Characterization and evolutionary analysis of *Brassica* species-diverged sequences containing simple repeat units

Lijuan Wei[1], Meili Xiao[1], Annaliese S. Mason[2], Bi Ma[1], Kun Lu[1], Jiana Li[1], Link Katrin[1], Donghui Fu[3,*]

[1] College of Agronomy and Biotechnology, Southwest University. Engineering Research Center of South Upland Agriculture of Ministry of Education, P. R. China. Chongqing 400716, China.

[2] Centre for Integrative Legume Research and School of Agriculture and Food Sciences, The University of Queensland, Brisbane 4072, Australia

[3] Key Laboratory of Crop Physiology, Ecology and Genetic Breeding, Ministry of Education, Agronomy College, Jiangxi Agricultural University, Nanchang 330045, China.

[*] Corresponding author

E-mail: fudhui@163.com

Telephone: +086-0791-83813142;

Fax: +086-0791-83813185

Short running title: *Brassica*-specific sequences containing SSR

Abstract

Brassica species, *B. napus* (canola), *B. rapa* and *B. oleracea*, are important sources of nutritionally valuable vegetable oil and protein-rich meal for animals and humans. Sequencing of the model plant *Arabidopsis thaliana*, has opened the way for investigations into the complex structure of the *Brassica* genomes, offering important insight into their evolution and composition. We use this sequence information for the characterization and functional analysis of SSR sequences that have diverged between the *Brassica* species. A total of 56 species-diverged sequences containing simple repeat units (SDS-SSR) of *B. napus* and its diploid progenitor species *B. rapa* and *B. oleracea* were isolated and characterized. Of these, 40 sequences showed homology with other *Brassica* sequences. Using the SSR Locator software, only 23 sequences were found to have SSRs, possibly due to the loss of SSR units in the process of species divergence. Sequence alignments with *Arabidopsis thaliana* revealed that these species-diverged SSR sequences were responsible for *Brassica* divergence for differences between *Brassica* species in several genomic regions. Six active genes related to transferase, protein, transcription factor and retroelements were found in the SDS-SSRs. These results will further improve our understanding of the characteristics of species-diverged SSR fragments and their contribution to genome differentiation.

Key words： *Brassica*, species-diverged sequences, SSR

Introduction

Genomic DNA sequences of related species are generally highly similar but the abundance and chromosomal distribution of homologous sequences varies. Specific genomic fragments may account for large differences between species (Springer et al., 2009). For instance, the genetic difference between humans and chimpanzees is only 5%, but the ecological divergence of the two species is very high (Britten, 2002). Sequences that have diverged between closely related genomes are the fragments that are characteristic of a species, and it is important to investigate such genomic fragments in order to obtain a better understanding of species divergence.

Genome-diverged fragments are mostly generated through sequence mutation and differentiation, chromosomal translocation or transposition, or chromosomal substitution (Anamthawat-Jonsson and Heslop-Harrison, 1993), either under special circumstances or artificial selection. These fragments are not only important for investigation of the origin, divergence and evolution of a species, but also facilitate understanding of the organization of and changes occurring in a genome (Gao et al., 2000). Sequences specific to *Arabidopsis thaliana* genome play an important role in investigation of functional genomics, including transcript profiling and reverse genetics (Hilson et al., 2004). Specific sequences can be used to identify related species that are not easily distinguished because of high morphological similarities (Aveskamp et al., 2009). Species-specific sequences can also affect the conformational transition of proteins (Li et al., 2011) as well as their superhelical arrangement (Hothorn et al., 2011). Genome-specific sequences have also been used to

improve the detection of DNA motifs (Marchal et al., 2003) and optimal detection of operons (Dam et al., 2007). Repetitive specific sequences have been used to analyze interspecies crosses in rice (Kang and Kang, 2008) and identify addition lines in sugar beet (Gao et al., 2000).

Most genome sequences consist of tandem, inverted and dispersed sequence repeats. Repetitive fragments occupy more than 75% of the wheat genome and nearly 90% of the rye genome (Flavell et al., 1974). Repetitive elements play a significant role in genome and gene evolution and are a major source of species-specific diverged DNA fragments during duplication, deletion and transposition events (Gao et al., 2000; Chung et al., 2007). Chromosome- or locus-specific repeats are universal in the human genome (Lisch et al., 1995). In cereal genomes repeats can contribute 15% to 45% (Flavell et al., 1981). In barley, the repeated species-specific sequences are A/T-rich nucleotides of widely varying lengths (Junghans and Metzlaff, 1988). Species-specific repetitive sequences are also abundant in microbial genomes (Remm et al., 2009). The number of species-specific repeats (>100 bp) in yeast (*Candida glabrata* CBS138) is 137, and the maximum number of copies per repeat is 36. In *Brassica*, the repetitive DNA sequence BAC pBNBH35 is specific for eight chromosomes of the B genome, and located in the centromeric region of each B-genome chromosome (Schelfhout et al., 2004) but is absent in the A and C genomes. The *Brassica* C genome-specific transposable element Bot1 involved in the S locus plays an important role in the divergence between the *Brassica* A and C genomes (Alix et al., 2008). In addition, species-specific genes are also related to differences between species (Smith and

Rausher, 2011). Hence, species-specific genes, like species-specific repetitive sequences, drive the divergence and evolution of a species.

A simple sequence repeat (SSR) or so-called "microsatellite" is a type of repetitive DNA region consisting of repeats of a 1 to 6 bp motif, and is also sometimes called a short tandem repeat sequence. It may exist in any coding or non-coding region in both eukaryotes and prokaryotes. Microsatellites are considered to be one of the most variable elements in a genome due to the high level of variation in microsatellite sequence motifs, lengths, and repeat numbers. The mutation rate of SSRs range from 10^{-2} to 10^{-6} events per locus per generation, which is significantly higher than that of any other region of a genome (Sia et al., 2000) due to factors such as allopolyploidization (Tang et al., 2009), stress (Nevo et al., 2005), and radiation (Eggert et al., 2009). Microsatellites comprise a force to drive divergence of species-specific sequences.

Brassica crops are important sources of nutritional and valuable vegetable oil and protein-rich meals for both animals and humans. They are also becoming increasingly important as bioenergy sources and industry materials for lubricants and surfactants (Li et al., 2010). The *Brassica* genus includes three diploid and three amphidiploid species. Genomes A, B, and C are derived from a common allohexaploid *Brassica* ancestor (Röbbelen, 1960). *Brassica napus* L. (AACC, $2n = 4x = 38$) is a broad-acre oil crop derived from interspecies hybridization between *Brassica rapa* (AA, $2n = 2x = 20$) and *Brassica oleracea* (CC, $2n = 2x = 18$) (U, 1935). The identification of species-diverged DNA fragments is very important in *Brassica* research, particularly in identifying the authenticity

of hybrids and in distinguishing specific chromosomes (La Mura et al., 2010; Pankin and Khavkin, 2011). However, studies on SDS-SSRs are still rare. The impact of these fragments on the divergence and speciation of *B. napus* and its diploid parental species, *B. rapa* and *B. oleracea* is also still unknown. The present study aimed to isolate and characterize these repetitive fragments in *B. oleracea*, *B. rapa*, and *B. napus*, to analyze differences between SDS-SSRs and their homologous fragments; and to determine the putative function if any of these regions. The results are expected to contribute to an understanding of the origin, divergence, and putative functions of species-diverged sequences.

Materials and methods

Primer design

B. rapa genome sequences were downloaded from public databases (*Brassica* database and NCBI website), and SSRs were identified using the SSR Locator software (da Maia et al., 2008). Specific primer pairs were designed with default values using Primer Premier 5.0 (Premier Biosoft International, Palo Alto, CA, USA). The annealing temperature varied from 55 °C to 65 °C, with an optimal value of 60 °C. The SSR loci with SSR primers were used as queries in BlastN analysis against the *B. rapa* database. Hits with one unique hit per SSR locus were selected. The selected SSRs were then used as queries in the BlastN analysis against the NCBI (http://www.ncbi.nlm.nih.gov/) *B. oleracea* and *B. napus* databases. To improve the screening frequency and specificity of the primer pairs, hits with

only one unique hit per SSR locus in the three species were retained. The corresponding SSR primer pairs were used in the subsequent molecular marker assay.

Material selection and DNA pool preparation

A total of 153 accessions (37 *B. oleracea*, 59 *B. napus*, and 57 *B. rapa* accessions) were selected for species-diverged fragment screening, selecting accessions for maximum geographic and genetic diversity. The nomenclature, code, origin, and other basic information for the selected accessions are listed in Supplementary Table 1. Total genomic DNA for all plants was extracted using the CTAB method (Rogers and Bendich, 1985). To acquire the specific fragments in the three *Brassica* species, the DNA of all accessions of one species was proportionately mixed to construct a DNA pool of the species. The final DNA concentration of each DNA pool was adjusted to 50 ng/μl and was used as for DNA templates in polymerase chain reaction (PCR) analysis.

Acquisition and sequencing of species-diverged bands

The putatively species-diverged bands were amplified with the aforementioned primers and the three species DNA pools as the templates. All PCR amplifications were performed using a 50 μl reaction mixture containing high-fidelity Taq polymerase (Sangon Biotech (Shanghai) Co., Ltd.). The “touchdown” PCR amplification program used was as follows: initial denaturation at 94 °C for 5 min; 5 cycles of 30 s at 94 °C, 45 s at 61 °C with a 1 °C

decrease in the annealing temperature per cycle, and 1 min at 72 °C; 30 cycles of 30 s at 94 °C, 45 s at 57 °C, and 1 min at 72 °C; and a final extension at 72 °C for 10 min.

Amplification products were identified using 6% polyacrylamide gel electrophoresis. Only bands existing in one species were selected and gel extraction was performed. Then PCR was again performed to efficiently remove the non-target bands. PCR products were separated and purified on a 1.5% agarose gel. Target bands were ligated into the pMD19-T plasmid vector (TaKaRa, Japan), transformed into *Escherichia coli* and cultivated overnight in a lysogeny broth (LB) medium at 16 °C. Two positive clones per band were selected and sequenced.

Diversity analysis

The species-diverged sequences were used for alignment analysis by Clustal W in the BioEdit software and the same sequences and sequences of unexpected length compared to the gel electrophoresis were removed (Hall, 1999). One quality sequence out of the two positive clones per band was selected for further analysis. Species-diverged sequences were used as queries against the *B. napus* (http://www.ncbi.nlm.nih.gov/), *B. rapa* (http://brassicadb.org/brad/), and *B. oleracea* databases (http://brassicadb.org/brad/) to identify all homologous fragments. Matching homologous sequences were selected based on an *E* value lower than e^{-20} and a rate of homology (hit length/query length) higher than 45%. Each species-diverged sequence and its hits were used to calculate the nucleotide diversity, the number and length of insertion and deletion (InDel) events, and Tajima's D

test with default values using DnaSP v5 (Librado and Rozas, 2009). SSRs with mono-, di-, tri-, tetra-, penta-, hexa-, hepta-, octa-, nona-, or decanucleotide motifs were identified using the SSR Locator software (da Maia et al., 2008).

In silico mapping of species-specific fragments

The five chromosomes of *A. thaliana* were classified into 24 conserved building blocks (labeled A to X) (Schranz et al., 2006). The sequence accessions of the corresponding block boundary loci and related sequences were downloaded from the *A. thaliana* database (http://www.arabidopsis.org/). To map the species-diverged sequences in the *A. thaliana* genome, a local BlastN (http://www.ncbi.nlm.nih.gov/BLAST/) analysis was performed against the *A. thaliana* block sequences with *E* values lower than $1.0E^{-5}$.

Functional annotation

Species-diverged sequences that contained putative genes were annotated using Blast2GO (http://www.hindawi.com/journals/ijpg/2008/619832/) (Conesa and Gotz, 2008). The annotation included the putative cellular compositions, molecular functions, and biological processes.

Results

Primer design and species-diverged SSR sequences screening

All *B. rapa* genome sequences (in total 283 Mbp) were downloaded from the *Brassica*

database (http://brassicadb.org/brad/). A total of 14,713 SSR loci were detected by the SSR Locator software. Through homologous alignment and screening against the *B. napus* database (http://www.ncbi.nlm.nih.gov/) and the *Brassica* database (http://brassicadb.org/brad/), 14,327 SSR loci with more than one hit were rejected. Thus, a total of 386 putative unique SSR loci in *B. rapa* (absent in *B. napus*, and *B. oleracea*) were obtained to design SSR primers. To quickly identify the species-diverged SSR fragments between the three species, three DNA pools consisting of mixed DNAs from 37 *B. oleracea*, 57 *B. rapa* and 59 *B. napus* accessions were constructed.

Though we selected unique loci, some bands were still detected simultaneously in all three *Brassica* species. A total of 108 bands amplified by 83 primer pairs were obtained. Finally, 208 clones from these fragments (average two clones per locus) were selected and sequenced. We removed redundant and truncated sequences that were identical or without expected lengths as identified by ClustalW inside the BioEdit software, leaving 56 fragments amplified by 37 SSR primers; 19 (33.9%) in *B. oleracea*, 24 (42.9%) in *B. napus* and 13 (23.2%) in *B. rapa* (Supplementary Table 2). Five representative species-diverged SSR fragments are presented in Figure 1.

Divergence analysis of species-diverged sequences

The species-diverged fragments were used in the BlastN analysis against the *B. napus*, *B. rapa*, and *B. oleracea* databases in NCBI. The homologies of 40 sequences of 56 species-diverged fragments were determined based on an *E* value of $\leq e^{-20}$ and a homology

ratio (hit length/query length) of ≥ 45% (Supplementary Table 3). Another 16 sequences hardly had any similarity with any other fragment in the three species, and thus may be novel unique sequences.

These sets of homologous fragments were used to estimate the degree and determine the characteristics of divergence of the species-diverged sequences. All parameters and characteristics, including nucleotide diversity and InDel polymorphisms, are listed in Supplementary Table 4 (except for E1 and E20). The nucleotide diversity (*Pi*) of most sets of sequences centered on 0.30 to 0.50, with an average of 0.40, indicating that these fragments exhibited a high divergence. Theta (per site) from Eta had almost the same trends as the nucleotide diversity, with an average number of 0.41. The number of nucleotide differences ranged from 1 to 225, with an average of 56, indicating that a difference of approximately 56 nucleotides existed on average between species-diverged fragments and other homologous sequences. The average number of InDel events was 15.4, and the average length of the Indel was 16 bp, demonstrating large variation between divergent fragments. Only five sets of sequences could be used to perform Tajima's D test. The results showed that half of Tajima's D test values were positive, but all not reached significant levels.

SSR distribution of species-diverged sequences and homologies

Although the primers used were designed based on microsatellites, only 23 out of 56 species-diverged sequences were found to have SSRs (Table 1), possibly due to the loss of

SSR units. The SSRs of the 23 sequences contained four types of motifs: mononucleotide (15.4%), dinucleotide (49.6%), trinucleotide (27.2%) and tetranucleotide (7.8%). Table 2 shows the distribution of SSR repeats in the species-diverged sequences. A total of 26 loci were detected, and dinucleotide SSRs comprised the highest proportion overall. The locus E56 had four SSRs, but the first three SSRs might have been amplified as the same locus by one primer pair, with additional primers starting at repeats 100 and 192 probably designed from different flanking sequences as different loci. The SSR motif TA was the most common accounting for 30.7% of the dimer motif loci and 15.0% of all 26 loci.

In addition, only seven SSRs were found in the 40 homologous sequences; much lower than the number of SSRs in the species-diverged SSR sequences. The detailed SSR distribution results are shown in Table 3. These seven SSRs contained six types of motifs, namely TA (14.3%), TC (14.3%), TCT (28.6%), CTC (14.3%), AAAG (28.6) and TGATC (28.6%).

In silico mapping of species- diverged sequences

To understand the evolution of the *Brassica*-diverged sequences in *A. thaliana,* a local BLAST analysis was performed to determine their location in *A. thaliana* blocks based on the highest identity ($E \leq 1.0E^{-5}$). Schranz et al. (2006) reported the presence of 24 genomic conservative blocks (labeled A to X) and block boundaries of five chromosomes in *A. thaliana.* Therefore, the positions of the homologous fragments of these diverged sequences in *A. thaliana* could be clearly determined. A total of 33 out of 56 sequences in

A. thaliana showed high homology and were distributed in 14 blocks: A, B, C, and D in chromosome 1 (Chr1), G and H in Chr2, F and N in Chr3, O and U in Chr4, and Q, R, S, and W blocks in Chr5 (Supplementary Table 5). To determine the detailed divergence pattern, these sequences were classified into three types according to species (*B. oleracea*, *B. napus*, and *B. rapa*). The *B. oleracea*-diverged sequences were derived from nine blocks (B, C, G, H, F, N, O, U, and S) of the five chromosomes of *A. thaliana*. *B. napus*-diverged sequences had high homologies with seven blocks (A, B, D, F, O, U, and R) located on four chromosomes of *A. thaliana* (excluding Chr2). However, *B. rapa* exhibited less similarity with *A. thaliana*, and only six blocks (B, O, U, R, Q, and W) showed homologies with *B. rapa*-diverged sequences. In general, the species-diverged regions were enriched in B block in Chr1, and O and U blocks in Chr4 and these regions were also highly variable.

Functional annotation

To investigate the putative function of the identified species-diverged fragments (56), the Blast2Go software was used to annotate these sequences on the basis of the cellular components, molecular function, and biological processes. Six SDS-SSR gene sequences out of 56 could be annotated and the other sequences may be nonfunctional or have had limited information in the database. The codes of the six sequences and their corresponding putative gene information are listed in Table 4. E2, E12, E20 and E41 were related to amidophosphoribosyl transferase, tetratricopetide-repeat thioredoxin-like protein, transcription factor une12 and retroelement pol poly polyprotein, respectively. E2 and E11

mainly existed in the plastid out of the cellular components. Four out of the remaining six sequences were associated with binding function including guanosine triphosphate (GTP) binding nucleotide acid, DNA and RNA binding, and transferase activity. As far as biological processes were concerned, these sequences participated in metabolic processes, primarily nucleotide, acid, and macromolecule metabolic processes. Morphogenesis was also noticeable among the biological processes in E2. E20 seemed to play a role in transcription, regulation of biological process and reproduction. To determine whether these genes were active, the annotated sequences were used as queries in a BlastN analysis against the *Brassica* expressed sequence tags (EST) database in NCBI. The results revealed that all genes had homologous fragments, and an average of one gene matched well with *Brassica* ESTs, indicating that these genes were active in the *Brassica* species.

Discussion

Divergent fragments between the three *Brassica* species contained 23 SSRs, more than their homologous sequences (7). SSRs are one of most mutable regions in a genome (Sia et al., 2000). Therefore, these SSRs may be an important force in promoting the generation and divergence of these sequences, affecting evolution and speciation in *Brassica*. SSRs in gene regions of eukaryotes are widely distributed and have been reported to evolve more quickly than other sequences (Huntley and Golding, 2000), and to play a positive role during adaptive evolution (Kashi and King, 2006). Although SSRs with unique loci in the genomes were selected, the *Brassica* databases did not have enough sequence information

to confirm the uniqueness of the SSR loci. The primer pairs designed based on the *B. rapa* genome could amplify the unique loci of *B. napus* and *B. oleracea*. In the results, 56 SDS-SSR from 37 SSRs of the three *Brassica* species were detected.

SSR markers have been extensively applied in genetic diversity analyses (Zhao et al., 2009), gene mapping (Hearnden et al., 2007), marker-assisted selection (Ma et al., 2007), and detection of cultivars purity (Xin et al., 2005) as they are codominant, abundant and highly polymorphic. Genome-specific markers, especially SSRs, are also convenient for detection of alien fragments or alien chromosomes in progeny resulting from wide hybridization (Huang and Brule-Babel, 2010) and within-species crosses (Harrison et al., 2011). Markers designed from genome-specific sequences have also been found to differentiate between the *B. rapa* and *B. oleracea* species (Kong et al., 2010). In wheat, starch-biosynthesis-specific primers have been developed based on intron sequences (Blake et al., 2004). Species-specific markers are also important in the comparative mapping and identification of introgressed regions (Schelfhout et al., 2004; Rodriguez-Suarez et al., 2011). In future, we aim to develop species-specific SSR markers from available SDS-SSR sequences

In the present study, *B. oleracea*-diverged fragments could be associated with nine blocks of five chromosomes of *A. thaliana*, confirming previous results of comparative mapping between *A. thaliana* and *B. oleracea* (Lim et al., 2007). *B. napus*-diverged fragments had high homologies to seven blocks on the *A. thaliana* chromosomes, but not to blocks on Chr2, which is consistent with the reported absence of any identical region

between the *B. napus* genetic map and *A. thaliana* Chr2 (Mayerhofer et al., 2005). *B. rapa*-diverged sequences had the least similarity with *A. thaliana.* Only six *A. thaliana* blocks matched well with these diverged SSR sequences. Overall, we could identify highly variable, diverged and overlapping regions for all three species (i.e., the B block in Chr1 and the O and U blocks in Chr4). These regions may have played a role in driving the divergence of the *Brassica* species.

Our results showed that evolution of the *Brassica* species was mainly influenced by its conservative blocks. A genome consists of a number of conservative blocks, and the genomic differences between related species mainly reflect the deviation in order or combination of these blocks. In rice, the genome could be divided into 30 conserved blocks, showing highly conserved structure relative to other related cereal species (Moore et al., 1995). The same evolutionary phenomenon can also be observed between human and mouse genomes, and their conserved segments were called "synteny blocks" (Pevzner and Tesler, 2003). Therefore, these conserved blocks may be considered as "evolutionary units," and it is possible that conserved block specific sequences may determine the divergence of species.

The putative functional annotation of six species-diverged genes showed that they were expressed mainly in the plastid, and were mainly involved in binding functions and transferase activity. According to the BlastN results of the SDS-SSR sequences against the EST database, all these genes in *Brassica* exhibited high activity. E2 fragments were found to be highly significant in species evolution, and especially in morphology determination.

These results are also consistent with a recent report on *B. rapa* genome sequencing (Wang et al., 2011), which suggested that multiple amplification occurred in the genes related to organ differentiation during *B. rapa* speciation. *B. rapa* is known to exhibit abundant variation in the roots, stems, and leaves. Species-specific sequences may be involved in the domestication of genera, such as that of sugar canes (Nakayama, 2004).

Species-diverged SSR fragments showed a high degree of variation between their homologous sequences. Only 23 sequences were found to have SSRs, probably because the repeat units of some SSRs of these species-diverged SSR fragments in *Brassica* were lost in the process of speciation. Based on the results of the functional annotation, six genes were found and the fragment, E2, was highly significant in species evolution, especially in morphology determination. The results presented here comprise a major step towards understanding species-diverged SSR fragments, and are expected to provide a better understanding of the origin, evolution, and divergence of species, as well as the organization of and changes in the *Brassica* genome. These fragments may also be developed as species-specific markers for use in species/chromosome identification.

Acknowledgements

This work was supported by National High Technology Research and Development Program of China (863 Program) (2011AA10A104) and Ministry of Agriculture, Modern Agricultural Industrial Technology System Program (CARS-13).

References:

Alix K, Joets J, Ryder CD, Moore J, Barker GC, Bailey JP, King GJ and Pat Heslop-Harrison JS (2008) The CACTA transposon Bot1 played a major role in *Brassica* genome divergence and gene proliferation. Plant J. 56: 1030-1044.

Anamthawat-Jonsson K and Heslop-Harrison JS (1993) Isolation and characterization of genome-specific DNA sequences in *Triticeae* species. Mol. Gen. Genet. 240: 151-158.

Aveskamp MM, Woudenberg JH, de Gruyter J, Turco E, Groenewald JZ and Crous PW (2009) Development of taxon-specific sequence characterized amplified region (SCAR) markers based on actin sequences and DNA amplification fingerprinting (DAF): a case study in the *Phoma exigua* species complex. Mol. Plant Pathol. 10: 403-414.

Blake NK, Sherman JD, Dvorak J and Talbert LE (2004) Genome-specific primer sets for starch biosynthesis genes in wheat. Theor. Appl. Genet. 109: 1295-1302.

Britten RJ (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. Proc. Natl. Acad. Sci. USA 99: 13633-13635.

Chung MC, Cheng YY, Fang SA and Lin YC (2007) A repetitive sequence specific to *Oryza* species with BB genome and abundant in *Oryza punctata Kotschy ex Steud*. Bot. Stu. 48: 263-272.

Conesa A and Gotz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genomics 2008: 619832.

da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FI and Costa de Oliveira A (2008) SSR Locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. Int. J. Plant Genomics 2008: 412696.

Dam P, Olman V, Harris K, Su Z and Xu Y (2007) Operon prediction using both genome-specific and general genomic information. Nucleic. Acids. Res. 35: 288-298.

Eggert LS, Beadell JS, McClung A, McIntosh CE and Fleischer RC (2009) Evolution of microsatellite loci in the adaptive radiation of Hawaiian honeycreepers. J. Hered. 100: 137-147.

Flavell RB, Bennett MD, Smith JB and Smith DB (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem. Genet. 12: 257-269.

Flavell RB, O'Dell M and Hutchinson J (1981) Nucleotide sequence organization in plant chromosomes and evidence for sequence translocation during evolution. Cold Spring Harb. Symp. Quant. Biol. 45 Pt 2: 501-508.

Gao D, Schmidt T and Jung C (2000) Molecular characterization and chromosomal distribution of species-specific repetitive DNA sequences from *Beta corolliflora*, a wild relative of sugar beet. Genome 43: 1073-1080.

Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41: 95-98.

Harrison E, Muir A, Stratford M and Wheals A (2011) Species-specific PCR primers for the rapid identification of yeasts of the genus *Zygosaccharomyces*. FEMS Yeast Res. 11: 356-365.

Hearnden PR, Eckermann PJ, McMichael GL, Hayden MJ, Eglinton JK and Chalmers KJ (2007) A genetic map of 1,000 SSR and DArT markers in a wide barley cross. Theor. Appl. Genet. 115: 383-391.

Hilson P, Allemeersch J, Altmann T, Aubourg S, Avon A, Beynon J, Bhalerao RP, Bitton F, Caboche M, Cannoot B, et al. (2004) Versatile gene-specific sequence tags for *Arabidopsis* functional genomics: transcript profiling and reverse genetics applications. Genome Res. 14: 2176-2189.

Hothorn M, Belkhadir Y, Dreux M, Dabi T, Noel JP, Wilson IA and Chory J (2011) Structural basis of steroid hormone perception by the receptor kinase BRI1. Nature 474: 467-471.

Huang XQ and Brule-Babel A (2010) Development of genome-specific primers for homoeologous genes in allopolyploid species: the waxy and starch synthase II genes in allohexaploid wheat (*Triticum aestivum* L.) as examples. BMC Res. Notes 3: 140.

Huntley M and Golding GB (2000) Evolution of simple sequence in proteins. J. Mol. Evol. 51: 131-140.

Junghans H and Metzlaff M (1988) Genome specific, highly repeated sequences of *Hordeum vulgate*:cloning, sequencing and squash dot test. Theor. Appl. Genet. 76: 728-732.

Kang HW and Kang KK (2008) Genomic characterization of *Oryza* species-specific CACTA-like transposon element and its application for genomic fingerprinting of rice varieties. Mol. Breeding 21: 283-292.

Kashi Y and King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22: 253-259.

Kong F, Ge C, Fang X, Snowdon RJ and Wang Y (2010) Characterization of seedling proteomes and development of markers to distinguish the *Brassica* A and C genomes. J. Genet. Genomics 37: 333-340.

La Mura M, Norris C, Sporle S, Jayaweera D, Greenland A and Lee D (2010) Development of genome-specific 5S rDNA markers in Brassica and related species for hybrid testing. Genome 53: 643-649.

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, et al. (2010) Building the sequence map of the human pan-genome. Nat. Biotechnol. 28: 57-63.

Li W, Wolynes PG and Takada S (2011) Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins. Proc. Natl. Acad. Sci. USA 108: 3504-3509.

Librado P and Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

Lim GA, Jewell EG, Li X, Erwin TA, Love C, Batley J, Spangenberg G and Edwards D (2007) A comparative map viewer integrating genetic maps for *Brassica* and *Arabidopsis*. BMC Plant Biol. 7: 40.

Lisch D, Chomet P and Freeling M (1995) Genetic characterization of the Mutator system in maize: behavior and regulation of *Mu* transposons in a minimal line. Genetics 139: 1777-1796.

Ma X, Wang K, Guo W and Zhang T (2007) Multiple SSR-PCR techniques and their application in cotton. Mol. Plant Breeding 5: 648-654.

Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B and Vanderleyden J (2003) Genome-specific higher-order background models to improve motif detection. Trends Microbiol. 11: 61-66.

Mayerhofer R, Wilde K, Mayerhofer M, Lydiate D, Bansal VK, Good AG and Parkin IA (2005) Complexities of chromosome landing in a highly duplicated genome: toward map-based cloning of a gene controlling blackleg resistance in *Brassica napus*. Genetics 171: 1977-1988.

Moore G, Devos KM, Wang Z and Gale MD (1995) Cereal genome evolution. Grasses, line up and form a circle. Curr. Biol. 5: 737-739.

Nakayama S (2004) Species-specific accumulation of interspersed sequences in genus *Saccharum*. Genes Genet. Syst. 79: 361-365.

Nevo E, Beharav A, Meyer RC, Hackett CA, Forster BP, Russell JR and Powell W (2005) Genomic microsatellite adaptive divergence of wild barley by microclimatic stress in 'Evolution Canyon', Israel. Biol. J. Linn. Soc. 84: 205-224.

Pankin AA and Khavkin EE (2011) Genome-specific SCAR markers help solve taxonomy issues: A case study with Sinapis arvensis (*Brassiceae*, *Brassicaceae*). Am. J. Bot. 98: e54-57.

Pevzner P and Tesler G (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc. Natl. Acad. Sci. USA 100: 7672-7677.

Remm M, Koressaar T and Joers K (2009) Automatic identification of species-specific repetitive DNA sequences and their utilization for detecting microbial organisms. Bioinformatics 25: 1349-1355.

Rodriguez-Suarez C, Ramirez MC, Martin A and Atienza SG (2011) Applicability of chromosome-specific SSR wheat markers for the introgression of *Triticum urartu* in durum wheat breeding programmes. Plant Genetic Resources-Charac. 9: 439-444.

Röbbelen G (1960) Contributions to the analysis of the *Brassica*-genome. Chromosoma 11: 205-228.

Schelfhout CJ, Snowdon R, Cowling WA and Wroth JM (2004) A PCR based B-genome-specific marker in *Brassica* species. Theor. Appl. Genet. 109: 917-921.

Schranz ME, Lysak MA and Mitchell-Olds T (2006) The ABC's of comparative genomics in the *Brassicaceae*: building blocks of crucifer genomes. Trends Plant Sci. 11: 535-542.

Sia EA, Butler CA, Dominska M, Greenwell P, Fox TD and Petes TD (2000) Analysis of microsatellite mutations in the mitochondrial DNA of *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA 97: 250-255.

Smith SD and Rausher MD (2011) Gene loss and parallel evolution contribute to species difference in flower color. Mol. Biol. Evol. 28: 2799-2810.

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV)

in genome content. Plos Genet. 5: e1000734.

Tang Z, Fu S, Ren Z and Zou Y (2009) Rapid evolution of simple sequence repeat induced by allopolyploidization. J. Mol. Evol. 69: 217-228.

U N (1935) 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.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, et al. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. 43: 1035-1039.

Xin Y, Zhang Z, Xiong Y and Yuan L (2005) Identification and purity test of super hybrid rice with SSR molecular markers. Rice Sci. 12: 7-12.

Zhao YG, Atta O and Lu CM (2009) Genetic Diversity of European and Chinese Oilseed *Brassica rapa* Cultivars from Different Breeding Periods. Agr. Sci. China 8: 931-938.

Figure legends

Figure 1 Amplification results for five representative species-diverged SSR fragments across three *Brassica* species DNA pools. Lane 1 of each group of primer pairs represents a *B. oleracea* DNA mixed sample, Lane 2 refers to the *B. napus* DNA pool, and Lane 3 corresponds to the *B. rapa* DNA pool. Arrows indicate the specific bands corresponding to the samples in the lanes.

Table 1 SSR distribution characteristics of the species-diverged sequences

Table 2 The distribution of SSR repeats in the species-diverged sequences

Table 3 SSR distribution of homologous sequences of the species-diverged fragments

Table 4 Sequences codes and corresponding putative gene information for six sequences isolated from species-diverged sequences

Supplementary Table 1 Nomenclature, code, and origin of the selected *Brassica* cultivars and accessions

Supplementary Table 2 Basic information on species-diverged sequences

Supplementary Table 3 The information of homologies sequences of species-diverged fragments

Supplementary Table 4 Degree and characteristics of divergence of the species-diverged sequences

Supplementary Table 5 Homology results between species-diverged sequences of *Brassica* species and *A. thaliana*

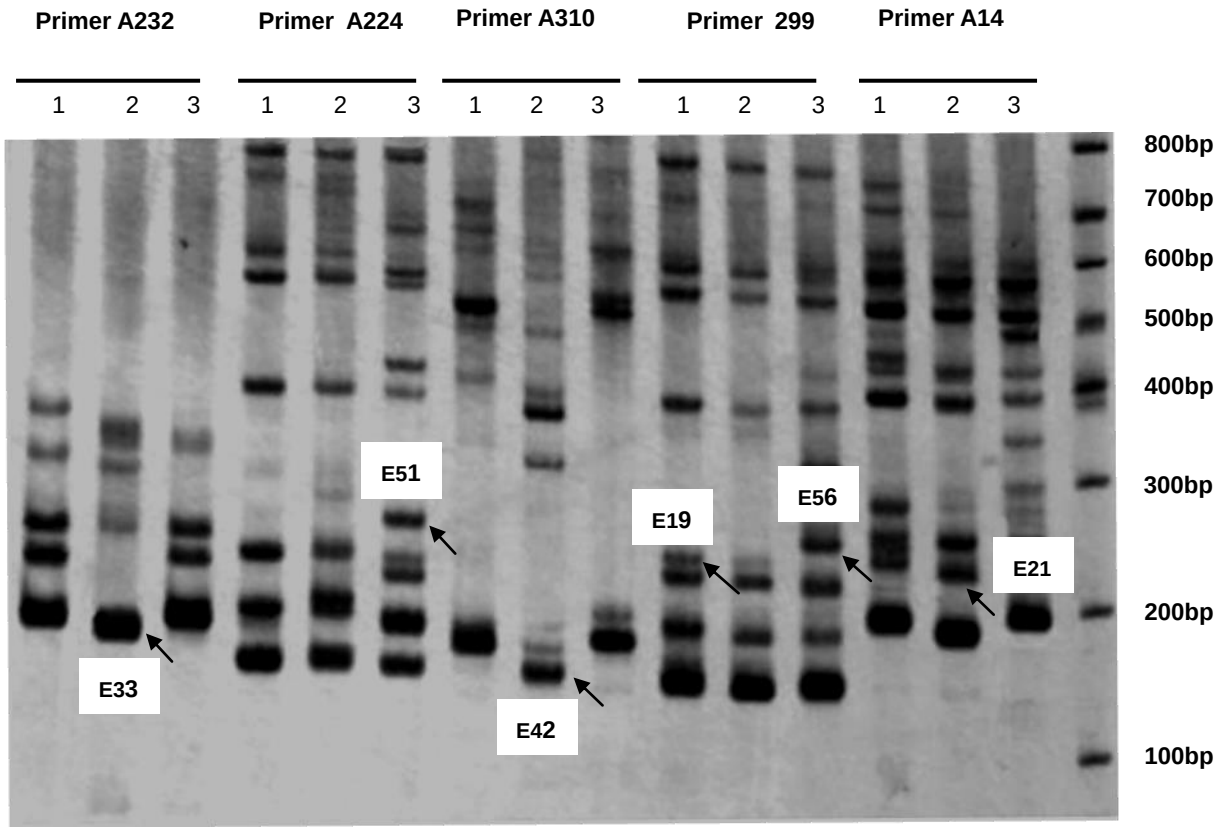

Figure 1 Amplification results for five representative species-diverged SSR fragments across three *Brassica* species DNA pools. Lane 1 of each group of primer pairs represents a *B. oleracea* DNA mixed sample, Lane 2 refers to the *B. napus* DNA pool, and Lane 3 corresponds to the *B. rapa* DNA pool. Arrows indicate the specific bands corresponding to the samples in the lanes

Table 1 SSR distribution characteristics of the species-diverged sequences

Code	Number of Loci	Motif	Start SSR1	End SSR1	Start SSR2	End SSR2	Start SSR3	End SSR3	Start SSR4	End SSR4
E1	1	(TGA)6	141	158						
E3	1	(AG)17	116	149						
E7	1	(CTT)6	138	155						
E8	1	(TA)13	54	79						
E9	1	(AT)12	85	108						
E10	1	(CT)9	129	146						
E14	1	(AT)9	49	66						
E15	1	(AT)24	49	96						
E16	1	(TA)14	31	58						
E19	1	(TAGT)5	36	55						
E20	1	(GGA)7	113	133						
E23	1	(GCA)6	53	70						
E32	1	(AGA)8	44	67						
E39	1	(TC)21	27	68						
E40	1	(CT)25	50	99						
E42	1	(TA)9	79	96						
E46	1	(TC)8	49	64						
E48	1	(TGT)11	176	208						
E49	1	(T)10	35	44						
E50	1	(C)11	156	166						
E52	1	(AC)8	18	33						
E53	1	(TA)12	116	139						
E56	4	(A)10-(A)11-(A)12-(AACT)6	100	109	118	128	136	147	192	215

Table 2 The distribution of SSR repeats in the species-diverged sequences

Motif	"a"/"b"	"a"repeats observed	"b" repeats observed	Total	(%) of SSR "-mer" type	% of repeat sequence
Monomer	A/T	2	1	3	75.0	11.5
	C/G	1	0	1	25.0	3.9
Dimer	TA	4	-	4	30.7	15
	AT	3	-	3	23.1	11.5
	AG/CT	2	1	3	23.1	11.5
	TC/GA	2	0	2	15.4	7.7
	AC/GT	1	0	1	7.7	3.9
Trinucleotide	TGA/TCA	1	0	1	14.3	3.9
	GGA/TCC	1	0	1	14.3	3.9
	GCA/TGC	1	0	1	14.3	3.9
	TGT/ACA	2	0	2	28.5	7.7
	CTT/AAG	1	0	1	14.3	3.9
	AGA/TCT	1	0	1	14.3	3.9
Tetranucleotide	TAGT/ACTA	1	0	1	50.0	3.9
	AACT/AGTT	1	0	1	50.0	3.9

Table 3 SSR distribution of homologous sequences of the species-diverged fragments

Motif		Repeats observed	(%)Group	(%)Overall
Dimer	TA	1	50.0	14.3
	TC	1	50.0	14.3
Trimer	TCT	2	66.7	28.6
	CTC	1	33.3	14.3
Tetranucleotide	AAAG	1	100.0	14.3
Pentamer	TGATC	1	100.0	28.6

Table 4 Sequences codes and corresponding putative gene information for six sequences isolated from species-diverged sequences

Code	Putative function	Cellular components	Molecular function	Biological process
E2	Amidophosphoribosyltransferase	Plastid	Binding; transferase activity	Biosynthetic process; nucleobase, nucleoside, nucleotide and nucleic acid metabolic process; cellular amino acid and derivative metabolic process; anatomical structure morphogenesis; multicellular organismal development;
E11	Protein	Plastid	RNA binding;transferase activity	Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
E12	Tetratricopetide-repeat thioredoxin-like protein	Cellular component	-	-
E20	Transcription factor une12	-	Transcription factor activity	Transcription; regulation of biological process; reproduction
E21	Protein	-	GTP binding;nucleotide binding; GTPase activity	-
E41	Retroelement pol polyprotein	-	DNA binding	DNA metabolic process