Theor Appl Genet (2013) 126:231–239
DOI 10.1007/s00122-012-1976-6

ORIGINAL PAPER

Development of InDel markers for *Brassica rapa* based on whole-genome re-sequencing

Bo Liu · Yan Wang · Wen Zhai · Jie Deng · Hui Wang · Yang Cui · Feng Cheng · Xiaowu Wang · Jian Wu

Received: 22 March 2012 / Accepted: 24 August 2012 / Published online: 13 September 2012

Abstract Genome-wide detection of short insertion/deletion length polymorphisms (InDels, <5 bp) in *Brassica rapa* (named the A genome) was performed by comparing whole-genome re-sequencing data from two *B. rapa* accessions, L144 and Z16, to the reference genome sequence of Chiifu-401-42. In total, we identified 108,558 InDel polymorphisms between Chiifu-401-42 and L144, 26,795 InDels between Z16 and Chiifu-401-42, and 26,693 InDels between L144 and Z16. From these, 639 InDel polymorphisms of 3–5 bp in length between L144 and Z16 were selected for experimental validation; 491 (77 %) yielded single PCR fragments and showed polymorphisms, 7 (1 %) did not amplify a product, and 141 (22 %) showed no polymorphism. For further validation of these intra-specific InDel polymorphisms, 503 candidates, randomly selected from the 639 InDels, were screened across seven accessions representing different *B. rapa* cultivar groups. Of these assayed markers, 387 (77 %) were polymorphic, 111 (22 %) were not polymorphic and 5 (1 %) did not amplify a PCR product. Furthermore, we randomly selected 518 InDel markers to validate their polymorphism in *B. napus* (the AC genome) and *B. juncea* (the AB genome), of which more than 90 % amplified a PCR product; 132 (25 %) showed polymorphism between the two *B. napus* accessions and 41 (8 %) between the two *B. juncea* accessions. This set of novel PCR-based InDel markers will be a valuable resource for genetic studies and breeding programs in *B. rapa*.

Communicated by R. Visser.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1976-6) contains supplementary material, which is available to authorized users.

B. Liu · Y. Wang · W. Zhai · J. Deng · H. Wang · Y. Cui · F. Cheng · X. Wang · J. Wu (✉)
Institute of Vegetables and Flowers,
Chinese Academy of Agricultural Sciences,
Beijing 100081, People's Republic of China
e-mail: wujian@caas.net.cn

B. Liu
e-mail: lb_bobo@yahoo.cn

Background

The genus *Brassica* includes a diverse range of crop species. Of particular importance are the diploids *Brassica rapa* (A genome) and *B. oleracea* (C genome), which include vegetable and oil crops, and the amphidiploids *B. napus* (AC genomes) and *B. juncea* (AB genomes), the sources of canola oil and condiment mustard. The genome of *B. rapa* has triplicated homologous counterparts of the corresponding segments in the Arabidopsis genome due to a whole-genome triplication that occurred 5–9 million years ago (Wang et al. 2011a). The *B. rapa* genome was selected as the initial reference genome for sequencing by the Multinational *Brassica* Genome Project, due to its relatively small size and high economic importance. The Chinese cabbage cultivar Chiifu-401-42 was adopted as the reference A genome, culminating in the recent release of a complete genome assembly representing at least 98 % of the gene space (Wang et al. 2011a). This resource now facilitates the development of sequence-based molecular markers in *B. rapa* to underpin genetic improvement of *Brassica* crops.

Molecular markers are a valuable tool in both basic and applied research for fingerprinting genotypes, analyzing genetic diversity, determining variety identity, marker-assisted breeding and phylogenetic analysis (Nagaraju et al. 2002; McCouch et al. 1997; Vos et al. 1995). Variation in

non-repetitive DNA has been exploited for generation of genetic markers using various approaches, including restriction fragment length polymorphism (RFLP, Botstein et al. 1980), random amplified polymorphism (RAPD, Williams et al. 1990) and amplified fragment length polymorphism (AFLP, Vos et al. 1995). Microsatellite, or simple sequence repeat (SSR, Tautz 1989), markers provided some advantages over earlier molecular markers due to their reproducibility, multi-allelic nature, codominant inheritance, relative abundance and relatively good genome coverage (Powell et al. 1996). Once developed, SSR markers have been relatively easy to analyze at moderate cost with separation and detection of allelic fragments by gel or capillary electrophoresis. One of the most important features of SSR markers is that they can detect multiple alleles. However, the complex and heterogenous mutation patterns of SSRs can introduce ambiguities in further data analysis (Vali et al. 2008), and genotyping errors may occur because of stutter bands and technical artefacts (null alleles, false alleles and size homoplasy) (Pompanon et al. 2005).

The discovery of large numbers of single-nucleotide polymorphisms (SNPs) in genome-scale sequencing initiatives provides an alternative approach to develop high-density markers. In contrast to SNPs, which have been studied extensively, other forms of natural genetic variation, such as insertion and deletion (InDel) polymorphisms, have received relatively little attention. In spite of the development of SNP genotyping technologies, InDel markers have practical value for those laboratories without infrastructure to perform SNP genotyping. Within the complete genomes of humans, rice and *Arabidopsis thaliana*, the InDel polymorphism frequency is 139, 1,050 and 151 InDels/Mb, respectively (Mills et al. 2006; Shen et al. 2004; Jander et al. 2002). A recent survey on *B. rapa* has indicated an InDel frequency of 4.83 InDels/kb. This study was based on comparison of 1,398 STSs within 557 BAC sequences of Chiifu-401-42 (Park et al. 2010). InDel polymorphisms are readily genotyped by fragment length polymorphisms, using the same experimental approaches routinely used for SSR markers (Bhattramakki et al. 2002). Moreover, the genomic density of InDels outnumbers that of SSRs by orders of magnitude. By harnessing the reference genome sequence of *B. rapa*, it is now possible to detect genome-wide InDel polymorphisms among different accessions using whole-genome re-sequencing to guide rapid and efficient development of PCR-based markers for high-resolution genetic analysis.

In this study, we used re-sequencing data from two *B. rapa* cultivars in comparison with the reference Chiifu-402-41 de novo assembled genome sequence to identify 26,693 short InDel polymorphisms across the genome. These could be genotyped with simple procedures based on size separation. Furthermore, we converted 639 InDel polymorphisms to PCR-based InDel assay markers and experimentally validated them, using the two re-sequenced accessions together with seven additional accessions representing different *B. rapa* subspecies and two *B. napus* and two *B. juncea* accessions. These stable InDel markers will be a useful resource for the *Brassica* research community.

Materials and methods

Plant materials and sequence data sets

The genome sequence of *B. rapa* line Chiifu-401-42 was retrieved from BRAD (http://brassicadb.org/brad/) and used as a reference. Using the sequencing-by-synthesis method, a total of 7.9-Gb Illumina pair-end reads were generated from *B. rapa* accession L144 (a rapid cycling line), as well as 1.4-Gb single-end reads and 2.9-Gb pair-end reads from accession Z16 (a heading Chinese cabbage line).

Seven *B. rapa* accessions from different cultivar groups: Chinese cabbage (ssp. *pekinensis*), turnip (ssp. *rapa*), Caixin (ssp. *parachinensis*), Zi Caitai (ssp. *chinensis* var. *purpurea* Bailey), Komatsuna (ssp. *perviridis*), Yellow sarson (ssp. *tricolaris*), rape seed (ssp. *oleifera*) and the two re-sequenced accessions, Z16 (ssp. *pekinensis*) and L144 (rapid cycling), were used for InDel polymorphism validation. Two *B. juncea* accessions from different groups, A12451 (var. *napiformis*) and A12907 (var. *foliosa*), together with two *B. napus* accessions (A12409 and A12412) were also tested for InDel polymorphism. One hundred BC_2DH lines were developed from a cross between the inbred rapid cycling accession L144, a donor parent, and the DH Chinese cabbage line Z16, a recurrent parent.

InDel detection

The process used to detect InDel polymorphisms (Fig. 1) involved three steps. For the first step of short reads alignment, the Short Oligo-nucleotide Alignment Program (SOAP) software (Li et al. 2008) was used to align single reads or paired reads to the reference with default parameters. The alignment with the least number of mismatches was nominated as the 'best hit' (Wang et al. 2008). No gaps were allowed in the alignment. The next step involved re-mapping the unmapped read pairs to the reference by allowing at most a 5-bp gap and less than two mismatches. This limited the InDels within a range of 1–5 bp in length. Gaps supported by at least three pair-end reads were retained (Wang et al. 2008). The third step involved

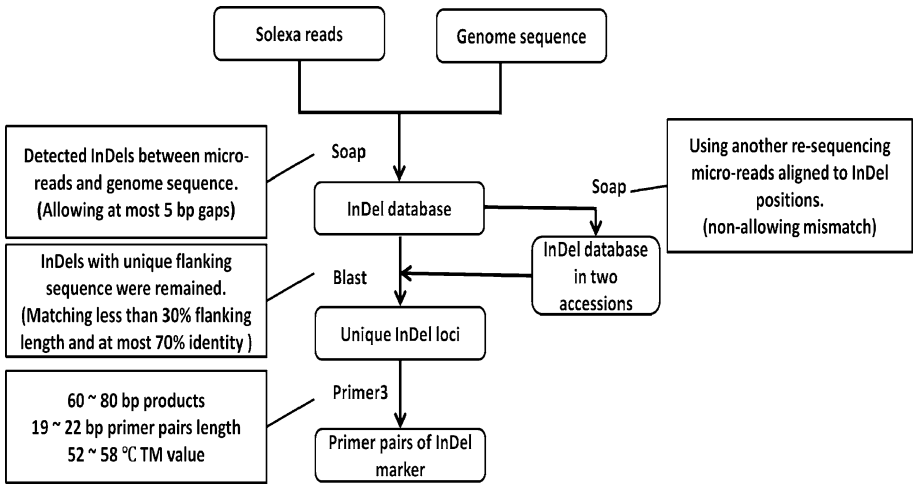

Fig. 1 Flowchart of the procedure used to detect InDel polymorphisms

identification of unique InDel polymorphisms. We extracted 150-bp flanking sequences on both sides of an InDel to search the reference genome sequence using a simple Visual C++ script. The InDel was removed if more than 30 % of the 150-bp flanking sequences on both sides matched multiple sites with at least 70 % identity. The scripts for InDel polymorphism detection from re-sequencing PE reads and the reference sequence are presented on BRAD (http://brassicadb.org/brad/tools/InDel/).

For the detection of InDel polymorphisms between the two re-sequenced *B. rapa* accessions, we used the reference Chiifu-401-42 genome as a 'bridge' to sequentially detect sequence polymorphisms between them. The single-end reads of Z16 were aligned to the reference sequence of Chiifu-402-41 by SOAP with no gaps allowed. To avoid false detection of polymorphisms, multiple-hit reads were filtered out from the dataset. The aligned reads dataset was compared against the InDel polymorphisms dataset identified between L144 and Chiifu-401-42. Only those InDel polymorphisms where the sequences were identical between Z16 and Chiifu-401-42 were deemed as InDel polymorphisms between Z16 and L144.

Primer 3 (Rozen and Skaletsky 2000) was used to design PCR primers, with a constraint of generating products of 60–180 bp. The length of primer pairs was limited to 19–22 bp. The T_m value was restricted to between 52 and 58 °C.

Experimental validation of DNA polymorphism

Genomic DNA was isolated from mature leaves as described by Wang et al. (2005). PCR was performed in 15-μl reaction volumes containing 0.4 U of *Taq* DNA polymerase with 1× PCR buffer (Tiangen, Beijing, China), 0.5 μM of each primer, 300 μM of each dNTP, 1.5–2.0 mM $MgCl_2$ and approximately 30 ng of genomic DNA as a template. Thermocycling was started at 94 °C for 7 min and followed by 35 cycles of 94 °C for 40 s, 57 °C for 40 s and 72 °C for 1 min, with a final extension at 72 °C for 10 min. The PCR products were separated on 8 % polyacrylamide gel and visualized by silver staining.

Results

Identification of short InDels between Chiifu-401-42 and L144

The genome of L144 (a rapid cycling line of *B. rapa*) was re-sequenced using the Illumina GA II platform. Reads in length of 44 nt were generated from two paired-ends (PE) libraries of 489 and 2,224-bp insert sizes. Approximately, 180 million high-quality reads (with an average quality ≥20) were generated, corresponding to about 7.9 gigabases (Gb) of sequence.

Using the SOAP software (Li et al. 2008), 5.3 Gb of sequence (accounting for 68 % of all data) was aligned to the reference genome, covering 88 % of the reference sequence with an 18.7× sequencing depth. About 79 % of the mapped PE reads were uniquely aligned (approximately 14.8× depth) and used to detect insertion/deletion (InDel) polymorphisms. Approximately, 21 % of the mapped PE reads were aligned to multiple sites. This was most likely due to the triplication of the *B. rapa* genome throughout the nested polyploidy events (Wang et al. 2011a).

To define an InDel, we identified sequence regions where at least three paired reads from L144 were aligned to the reference genome sequence. Given the short-read sequencing strategy, it was necessary to limit the InDel length detection to ≤5 bp to avoid alignment errors. Based on this criterion, we surveyed 4.2 Gb of uniquely aligned PE reads for short InDel polymorphisms. This revealed a total of 108,558 putative short InDels (1–5 bp) between

Chiifu-402-41 and L144, corresponding to a frequency of 434.2 InDels/Mb within an approximately 249.7 Mb physical distance. Among the 1- to 5-bp short InDels detected, there was a skewed distribution of InDel lengths: most (63 %) were single-nucleotide, 19 % were two-nucleotide and only 10 % were three-nucleotide InDels (Table 1). In addition to natural variation in InDel size, some of this observed bias may be due to the detection method. Due to the deep re-sequencing at 14.8× depth, we detected a large number of short InDels with a dense distribution across each of the ten *B. rapa* chromosomes (Fig. 2). The normalized average occurrence of InDels varied across the chromosomes, falling within the range of approximately 379–517 InDels/Mb of genomic DNA (Table 2). Based on this distribution of InDels, it was possible to construct high-density genetic maps and select InDels within specific regions for fine mapping.

We next examined the distribution of the 108,558 InDels relative to genes of *B. rapa* and found that 15,824 (15 %) were located within annotated genes (Table 3). Some 2,822 identified InDels were located within exons, where gene function may be expected to be influenced. Of these, 1,163 (41 %) were non-3-nucleotide InDels that were predicted to cause frameshifts that would lead to the premature termination of the encoded proteins.

Identification of short InDel polymorphisms between two re-sequenced *B. rapa* accessions, L144 and Z16

For cost-effective detection of InDel polymorphisms between the two re-sequenced *B. rapa* accessions, we used the reference Chiifu-401-42 genome as a 'bridge' to sequentially detect sequence polymorphisms between them. The InDel polymorphisms identified between L144 and Chiifu-401-42 were used as seeds to detect InDels between L144 and another *B. rapa* accession, Z16, for which 1.4 Gb of 35-bp single-end reads had been generated.

Table 1 The number and distribution frequency of short InDels (1–5 bp) identified in the *B. rapa* genome

InDel size (bp)	L144 versus Ref		L144 versus Z16	
	Count	Frequency (InDels/Mb)	Count	Frequency (InDels/Mb)
1	68,074	272.3	16,521	66.0
2	20,338	81.4	4,915	19.7
3	10,367	41.5	2,715	10.9
4	6,893	27.6	1,814	7.3
5	2,886	11.5	728	2.9
Total	108,558	434.0	26,693	106.8

Fig. 2 Distribution of DNA polymorphisms identified between *B. rapa* accessions Chiifu-402-41 and L144 along each chromosome. The *horizontal scale* indicates the physical distance (from the *B. rapa* physical map; http://brassicadb.org/brad/). The *vertical scale* indicates the number of InDel polymorphisms within each 1M region between Chiifu-402-41 and L144

The single-end reads of Z16 were aligned to the reference sequence of Chiifu-402-41 with no gaps allowed, to search accurately between the single-end reads and the sequence of Chiifu-401-42. To avoid false detection of polymorphisms, multiple-hit reads were filtered out from the dataset. The aligned reads dataset was compared against the InDel polymorphisms dataset identified between L144 and Chiifu-401-42. Only the InDel polymorphisms where the sequences were identical between Z16 and Chiifu-401-42 were deemed as InDel polymorphisms between Z16 and L144. Using this procedure, we identified a total of 26,693 (1–5 bp) InDel polymorphisms between Z16 and L144 that were distributed across the ten chromosomes, varying from 3,510 on the chromosome A09 to 1,812 on chromosome A04 (Table 1). The frequency of InDel polymorphisms between Z16 and L144 varied from 71.7 InDels/Mb on the chromosome A06 to 148.2 InDels/Mb on chromosome A07 (Table 2).

As Illumina sequencing technology has developed, the read length has increased from 35 to 150 bp. To validate our strategy for identifying InDel polymorphisms, we repeated the analysis with long pair-end reads; approximately, 2.9 Gb of 80-bp reads from Z16 were used to identify InDel polymorphisms between Z16 and L144. Nearly 1.6 Gb of the read sequences were aligned to the reference genome using the SOAP software with default parameters. Employing the same method used to identify InDel polymorphisms between L144 and Chiifu-401-42, a total of 26,795 InDels were identified between Z16 and Chiifu-401-42. In total, 120,675 InDel polymorphisms were detected between Z16 and L144 by comparing the two batches of InDel datasets identified between L144 and Chiifu-401-42, and Z16 and Chiifu-401-42. The comparison between two InDel datasets also revealed 7,339 common InDels where L144 and Z16 had the same sequences compared to Chiifu-401-42. Nearly, 98 % of the 26,693 InDels identified from 35-bp single-end reads were included in the 120,675 InDel polymorphisms dataset. A further 94,516 InDels were identified using the 80-bp pair-end reads between Z16 and L144.

Experimental validation of short InDel polymorphisms

To validate the InDels identified between L144 and Z16, we selected 639 of the 26,693 InDels and converted them to PCR-based markers. To facilitate screening using PAGE, only InDels of 3–5 bp in length were selected.

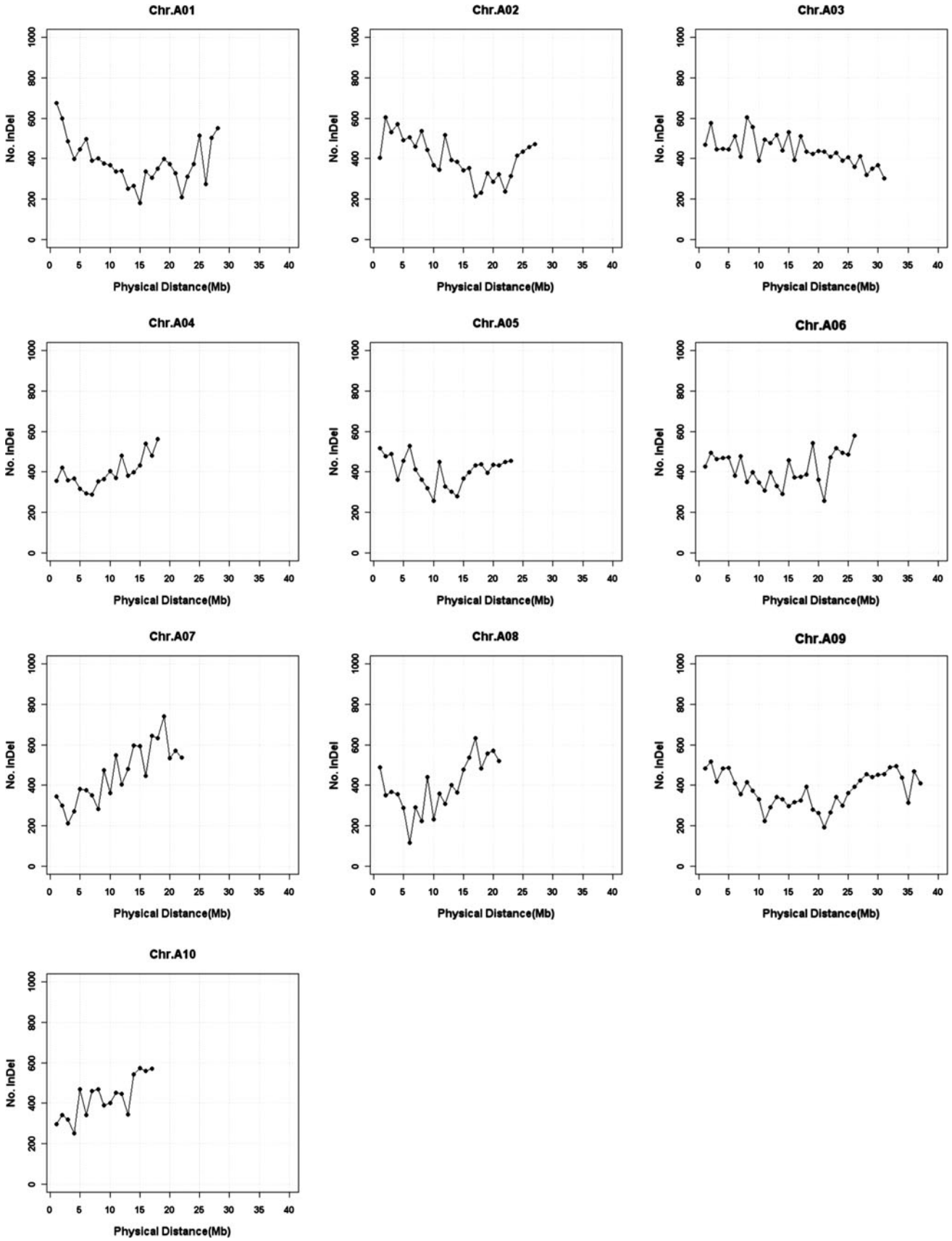
Chr.A01
Chr.A02
Chr.A03
Chr.A04
Chr.A05
Chr.A06
Chr.A07
Chr.A08
Chr.A09
Chr.A10
No. InDel
Physical Distance(Mb)
0
200
400
600
800
1000
0 5 10 15 20 25 30 35 40

Table 2 Short InDel polymorphisms (1–5 bp) identified in each chromosome of *B. rapa*

Chr	CD[a] (Mb)	L144 versus Ref		L144 versus Z16	
		No. InDels	Frequency (InDels/Mb)	No. InDels	Frequency (InDels/Mb)
A01	28.6	10,832	378.6	2,663	93.1
A02	27.9	10,965	393.7	2,577	92.4
A03	31.7	13,701	431.9	3,055	96.4
A04	19.0	7,167	379.2	1,812	95.4
A05	23.9	9,340	390.1	2,515	105.2
A06	26.3	10,914	415.5	1,887	71.7
A07	22.6	10,085	446.4	3,350	148.2
A08	21.6	8,363	387.2	2,005	92.8
A09	37.1	14,025	517.2	3,510	94.6
A10	17.6	7,230	410.8	1,923	109.3

[a] The physical distances of *B. rapa* chromosomes

Table 3 Location distribution of short InDels identified in *B. rapa*

Location	Count
Coding exon	2,822
Intron	13,002
Intergenic	92,734
Total	108,558

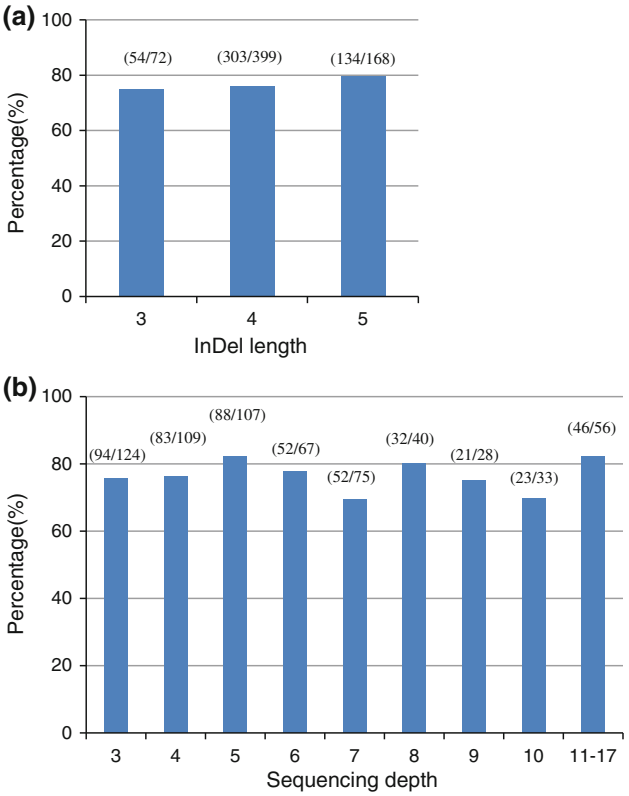

Fig. 3 The relationship between polymorphism rate and the characters of markers of *B. rapa*. **a** The relationship between polymorphism rate and InDel length. **b** The relationship between polymorphism rate and sequencing depth. Analyses were carried out on 639 InDel markers validated between L144 and Z16. The *number in the bracket* indicates the number of polymorphism markers in total markers

To reduce non-specific amplification, InDels located within assembled sequences represented by multiple homologs were excluded (data not shown). The selected InDels were evenly distributed across the *B. rapa* genome, with most residing in non-coding sequences. Based on this selection, we designed primer pairs to amplify fragments of 80–140 bp surrounding the InDels. In the PCR analysis, 614 of the 639 primer pairs (96 %) gave reliable amplification using genomic DNA of both L144 and Z16 as the template, and 491 (77 %) revealed identifiable polymorphisms between L144 and Z16 using PAGE, including 18 primer pairs that produced an amplicon in only one genotype (3 %). These 18 primer pairs were not propitious to genetic analysis. Another 141 primer pairs (22 %) gave monomorphic results, and no amplification was observed for 7 primer pairs (1 %). According to the characters of InDel markers (sequencing depth and InDel length), the polymorphism rate increased slightly along with increasing InDel length, and the polymorphism rate varied from 75 % on 3-bp length to 80 % on 5-bp length (Fig. 3a). However, there was no distinct relationship between reads coverage of InDel regions with polymorphism rate (Fig. 3b). The result indicated that the read depth of InDel polymorphisms position (at least three paired reads) was enough to obtain high polymorphism rate InDel markers.

To check the universal applicability of the InDel markers, we screened seven *B. rapa* accessions representing a range of different subspecies for polymorphisms using 503 of the 3- to 5-bp InDels. In this analysis, 387 of the InDels (77 %) identified from L144 and Z16 were polymorphic across the wider taxa, with 111 (22 %) being monomorphic and only 5 having no amplification. Of the 387 polymorphic InDels, 325 appeared to be conserved with those identified between Z16 and L144 (Supplementary Document). In addition, 59 InDels were monomorphic between Z16 and L144, but revealed identifiable polymorphisms in the seven *B. rapa* accessions. This indicates that these InDels have universal applicability in *B. rapa* subspecies.

Two *B. napus* accessions and two *B. juncea* accessions from different varieties (var. *napiformis* and var. *foliosa*) were selected to validate the InDel markers that were polymorphic among *B. rapa* accessions. From 518 tested InDels, 132 (25 %) showed polymorphism between the two *B. napus* accessions and 41 (8 %) between the two *B. juncea* accessions. Some 352 InDel markers (68 %) were monomorphic between the two *B. napus* accessions,

and 443 (86 %) between the two *B. juncea* accessions. There were 34 primer pairs that produced no amplification in either species (Supplementary Document).

To assess the value of the InDel markers for genetic analysis, 100 *B. rapa* BC_2DH lines (ILs) were selected from a Z16/L144 BC_2DH population to show introgression patterns on chromosome A03 (Fig. 4). There were 26 InDel markers evenly distributed on chromosome A03. Genotyping of these 26 InDels among the 100 BC_2DH lines showed that 68 of 100 lines were genetically same as the recurrent parent (Z16). Twelve lines contained more than half of the genomic segments from the donor parent (L144), while 20 lines contained less than half. This indicates that the developed InDel markers are useful for identifying the genetic composition of *B. rapa* lines and provide a valuable source of allelic diversity for genetic and molecular dissection of traits.

Given that the reference genome sequence was assembled from 72× Illumina short-read and BAC-end sequences (Wang et al. 2011a), the majority of markers were located in euchromatic, non-repetitive regions. To facilitate the construction of a genetic map with evenly distributed markers, we defined 518 bins representing every 500-kb block within the reference *B. rapa* genome. The InDel markers we identified were distributed in 363 (70 %) of these bins (Fig. 5). Using these markers, a high-density linkage map that contained 414 InDel markers and 94 SSRs was constructed for a DH population derived from crossing between Z16 and L144 (Wang et al. 2011b). Based on alignment to this reference linkage map, these InDel markers appear to be evenly distributed across the ten linkage groups (Fig. 5), providing encouragement that it will be possible to select appropriate markers for fine mapping and construction of high-density genetic maps.

Discussion

We have shown the feasibility and facility of using genomic re-sequencing data to identify putative InDel polymorphisms. These InDel loci can then be validated to generate informative genetic markers with a reliably high rate of polymorphism. More importantly, since the coordinates of the InDels are known in relation to a reference genome, it is possible to develop genetic markers within specific genome regions to assist in efficient construction of genetic maps and for fine mapping of, for example, BC_2DH lines as used in this study or near isogenic lines (NILs).

InDel markers may be assayed using the same separation and detection technologies as SSR markers. In this study, we selected 639 InDels of 3–5 bp to develop PCR-based markers, of which 553 (87 %) were polymorphic either between the two re-sequenced accessions (491) or among seven accessions from different subspecies of *B. rapa* (387). Only 25 (4 %) of the primer pairs yielded no amplification product from either of the two re-sequenced accessions. This can be explained by sequence variations in the primer binding sites among *B. rapa* accessions as we designed primers based on the reference genome sequence.

The 108,558 short InDel polymorphisms identified between Chiifu-402-41 and L144 represent an average of 434.23 InDels/Mb across the entire *B. rapa* genome. This is a low estimate because we limited the InDel size to

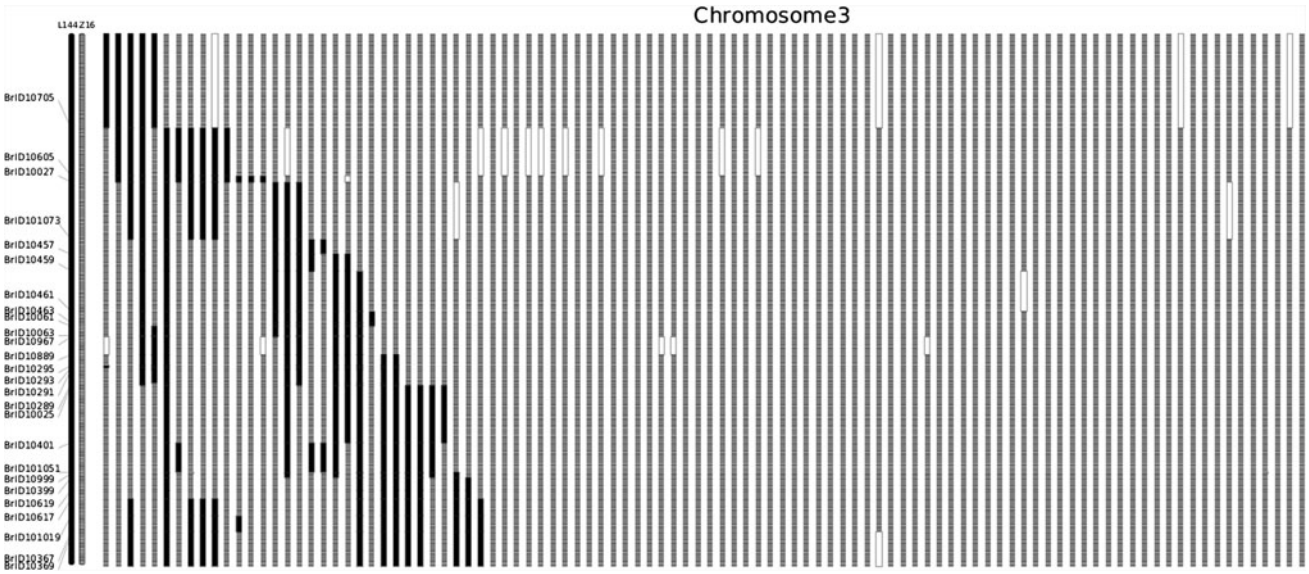

Fig. 4 Graphical genotypes of 100 *B. rapa* BC_2DH lines on chromosome A03. *Black bars* represent the donor parent; *gray bars* represent the recurrent parent. *White bars* represent missing data. The names of 26 InDel markers are indicated on the *left margin*

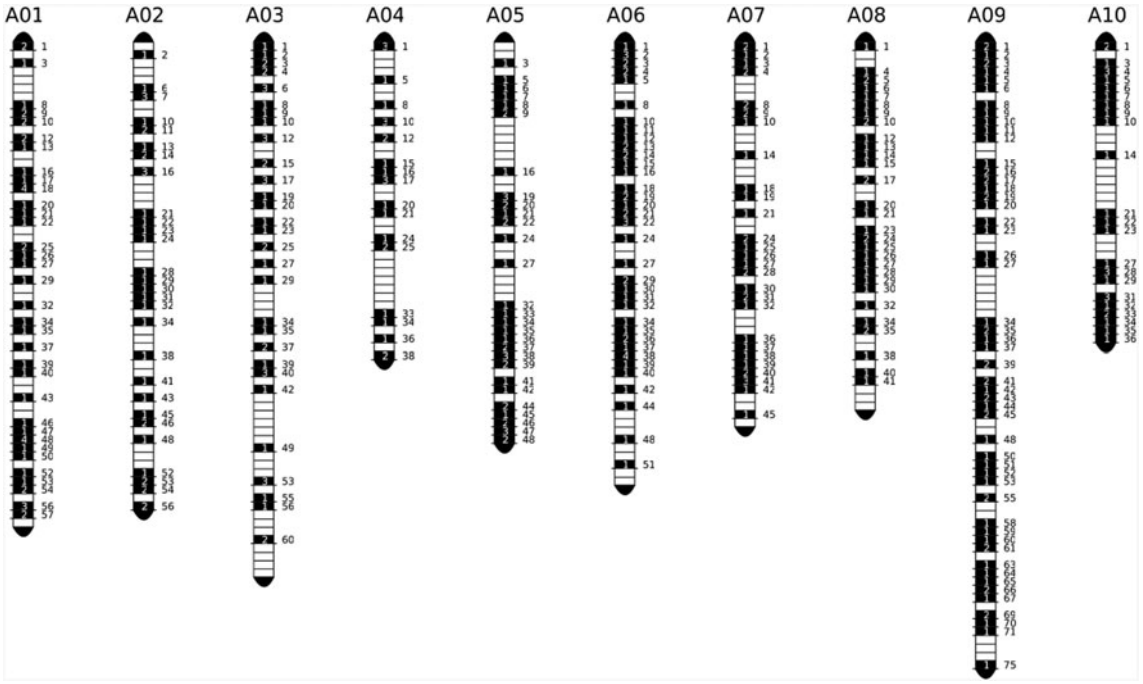

Fig. 5 Distribution of experimentally validated polymorphic InDel markers on each chromosome of *B. rapa*. Bin names are listed to the right of the chromosomes. Each bin comprised 500-kb physical distance. The number of InDel markers in each filled bin is indicated in the *boxes*. *White boxes* indicated regions that do not include any markers

1–5 bp during the identification process. In addition, we used short reads (35 bp) for the analysis, although we have shown that by using longer reads (80 bp) there is a substantial increase in detected InDels between the two re-sequenced accessions. As expected, the density of 434.23 InDels/Mb is ten times lower than the previously reported 4,830 InDels/Mb for InDels of 1–100 bp (Park et al. 2010). However, one factor that should be taken into account is that in the previous study the InDels and SNPs were identified among eight genotypes rather than between two genotypes as used here. Compared to the estimate of 151 InDels/Mb (1–100 bp in size) in *A. thaliana* (Jander et al. 2002), the frequency of short InDel polymorphisms in *B. rapa* is more than three times greater, although approximately 2.4 times less than that observed in rice (Shen et al. 2004). This high frequency of polymorphism in *B. rapa* may be explained by the outcrossing reproductive mode of the species, which is mediated by strong self-incompatibility. It has been noted previously that outcrossing species have higher levels of sequence variation than self-fertile species (Pollak 1987).

Second-generation sequencing platforms are more cost-effective than traditional Sanger sequencing (Datta et al. 2010). However, there remains a relatively high entry cost for generating datasets representing high sequencing depth, such as the 18-fold depth for L144 in this study. We therefore demonstrated the value of using a low depth set with the 1.4-Gb single-end reads from *B. rapa* accession Z16, corresponding to about threefold depth, to identify InDel polymorphisms between L144 and Z16. A key aspect of our strategy was to use the reference Chiifu-401-42 genome to bridge the polymorphisms identified in the two re-sequenced accessions. Once the location of polymorphisms between one re-sequenced accession and the reference was established, those between the two re-sequenced accessions are readily distinguished at corresponding positions where the second accession is identical to the reference. We identified 26,693 InDel polymorphisms using this procedure. The subsequent validation of the markers indicated that this strategy was a practical route for marker development. Since the *B. rapa* genome underwent a hexaploidy event after its divergence from *A. thaliana* (Wang et al. 2011a), to ensure locus-specific amplification, InDels located within the assembled sequence represented by multiple homologs were excluded. However, there was a low proportion of non-specific amplification, possibly because of amplification from homologs within un-assembled sequence regions. We experimentally validated the detected InDel polymorphisms between the two re-sequenced accessions by converting them into PCR-based markers. Furthermore, a high-density linkage map was constructed using the validated InDel markers for the RCZ16_DH population (Wang et al. 2011b). In contrast to the high polymorphism rate of InDels among the seven *B. rapa* accessions representing different subspecies, the tested two *B. napus* and two *B. juncea* accessions showed

much lower polymorphism rates. However, nearly 93 % of the test InDels showed amplified fragments, indicating that those markers will be potentially useful for genetic research on *B. napus* or *B. juncea* when there are more accessions being screened.

In contrast to the high frequency of 15,260 SNPs/Mb observed in *B. rapa* (Park et al. 2010), the frequency of 1-bp InDels is only approximately 272.3 InDels/Mb across the whole *B. rapa* genome (Table 1). Therefore, in the same experimental conditions, SNPs are better for developing polymorphism markers than InDels of 1-bp size.

Acknowledgments The authors wish to thank Graham J King for valuable suggestions for manuscript correction. This work was supported by National High Technology Research and Development Program of China (863 Program, No. 2008AA10Z155) and Core Research Budget of the Non-profit Governmental Research Institution (ICS, CAAS, 1610032011011). This work was done in the Key Laboratory of Biology and Genetic Improvement of Horticulture Crops, Ministry of Agriculture, People's Republic of China and the Sino-Dutch Joint Lab of Horticulture Genomics Technology in Beijing.

References

Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register JC 3rd, Tingey SV, Rafalski A (2002) Insertion–deletion polymorphisms in 3′ regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol Biol 48(5–6):539–547

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32(3):314–331

Datta S, Datta S, Kim S, Chakraborty S, Gill RS (2010) Statistical analyses of next generation sequence data: a partial overview. J Proteomics Bioinforma 3(6):183–190

Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL (2002) Arabidopsis map-based cloning in the post-genome era. Plant Physiol 129(2):440–450

Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: Short Oligonucleotide Alignment Program. Bioinformatics 24(5):713–714

McCouch SR, Chen X, Panaud O, Temnykh S, Xu Y, Cho YG, Huang N, Ishii T, Blair M (1997) Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol Biol 35(1–2):89–99

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16(9):1182–1190

Nagaraju J, Kathirvel M, Kumar RR, Siddiq E, Hasnain SE (2002) Genetic analysis of traditional and evolved Basmati and non-Basmati rice varieties by using fluorescence-based ISSR-PCR and SSR markers. Proc Natl Acad Sci USA 99(9):5836

Park S, Yu HJ, Mun JH, Lee SC (2010) Genome-wide discovery of DNA polymorphism in *Brassica rapa*. Mol Genet Genomics 283(2):135–145

Pollak E (1987) On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics 117(2):353–360

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nat Rev Genet 6:847–859

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1(7):215–222

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

Shen YJ, Jiang H, Jin JP, Zhang ZB, Xi B, He YY, Wang G, Wang C, Qian L, Li X, Yu QB, Liu HJ, Chen DH, Gao JH, Huang H, Shi TL, Yang ZN (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol 135(3):1198–1205

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17(16):6463–6471

Vali U, Brandstrom M, Johansson M, Ellegren H (2008) Insertion–deletion polymorphisms (indels) as genetic markers in natural population. BMC Genet 9:8

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23(21):4407–4414

Wang X, Lou P, Bonnema G, Yang B, He H, Zhang Y, Fang Z (2005) Linkage mapping of a dominant male sterility gene Ms-cd1 in *Brassica oleracea*. Genome 48(5):848–854

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Yang H (2008) The diploid genome sequence of an Asian individual. Nature 456(7218):60–65

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, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li 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, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Jin H, Parkin IA, Batley J, Kim JS, Just J, Li J, Xu J, Deng J, Kim JA, Yu J, Meng J, Min J, Poulain 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, 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 Z, Li Z, Xiong Z, Zhang Z (2011a) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43(10):1035–1039

Wang Y, Sun S, Liu B, Wang H, Deng J, Liao Y, Wang Q, Cheng F, Wang X, Wu J (2011b) A sequence-based genetic linkage map as a reference for *Brassica rapa* pseudochromosome assembly. BMC Genomics 12:239

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18(22):6531–6535