) A consensus linkage map for rapeseed (*Brassica napus* L.): construction and integration of three individual maps from DH populations. Theor Appl Genet 103:491–507

Long Y, Shi J, Qiu D, Li R, Zhang C, Wang J, Hou J, Zhao J, Shi L, Park B-S, Choi SR, Lim YP, Meng J (2007) Flowering time quantitative trait loci analysis of oilseed *Brassica* in multiple environments and genomewide alignment with *Arabidopsis*. Genetics 177:2433–2444

Lowe A, Moule C, Trick M, Edwards K (2004) Efficient large-scale development of microsatellites for marker and mapping applications in *Brassica* crop species. Theor Appl Genet 108:1103–1112

McCouch SR, Teytelman L, Xu YB, Lobos KB, Clare K, Walton M, Fu BY, Maghirang R, Li ZK, Xing YZ, Zhang QF, Kono I, Yano M, Fjellstrom R, DeClerck G, Schneider D, Cartinhour S, Ware D, Stein L (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). DNA Res 9:199–207

Mitchell S, Kresovich S, Jester C, Hernandez C, Szewc-McFadden A (1997) Application of multiplex PCR and flourescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. Crop Sci 37:617–624

Padmaja KL, Arumugam N, Gupta V, Mukhopadhyay A, Sodhi YS, Pental D, Pradhan AK (2005) Mapping and tagging of seed coat colour and the identification of microsatellite markers for marker-assisted manipulation of the trait in *Brassica juncea*. Theor Appl Genet 111:8–14

Piquemal J, Cinquin E, Couton F, Rondeau C, Seignoret E, Doucet I, Perret D, Villeger M, Vincourt P, Blanchard P (2005) Construction of an oilseed rape (*Brassica napus* L.) genetic map with SSR markers. Theor Appl Genet 111:1514–1523

Plieske J, Struss D (2001) Microsatellite markers for genome analysis in *Brassica*. I. Development in *Brassica napus* and abundance in *Brassicaceae* species. Theor Appl Genet 102: 689–694

Qiu D, Morgan C, Shi J, Long Y, Liu J, Li R, Zhuang X, Wang Y, Tan X, Dietrich E, Weihmann T, Everett C, Vanstraelen S, Beckett P, Fraser F, Trick M, Barnes S, Wilmer J, Schmidt R, Li J, Li D, Meng J, Bancroft I (2006) A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content. Theor Appl Genet 114:67–80

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386

Saito M, Kubo N, Matsumoto S, Suwabe K, Tsukada M, Hirai M (2006) Fine mapping of the clubroot resistance gene, *Crr3*, in *Brassica rapa*. Theor Appl Genet 114:81–91

Shoemaker RC, Grant D, Olson T, Warren WC, Wing R, Yu Y, Kim H, Cregan P, Joseph B, Futrell-Griggs M, Will N, Davito J, Walker J, Wallis J, Kremitski C, Scheer D, Clifton SW, Graves T, Nguyen H, Wu X, Luo M, Dvorak J, Nelson R, Cannon S, Tomkins J, Schmutz J, Stacey G, Jackson S (2008) Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. Genome 51:294–302

Shultz J, Kazi S, Bashir R, Afzal J, Lightfoot D (2007) The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. Theor Appl Genet 114:1081–1090

Snowdon R, Lühs W, Friedt W (2006) Oilseed rape. In: Kole C (ed) Genome mapping and molecular breeding, vol 2: oilseeds. Springer Verlag, Heidelberg, pp 55–114

Song Q, Shi J, Singh S, Fickus E, Costa J, Lewis J, Gill B, Ward R, Cregan P (2005) Development and mapping of microsatellite (SSR) markers in wheat. Theor Appl Genet 110:550–560

Suwabe K, Iketani H, Nunome T, Kage T, Hirai M (2002) Isolation and characterization of microsatellites in *Brassica rapa* L. Theor Appl Genet 104:1092–1098

Suwabe K, Tsukazaki H, Iketani H, Hatakeyama K, Kondo M, Fujimura M, Nunome T, Fukuoka H, Hirai M, Matsumoto S (2006) Simple sequence repeat-based comparative genomics between *Brassica rapa* and *Arabidopsis thaliana*: the genetic origin of clubroot resistance. Genetics 173:309–319

Suwabe K, Morgan C, Bancroft I (2008) Integration of *Brassica* A genome genetic linkage map between *Brassica napus* and *B. rapa*. Genome 51:169–176

Szewc-McFadden A, Kresovich S, Bliek S, Mitchell S, McFerson J (1996) Identification of polymorphic, conserved simple sequence repeats (SSRs) in cultivated *Brassica* species. Theor Appl Genet 93:534–538

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441–1452

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

Uzunova M, Ecke W (1999) Abundance, polymorphism and genetic mapping of microsatellites in oilseed rape (*Brassica napus* L.). Plant Breed 118:323–326

Van Ooijen JW, Voorrips RE (2001) JoinMap®3.0. Software for the calculation of genetic linkage maps. Plant Research International, Wageningen

Varghese J, Rudolph B, Uzunova M, Ecke W (2000) Use of 5′-anchored primers for the enhanced recovery of specific microsatellite markers in *Brassica napus* L. Theor Appl Genet 101:115–119

Xiao S, Xu J, Li Y, Zhang L, Shi S, Wu J, Liu K (2007) Generation and mapping of SCAR and CAPS markers linked to the seed coat color gene in *Brassica napus* using a genome-walking technique. Genome 50:611–618

Yang T, Kim J, Lim K, Kwon S, Kim J, Jin M, Park J, Lim M, Kim H, Kim S (2005) The Korea *Brassica* genome project: a glimpse of the *Brassica* genome based on comparative genome analysis with *Arabidopsis*. Comp Funct Genom 6:138–146

Yi G, Lee J, Lee S, Choi D, Kim B (2006) Exploitation of pepper EST–SSRs and an SSR-based linkage map. Theor Appl Genet 114:113–130

Zhang X, Wessler S (2004) 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

Zhang T, Yuan Y, Yu J, Guo W, Kohel R (2003) Molecular tagging of a major QTL for fiber strength in Upland cotton and its marker-assisted selection. Theor Appl Genet 106:262–268

Zhao JY, Becker HC, Zhang DQ, Zhang YF, Ecke W (2006) Conditional QTL mapping of oil content in rapeseed with respect to protein content and traits related to plant development and grain yield. Theor Appl Genet 113: 33–38

Zhou W, Kolb F, Bai G, Domier L, Boze L, Smith N (2003) Validation of a major QTL for scab resistance with SSR markers and use of marker-assisted selection in wheat. Plant Breed 122:40–46