Theor Appl Genet (2013) 126:2017–2027
DOI 10.1007/s00122-013-2114-9

ORIGINAL PAPER

Genome-wide SNP discovery in mungbean by Illumina HiSeq

Kyujung Van · Yang Jae Kang · Kwang-Soo Han · Yeong-Ho Lee · Jae-Gyun Gwag · Jung-Kyung Moon · Suk-Ha Lee

Received: 23 July 2012 / Accepted: 20 April 2013 / Published online: 15 May 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract Mungbean [*Vigna radiata* (L.) Wilczek], a self-pollinated diploid plant with $2n = 22$ chromosomes, is an important legume crop with a high-quality amino acid profile. Sequence variation at the whole-genome level was examined by comparing two mungbean cultivars, Sunhwanokdu and Gyeonggijaerae 5, using Illumina HiSeq sequencing data. More than 40 billion bp from both mungbean cultivars were sequenced to a depth of 72×. After de novo assembly of Sunhwanokdu contigs by ABySS 1.3.2 (N50 = 9,958 bp), those longer than 10 kb were aligned with Gyeonggijaerae 5 reads using the Burrows–Wheeler Aligner. SAMTools was used for retrieving single nucleotide polymorphisms (SNPs) between Sunhwanokdu and Gyeonggijaerae 5, defining the lowest and highest depths as 5 and 100, respectively, and the sequence quality as 100. Of the 305,504 single-base changes identified, 40,503 SNPs were considered heterozygous in Gyeonggijaerae 5. Among the remaining 265,001 SNPs, 65.9 % (174,579 cases) were transitions and 34.1 % (90,422 cases) were transversions. For SNP validation, a total of 42 SNPs were chosen among Sunhwanokdu contigs longer than 10 kb and sharing at least 80 % sequence identity with common bean expressed sequence tags as determined with est2genome. Using seven mungbean cultivars from various origins in addition to Sunhwanokdu and Gyeonggijaerae 5, most of the SNPs identified by bioinformatics tools were confirmed by Sanger sequencing. These genome-wide SNP markers could enrich the current molecular resources and might be of value for the construction of a mungbean genetic map and the investigation of genetic diversity.

Communicated by B. Diers.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2114-9) contains supplementary material, which is available to authorized users.

K. Van · Y. J. Kang · K.-S. Han · Y.-H. Lee · S.-H. Lee (✉)
Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Korea
e-mail: sukhalee@snu.ac.kr

J.-G. Gwag
National Agrobiodiversity Center, National Academy of Agricultural Science, Suwon 441-707, Korea

J.-K. Moon
Upland Crop Division, National Institute of Crop Science, Suwon 441-857, Korea

S.-H. Lee
Plant Genomics and Breeding Research Institute, Seoul National University, Seoul 151-921, Korea

Introduction

Mungbean [*Vigna radiata* (L.) Wilczek] is an important legume crop in Asia, particularly in India and in South, Southeast and East Asia, where tropical and subtropical weather conditions are prevalent (Lambrides and Godwin 2007; Chankaew et al. 2011). The South and Southeast Asian diet is mostly cereal-based, and thus mungbean with its highly digestible protein (ca. 25 %) can be an important nutritional complement for the balanced nutrition of people in these regions. In addition, mungbean sprouts are popular in Asian cuisine because they are a good sources of vitamins and minerals (Lambrides and Godwin 2007; Somta and Srinives 2007). Therefore, mungbean studies have focused on seed traits, such as size, shape, color, sprouting quality and protein quality and quantity. The seed yield has been increased by improving resistance against diseases

such as mungbean yellow mosaic virus (MYMV) and Cercospora leaf spot (Lambrides and Godwin 2007; Chankaew et al. 2011).

Although mungbean shows a photoperiod response to short-day length, this crop can be grown at a wide range of latitudes (Imrie 1996). In addition, this warm-season grain legume can mature in approximately 60 days and adapt to relatively harsh conditions such as drought. Similar to other legumes, mungbean is able to fix atmospheric nitrogen through its symbiotic relationship with the soil-borne bacteria *Rhizobium* (Lawn 1979; Tangphatsornruang et al. 2009; Chankaew et al. 2011). Thus, mungbean is a valuable legume crop not only from an economic, but also from a nutritional standpoint.

Cultivated *Vigna* spp. are classified into three groups, two originating from Africa and one from Asia. A phylogenetic study with molecular markers showed that the African *Vigna* spp. were more diverse than the Asian *Vigna* spp., which evolved more recently (Fatokun et al. 1993; Lambrides and Godwin 2007). *V. radiata* has been divided into three subgroups: subspecies *radiata* (green grams and golden grams including the cultivated mungbean), subspecies *sublobata* and subspecies *glabra* (Verdcourt 1970). Archaeological studies have suggested that mungbean was domesticated in the Indian subcontinent about 3,500 years ago, and that cultivated mungbean spread to other parts of Asia and into North Africa during the early domestication process (Vishnu-Mittre 1974; Smartt 1984). Domestication and selection processes promoted the distribution of cultivated mungbean throughout Southern and Eastern Asia, Africa and Austronesia (Lambrides and Godwin 2007).

The cultivated *Vigna* species belongs to the subgenus *Ceratotropis* and is a self-pollinated diploid plant with $2n = 22$ chromosomes (Lambrides and Godwin 2007; Somta and Srinives 2007; Tangphatsornruang et al. 2009). Although mungbean chromosomes are difficult to study due to their small size, the genome size of mungbean was estimated as 1.2 pg/2C or 579 Mb by flow cytometric analysis (Arumuganathan and Earle 1991). Murray et al. (1979) estimated that 65 % of the nuclear DNA content consisted of single-copy sequences, and that 46 % of the single-copy mungbean DNA was interspersed within repetitive sequences. A restriction fragment length polymorphism (RFLP) study by Menancio-Hautea et al. (1993) suggested the duplication of the mungbean genome because duplicate markers were positioned at more than one linkage group.

Construction of a mungbean genetic map is an important step for the improvement of mungbean cultivars. Lambrides and Godwin (2007) summarized genetic linkage maps in the genus *Vigna* using mapping populations mainly derived from a cross between cultivated mungbean genotypes and wild progenitors. Humphry et al. (2002) constructed the most comprehensive mungbean map of a Berken × ACC41 population, and many mungbean mapping studies including this genetic map followed the first published mungbean map constructed by the University of Minnesota (Menancio-Hautea et al. 1993). The first generation of mungbean genetic maps was constructed with RFLP and random amplified polymorphic DNA (RAPD) markers (Somta and Srinives 2007). Although simple sequence repeat (SSR) or microsatellite markers have been developed recently, the number of published SSR markers is low because of the lack of polymorphisms in this species (Gwag et al. 2006; Somta et al. 2008; Tangphatsornruang et al. 2009). Therefore, instead of using SSR markers directly from mungbean, SSR markers generated from common bean, cowpea and Azuki bean, etc. could be used for comparative genomic studies (Chaitieng et al. 2006; Isemura et al. 2007; Lambrides and Godwin 2007; Somta and Srinives 2007; Datta and Gupta 2009; Chankaew et al. 2011).

Single nucleotide polymorphisms (SNPs) including insertions/deletions (indels) are very feasible for multiple applications such as increasing marker density, quantitative trait loci (QTL) mapping and high-throughput marker-assisted selection (MAS) (Rafalski 2002; Zhu et al. 2003; Choudhary et al. 2012). SNPs are not only abundant in plant and human genomes, but they can also lead to dramatic changes in phenotype when they are located within the exon (Van et al. 2004, 2012). There are currently no published reports describing mungbean SNP markers except for a recent transcriptome analysis that identified a significant number of mungbean SNPs (Moe et al. 2011). The rapid development of next-generation sequencing (NGS) instruments will facilitate genome-wide SNP discovery, and several companies are already working toward the production of longer read lengths and greater numbers of sequence reads. For example, with a run form at 2 × 150 bp, the Illumina HiSeq 2000 produces 1.13 Tb of output per run and 81 Gb daily (http://www.illumina.com/portfolio). Whole-genome or subgenome sequences of many plant species have recently become available with the development of NGS technologies (Feuillet et al. 2011; Van et al. 2011). This suggests that mungbean SNPs could be identified using NGS technologies and applied for crop improvement and breeding programs.

In this study, whole genomes of two mungbean cultivars, Sunhwanokdu and Gyeonggijaerae 5, were sequenced to develop SNPs using Illumina HiSeq and de novo assembly. More than 300,000 SNPs between these two mungbean cultivars were identified, and some sequence variations between Sunhwanokdu and Gyeonggijaerae 5 were closely examined to validate our SNP identification strategy. The genome-wide SNPs identified from the present study could be used for their potential application for mungbean improvement via QTL mapping, association mapping and MAS. Also, all of these SNPs will contribute to the development of the mungbean genetic map.

Materials and methods

Plant materials

Sunhwanokdu was first introduced as VC1973A, which was generated from a cross between Bangasa and EG-MG-16 by the Asian Vegetable Research and Development Center in 1980 (Hong et al. 1983). In 1983, Sunhwanokdu became a recommended mungbean cultivar in Korea after it was selected from the pure line ‘Suwon 3’ (Table 1). Gyeonggijaerae 5 is a local accession collected from the Gyeonggi Province, Korea (Table 1). Sunhwanokdu and Gyeonggijaerae 5 show different plant morphological phenotypes. The Sunhwanokdu hypocotyls are green, whereas Gyeonggijaerae 5 has purple hypocotyls. Seeds of Gyeonggijaerae 5 are easily shattered, but Sunhwanokdu seeds show moderate shatter rates (Hong et al. 1983). One of the most important agronomic traits of Sunhwanokdu is its high yield, and this variety is also fairly resistant to mungbean mottle virus (MMV) and powdery mildew (*Erysiphe polygoni*). However, Gyeonggijaerae 5 shows moderate to severe susceptibility to infection by MMV, *Cercospora* leaf spot and powdery mildew (Hong et al. 1983).

DNA isolation and whole-genome sequencing

After Sunhwanokdu and Gyeonggijaerae 5 were purified by self-pollinating for a few generations, genomic DNA was prepared from 2-week-old leaves of homozygous individuals from Sunhwanokdu and Gyeonggijaerae 5 using a cetyltrimethyl ammonium bromide method (Gelvin and Schilperoort 1995). Prior to being processed with the Illumina HiSeq Sequencing System (Illumina, Inc., San Diego, CA, USA), the extracted DNAs were treated with RNase A and proteinase K followed by ethanol precipitation. After the construction of sequencing libraries according to the manufacturer’s recommended protocols (Illumina), paired-end shotgun sequencing was performed with fragments of the library.

Sequence analysis and SNP discovery

After the production of raw sequences from Sunhwanokdu and Gyeonggijaerae 5 by Illumina HiSeq, sequences were trimmed based on quality, if the quality scores of the end sequences of the raw data were dramatically decreased to less than ten. These trimmed Sunhwanokdu sequences were assembled de novo using the ABySS 1.3.2 software (http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases/1.3.2).

Gyeonggijaerae 5 reads were aligned to contigs of Sunhwanokdu generated by ABySS 1.3.2 using the Burrows–Wheeler Aligner (BWA)’s short read aligner for Illumina reads (Li and Durbin 2009; http://bio-bwa.sourceforge.net) with a read quality threshold of 20 for end trimming. To remove false positives of SNPs derived from duplication and sequencing error, SNP calling was performed using SAMTools (Li et al. 2009; http://samtools.sourceforge.net/index.shtml) with the lowest and highest depths of 5 and 100, respectively. SNP filtering was performed with a Phred quality score cutoff of 100. Nucleotide diversity (θ) was calculated according to Van et al. (2004).

Validation of putative SNPs by Sanger sequencing

To validate the putative SNPs identified by bioinformatics, sequence variations between Sunhwanokdu and Gyeonggijaerae 5 were selected using the following procedures. Common bean ESTs (Phaseolus_vulgaris_release_

Table 1 List of mungbean cultivars used in this study

Cultivar	IT No.[a]	Origin	Purpose
Sunhwanokdu	IT144011	Korea	Reference mungbean genome and SNP discovery
Gyeonggijaerae 5	IT182269	Korea	Counterpart of Sunhwanokdu for SNP discovery
PI 381374	IT145300	Nigeria	SNP validation
Yellowgram	IT154085	Pakistan	SNP validation
EC15172	IT163201	USA	SNP validation
NPE28	IT163278	Netherlands	SNP validation
EC15043	IT182239	Turkey	SNP validation
Binh Khe D.X.	IT208075	Vietnam	SNP validation
Jangannokdu	IT216796	Korea	SNP validation

[a] All mungbean accessions were provided by the National Agrobiodiversity Center (http://www.genebank.go.kr), National Academy of Agricultural Science, Korea

1.fasta; http://plantta.tigr.org/, Childs et al. 2007) were mapped against the Sunhwanokdu contigs, using the est2-genome software (http://bioweb2.pasteur.fr/docs/EMBOSS/est2genome.html) with ≥80 % sequence similarity. The 1-kb regions harboring SNPs were retrieved with a Python script, and Primer3 (http://frodo.wi.mit.edu/) was used to design primers for targeting putative SNPs within these 1-kb regions.

In addition to Sunhwanokdu and Gyeonggijaerae 5, seven mungbean genotypes of various origins were selected for SNP validation (Table 1), and the genomic DNAs from these seven mungbean genotypes were prepared as described above. Genomic DNA from Sunhwanokdu was used to perform a gradient PCR to determine the optimal annealing temperature of each primer. Then, 10 μM each of forward and reverse primer, 2 μl of 10× reaction buffer [750 mM Tris–HCl (pH 8.5), 200 mM $(NH_4)_2SO_4$ and 25 mM $MgCl_2$], 1.6 μl of 2.5 μM dNTPs and 0.3 μl of 2 U *Taq* DNA polymerase (VIVAGEN, Sungnam, Korea) were added to the extracted genomic DNA in a total volume of 20 μl. The PTC-225 Peltier Thermal Cycler (MJ Research, Inc., Watertown, MA, USA) was used to perform PCR with the following conditions: an initial denaturation at 94 °C for 5 min, 30 cycles each of denaturation, annealing and extension at 94 °C for 30 s, primer annealing at 50–70 °C for 30 s, 72 °C for 1 min, final extension at 72 °C for 10 min and storage at 18 °C. The amplified PCR products were separated by gel electrophoresis on 1.0 % ethidium bromide-stained agarose gels. Those primer sets producing a single discrete amplicon with the DNA from Sunhwanokdu were used to amplify the genomic DNA from the other eight mungbean genotypes using the same conditions as described above with optimized primer annealing temperatures.

PCR products showing a single discrete band were confirmed on the agarose gel and purified using 2 U exonuclease I and 1 U shrimp alkaline phosphatase at 37 °C for 1 h and at 75 °C for 15 min. These purified fragments were used as templates in sequencing reactions with a BigDye Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA). The reaction mixture consisted of 0.2 μl of BigDye Terminator, 3.2 μM primer (one of the primers used in the PCR amplification), 5× reaction buffer [400 mM Tris–HCl (pH 9.0) and 10 mM $MgCl_2$] and 1 μl of the purified fragment as template DNA. The labeling reaction mixture was ethanol-precipitated and resuspended in 10 μl of water. Sequence analysis was conducted with an ABI 3730 DNA sequencer (Applied Biosystems, Foster City, CA, USA). To confirm the SNPs among the nine mungbean genotypes, ABI trace files were aligned, and sequence variations were identified using ABI Prism SeqScape Software version 2.5 (Applied Biosystems, Foster City, CA, USA).

Results

Mungbean whole-genome sequencing by NGS

Whole-genome sequencing of the mungbean cultivars Sunhwanokdu and Gyeonggijaerae 5 was performed using a shotgun paired-end library (500-bp insert) and Illumina HiSeq technology. More than 41.8 billion bp was sequenced from Sunhwanokdu, and more than 42.0 billion bp from Gyeonggijaerae 5 to a depth of 72× (Table 2). After evaluation of the sequencing qualities of the Illumina HiSeq reads, about 39.5 billion bp from both cultivars were used for further analysis (68 × of sequencing depth).

Generation of Sunhwanokdu contigs and mapping of Gyeonggijaerae 5 Illumina HiSeq reads to the Sunhwanokdu contigs

Because a reference genome for mungbean is not available, de novo assembly was performed with trimmed Illumina HiSeq reads from Sunhwanokdu using ABySS 1.3.2 (Fig. 1). Various sizes of Sunhwanokdu contigs (N50 = 9,958 bp) were produced with 33.14 % of GC content. From a total of 137,520 contigs longer than 200 bp, 13,066 contigs were longer than N50 (9,958 bp). The largest Sunhwanokdu contig was 92,050 bp. The total length of contigs over 200 bp was 464.6 Mb, which covered 80.24 % of the mungbean genome. Only contigs longer than 10 kb were used for further SNP analysis. Using BWA software, trimmed Gyeonggijaerae 5 reads were successfully mapped to Sunhwanokdu contigs over 10 kb (Fig. 1).

Detection and distribution of SNPs and indels between Sunhwanokdu and Gyeonggijaerae 5

The procedure used for SNP discovery is described in Fig. 1. After the alignment of Sunhwanokdu contigs longer than 10 kb with Gyeonggijaerae 5 Illumina HiSeq reads, all

Table 2 Summary of next-generation sequencing used for SNP discovery

	Sunhwanokdu	Gyeonggijaerae 5
Raw data		
Total produced bases (bp)	41,820, 704,200	42,033,722,200
Sequencing depth[a]	72.23	72.60
After trimming		
Total produced bases (bp)	39,493,182,736	39,549,766,450
Sequencing depth[a]	68.21	68.31

[a] Sequencing depths were estimated based on the 579-Mb genome size of *Vigna radiata* (Somta and Srinives 2007)

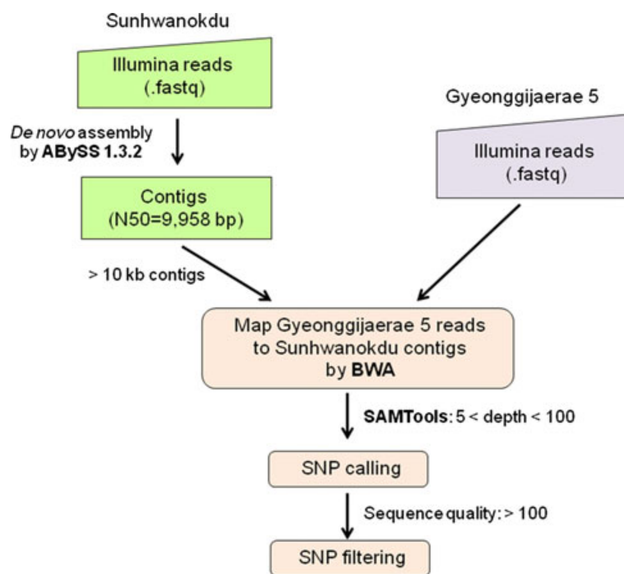

Fig. 1 Flow chart for mungbean SNP detection using Illumina Hiseq reads of *Vigna radiata* Sunhwanokdu as a reference

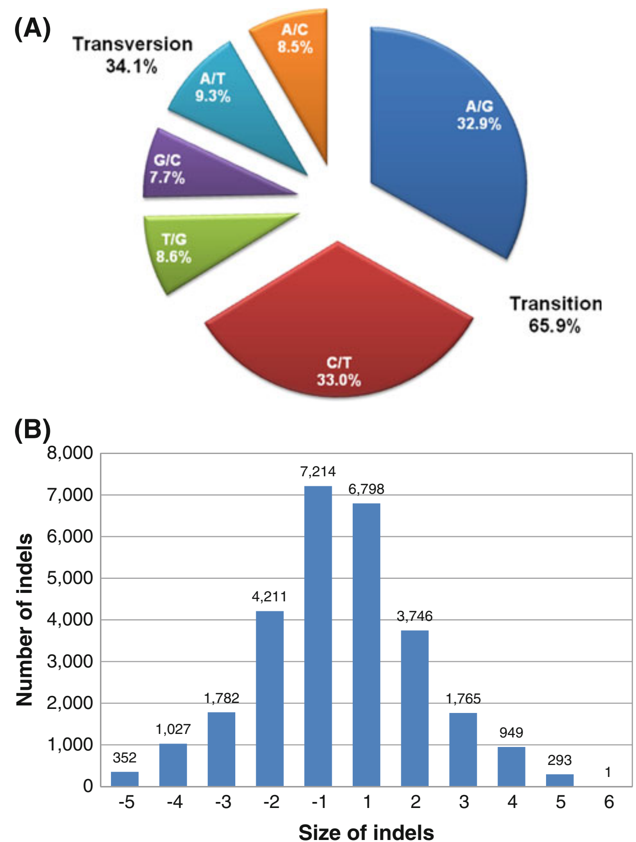

Fig. 2 Discovery of SNPs including indels of mungbean. **a** Patterns of single nucleotide substitutions between Sunhwanokdu and Gyeonggijaerae 5. **b** Distribution of the lengths of indels between Sunhwanokdu and Gyeonggijaerae 5. The sizes of the indels were determined based on Sunhwanokdu sequences as a reference. The *number* at the *top* of *each bar* indicates the number of indels

possible sequence variants including single nucleotide substitutions and indels were obtained using the SAMTools program. To decrease the number of false-positive SNPs and indels identified, the following filtering criteria were applied: minimum read depth of 5 and maximum read depth of 100 for the Sunhwanokdu contigs; sequencing quality of 100 (Fig. 1). Of 13,066 contigs (231.5 Mb) greater than 10 kb in length, 9,098 contigs yielded SNPs having a total length of 165,840 kb, and SNPs were positioned on 71.6 % of the sequences evaluated. A total of 305,504 single-base substitutions between Sunhwanokdu and Gyeonggijaerae 5 were detected (Supplementary Table 1). The six types of bi-allelic SNPs involved transitions (two possibilities, 65.9 %) and transversions (four possibilities, 34.1 %) (Fig. 2a). Of the 174,579 transitions detected, the numbers of A/G and C/T variants were 87,252 and 87,327, respectively, which accounted for 33 % of the total number of SNPs. Of the 90,422 transversions detected, 22,607 (8.5 %) were A/C, 24,599 (9.3 %) were A/T, 20,433 (7.7 %) were G/C and 22,783 (8.6 %) were T/G (Fig. 2a). The pattern of single nucleotide substitutions (transitions or transversions) in the remaining 40,503 SNP loci could not be determined because ABySS produced monogenic Sunhwanokdu sequences, and Gyeonggijaerae 5 appeared to be heterozygous.

Whole-genome NGS enabled the detection of indels of various sizes in Sunhwanokdu and Gyeonggijaerae 5. A total of 32,551 indels ranging from 1 to 6 bp were identified (Fig. 2b). Single and dinucleotide indels were the most frequent (78.1 %), and indels ≥3 bp were the second most frequent (21.9 %). Overall, 305,504 single-base changes and 32,551 indels were observed in the contigs longer than 10 kb, with an estimated detection rate of 1 SNP per 685 bp in the set of contigs analyzed. Considering the numbers of single-base changes and small indels, nucleotide diversity (θ) was estimated to be 0.00146 in the 231,500 kb of the mungbean sequences analyzed.

Validation of putative SNPs

To confirm the SNPs identified by our bioinformatics strategy, some single nucleotide substitutions located in genic regions were selected and validated by the traditional Sanger sequencing method. Because a limited number of mungbean transcripts (5,254 contigs) are currently available (Moe et al. 2011), in the present study, SNP confirmation was performed using 9,749 common bean ESTs. The common bean ESTs from various tissues were mapped onto a total of 13,066 Sunhwanokdu contigs longer than 10 kb using est2genome (Supplementary Fig. S1). Putative

Table 3 SNP validations by Sanger sequencing with Sunhwanokdu, Gyeonggijaerae 5 and seven other mungbean genotypes

SNP ID	Illunima HiSeq		Common bean EST accession No.	Description[a]	Sanger sequencing									Minor allele frequency
	Sunhwa-nokdu (Ref.)	Gyeonggi-jaerae 5 (Var.)			Sunhwa-nokdu	Gyeonggi-jaerae 5	PI 381374	Yellowgram	EC15172	NPE28	EC15043	Binh Khe D.X.	Jangan-nokdu	
>SNP_1	A	T	CV541385		A	T	A	T	T	T	A	T	A	0.44
>SNP_2	C	A	TA1119_3885	*Arabidopsis thaliana* genomic DNA, chromosome 3, TAC clone: K1 4A17-related cluster	C	C	C	C	C	C	C	C	C	NA
>SNP_3	G	A	CB542371	PsbP domain protein, putative-related cluster	G	A	G	A	A	G	G	A	A	0.44
>SNP_4	G	A	TA1163_3885	Putative phosphatase-related cluster	G	A	A	G	G	G	A	G	A	0.44
>SNP_5	T	C	TA1193_3885	Ser/Thr-specific protein phosphatase 2A B regulatory subunit beta isoform-related cluster	T	C	T	T	T	T	T	T	T	0.11
>SNP_6	A	T	TA483_3885	Hypothetical protein-related cluster	A	T	T	A	T	T	T	A	T	0.33
>SNP_7	T	C	TA501_3885	Inorganic pyrophosphatase-like protein-related cluster	T	C	T	T	C	T	T	T	C	0.33
>SNP_8	C	G	CV541793	PDR-type ABC transporter 1-related cluster	C	G	G	C	C	C	G	G	C	0.44
>SNP_9	C	T	TA1582_3885	Hypothetical protein At2g23140-related cluster	C	T	C	C	C	C	C	T	T	0.33
>SNP_10	C	A	TA960_3885	Aspartate aminotransferase isozyme 5-related cluster	C	A	C	C	C	C	C	A	C	0.22
>SNP_11	C	A	CB543573	F8M12.23 protein-related cluster	C	A	C	C	C	C	C	C	A	0.22
>SNP_12	C	T	TA2747_3885	F12A21.4-related cluster	C	T	C	C	C	C	C	C	T	0.22
>SNP_13	A	G	CV542171	FRO1-like protein-related cluster	A	G	A	A	G	G	G	A	A	0.44
>SNP_14	A	G	CV530630	Avr9/Cf-9 induced kinase 1-related cluster	A	G	G	G	G	G	A	G	A	0.33
>SNP_15	G	A	TA1514_3885	Ubiquitin-conjugating enzyme protein E2-related cluster	G	A	A	G	A	A	G	A	A	0.33
>SNP_16	A	T	CV530057	Hypothetical protein SEC2-related cluster	A	T	A	A	T	T	T	A	A	0.44
>SNP_17	T	C	TA2548_3885	Putative PAP-specific phosphatase, mitochondrial precursor (3′(2′), 5′-bisphosphate nucleotidase) (EC 3.1.3.7) (3′(2′), 5-bisphosphonucleoside 3′(2′)-phosphohydrolase)-related cluster	T	C	C	T	C	C	T	T	T	0.44
>SNP_18	A	G	CB556027		A	G	G	G	G	G	G	G	G	0.11
>SNP_19	A	C	CV541675	Hypothetical protein F28M20.240-related cluster	A	C	C	C	C	C	C	C	C	0.11
>SNP_20	T	C	CB542628	Beta-mannosidase enzyme-related cluster	T	C	T	C	C	C	T	C	T	0.44
>SNP_21	A	T	CV535508	F20N2.19-related cluster	A	T	T	A	A	A	A	T	A	0.33
>SNP_22	T	C	TA36_3885	Chloroplast nucleoid DNA binding protein, putative-related cluster	T	C	T	C	C	C	T	C	T	0.44
>SNP_23	G	T	CB539744	OSJNBa0011F23.7 protein-related cluster	G	T	G	G	G	G	G	G	T	0.22
>SNP_24	A	G	CV540955	Putative proteophosphoglycan-related cluster	A	G	G	A	G	G	G	G	A	0.33
>SNP_25	G	T	CV539366		G	T	G	G	G	G	G	G	T	0.22
>SNP_26	G	T	CV541599		G	T	G	G	G	G	G	G	G	0.11

Table 3 continued

SNP ID	Illunima HiSeq		Common bean EST accession No.	Description[a]	Sanger sequencing									Minor allele frequency
	Sunhwa-nokdu (Ref.)	Gyeonggi-jaerae 5 (Var.)			Sunhwa-nokdu	Gyeonggi-jaerae 5	PI 381374	Yellowgram	EC15172	NPE28	EC15043	Binh Khe D.X.	Jangan-nokdu	
>SNP_27	AA	TG	CV529800	Hypothetical protein-related cluster	AA	TG	AA	AA	AA	AA	AA	TG	TG	0.33
>SNP_28	G	C	CV530995	Hypothetical protein precursor-related cluster	G	C	C	C	G	C	C	C	C	0.22
>SNP_29	G	A	CV537004	Probable U6 snRNA-associated Sm-like protein LSm4-related cluster	G	A	A	G	A	A	A	G	G	0.44
>SNP_30	A	G	TA231_3885	60S ribosomal protein L18a-related cluster	A	G	G	G	G	G	G	G	G	0.11
>SNP_31	C	T	TA1135_3885	*Arabidopsis thaliana* genomic DNA, chromosome 5, TAC clone: K1 4A3-related cluster	C	T	C	C	C	C	C	C	C	0.11
>SNP_32	A	G	TA983_3885	Cluster related to UPI00004D6311	A	G	G	A	G	G	G	A	A	0.44
>SNP_33	G	T	CV539247		G	T	T	G	T	T	G	G	T	0.44
>SNP_34	A	G	TA708_3885	Putative finger transcription factor-related cluster	A	G	A	A	A	A	A	G	A	0.22
>SNP_35	T	C	TA276_3885	Drought-induced protein-related cluster	T	T	T	–	T	–	–	–	–	0.44
>SNP_36	C	G	CV535092	Amidase-like protein-related cluster	C	G	C	C	C	C	C	G	C	0.22
>SNP_37	T	C	CV529800	Hypothetical protein-related cluster	T	Multiple	T	T	T	T	Multiple	T	T	NA[b]
>SNP_38	T	C	TA860_3885	Epsilon1-COP-related cluster	T	C	T	T	T	T	T	C	C	0.33
>SNP_39	A	G	CB539173	Putative PTS protein-related cluster	A	G	A	G	G	A	A	G	G	0.44
>SNP_40	G	A	CV539204	Nuclear transcription factor Y subunit A-7-related cluster	G	A	G	G	A	A	G	A	G	0.44
>SNP_41	G	T	TA2260_3885	Putative aspartate aminotransferase-related cluster	G	T	G	G	T	G	G	G	G	0.22
>SNP_42	G	A	CV531982	Putative postsynaptic protein CRIPT-related cluster	G	A	A	G	G	G	G	G	A	0.33

NA not applicable, – deletion of 'T.'

[a] Arabidopsis proteins similar to common bean ESTs were described

mungbean exon regions were extracted based on at least 80 % similarity between the two sequences.

The results of in silico SNP discovery in mungbean were verified with a subset of 42 candidate SNPs (SNP_1 to SNP_42) randomly selected from 42 different contigs. The common bean EST accession numbers and the descriptions of these 42 different contigs are listed in Table 3. Supplementary Table S2 lists the partial sequences of the 42 contigs that were used for designing the primers; the putative SNPs were positioned around 500-bp flanking contig sequences. SNP flanking primers were designed with Primer3 (Supplementary Table S3), and the SNPs were confirmed by PCR amplification followed by Sanger sequencing (Supplementary Fig. S1). In addition to Sunhwanokdu and Gyeonggijaerae 5, seven additional mungbean cultivars from various origins were used for SNP validation (Table 1). The target mungbean DNA of all 42 randomly selected contig sequences was successfully amplified by PCR after melting temperatures were optimized by gradient PCR. Of these 42 putative SNPs, 39 (92.9 %) were validated (Table 3), indicating a fairly high SNP validation rate. SNP_35 and SNP_37 were miscalled, but it could not be determined whether these two SNPs were false positives. Although SNP_35 was not detected between Sunhwanokdu and Gyeonggijaerae 5 by Sanger sequencing, a 'T' deletion was identified when seven other mungbean genotypes were included for SNP validation. In the case of SNP_37, the Sanger method could not confirm the T/C SNP because multiple PCR bands were generated from not only Gyeonggijaerae 5, but also EC 15043.

Minor allele frequency (MAF) of each SNP locus is also listed in Table 3. SNPs with an MAF greater than 0.2 were selected as markers with normal allele frequencies (Lu et al. 2009). Of the 39 SNPs (excepting SNP_2, SNP_35 and SNP_37), six (15.4 %) SNP loci showed abnormal allele frequencies (0.11), in which only one mungbean genotype has a different single nucleotide at the position compared to the other eight genotypes. Nine (23.1 %) and ten (25.6 %) SNP loci had MAFs of 0.22 and 0.33, respectively. A total of 14 SNP loci (35.9 %) had MAFs of 0.44, indicating almost equal allele frequencies among the nine mungbean genotypes.

Discussion

Mungbean, as an important legume crop in terms of both nutrition and economics, has been researched extensively to develop MAS strategies for crop improvement. However, compared to other crops, molecular markers and saturated genetic maps of mungbean are less developed. NGS technology may enable the identification of whole-genome molecular markers for mungbean breeding programs. Although more than 30 crop species have been sequenced by the traditional Sanger method (BAC-by-BAC approach), NGS, or both, the whole-genome sequence of mungbean has not yet been completed. Therefore, the present study focused on the identification of putative mungbean SNPs using NGS for an SNP discovery pipeline and validation of some of the SNPs.

In addition to having the most abundance as genetic variation among individuals, SNPs are becoming more popular as genetic markers in various crops including maize and soybean because high-throughput genotyping methods have been developed (Mammadov et al. 2010). However, the recently developed technology, sequencing-by-synthesis (SBS), could not be applied to genome-wide and numerous SNP identifications in some crops due to genome complexity and the lack of a reference genome sequence. The rapid development of NGS technology and bioinformatics tools allows the collection of reference sequences of highly repetitive and duplicated genomes such as maize and soybean (Mammadov et al. 2010). Since mungbean has a non-duplicated genome about half the size of the soybean genome, genotyping using SBS technology combined with resequencing could be more feasible than with any other crop. Thus, the establishment of reference sequences is very important prior to genome-wide SNP discovery in the mungbean genome.

Although reference genome sequences for rice and maize are available, high-quality resequencings of various genetic resources within species are still required for future crop improvement. In addition, Xu et al. (2012) suggested that de novo sequencing of multiple maize genomes should be conducted to obtain a high-quality reference genome sequence. Resequencing of various accessions with a high-quality reference genome would be useful for studying domestication and improvement (Huang et al. 2010, 2012; Lai et al. 2010; Lam et al. 2010). Also, a high-resolution genetic map of a segregating population can be constructed by the resequencing of parents and a high-quality reference genome. Furthermore, NGS data from a resequenced population could be very useful because these sequences can fill gaps in the reference genome, and the information related to crop breeding program, such as genome-wide recombination frequency variation and segregation distortion, can be acquired (Xu et al. 2012). Thus, the de novo assembly of Sunhwanokdu and resequencing of Gyeonggijaerae 5 are very important for the establishment of a mungbean reference sequence and the construction of a high-resolution genetic map.

Errors occurring during the mapping of short reads to the reference genome and de novo assembly are typical sources of error in SNP discovery (You et al. 2011). Although the percentage of sequencing errors is less than 1 % in NGS platforms, sequencing errors are still a source

of false SNP calling. In particular, false calling is generated from indels in all three platforms, Illumina Solexa, SOLiD and Roche 454 (Harismendy et al. 2009). Huse et al. (2007) suggested that sequencing errors can be reduced if indels and homopolymers are filtered out. In the present study, of the high-quality SNPs that were aligned with mungbean putative exon regions, one SNP per Sunhwanokdu contig was selected for SNP validation (Fig. 1). The 42 SNP loci in 42 contigs were tested for SNP confirmation, and our validation rate (92.9 %) was higher than those reported previously (Barbazuk et al. 2007; Novaes et al. 2008; Deschamps et al. 2010; Hyten et al. 2010a, b). Of the 42 putative SNPs, only three were false positives. One false-positive SNP (SNP_35) might be miscalled by a long stretch of "T" repeats as a homopolymer in the process of contig assembly (Table 3). The other false-positive SNP (SNP_37) was not called correctly because Gyeonggijaerae 5 produced multiple PCR bands instead of one band containing a T/C base substitution (Table 3). The numbers of SNPs and false-positive SNP rates are dependent on SNP filtering criteria because read mapping and SNP calling are determined by bioinformatics tools in NGS platforms (You et al. 2011). Our stringent SNP criteria were sufficient to identify the appropriate numbers of SNPs and reduce the false-positive SNP rates. Furthermore, our SNP discovery pipeline is very successful and applicable for less-studied/orphan crop species.

Since single allele frequency has a large impact on linkage disequilibrium (LD) and leads to inaccurate estimations, any marker with an MAF less than 0.05 is eliminated from the genome-wide LD estimation and association mapping (Iqbal et al. 2012). Before selecting markers for a subset of SNPs, the MAF of each SNP locus was evaluated. The subset of SNP markers was selected for studying LD, association mapping, routine genetic diversity, mapping studies and SNP chip/array, based on SNP markers with MAF values greater than 0.200 (relatively frequent) (Lu et al. 2009; Iqbal et al. 2012; Peace et al. 2012; Semagn et al. 2012). Although nine mungbean genotypes of various origins were used for SNP validation, 33 of 39 confirmed SNP loci had MAFs higher than 0.200. Higher MAFs were observed by consequently increasing the number of SNP loci (Table 3). In particular, a total of 16 SNP loci showed MAFs of 0.44, indicating that these nine mungbean genotypes are highly diverse. Thus, the validated SNP loci along with our various mungbean accessions might be suitable for studying mungbean genetic diversity.

In this study, we were able to obtain large numbers of SNPs and indels between two mungbean cultivars, Sunhwanokdu and Gyeonggijaerae 5, using NGS technology. The large number of marker data accumulated in this study could be helpful for a mungbean breeding program. There are many advantages of using SNP markers in crop improvement; they are the most abundant form of genetic variations, and high-throughput automated systems for SNP genotyping like the Illumina GoldenGate SNP genotyping assay are available. Thus, SNPs are helpful for developing high-density genetic maps and for advancing generations by MAS (Shirasawa et al. 2010; Choudhary et al. 2012). A mungbean genetic map can be constructed from the large number of SNPs between Sunhwanokdu and Gyeonggijaerae 5. With various mungbean accessions in addition to these two mungbean cultivars, genetic diversity and high-resolution association and linkage mapping of target traits could be achieved for mungbean. Furthermore, the rapid development of high-throughput automation systems for SNP genotyping will enable the application of the identified SNPs and indels to QTL analysis and MAS for mungbean cultivar development. The present results suggest the potential for a wider application of comparative genomics and genetic diversity for crop improvement purposes.

Acknowledgments This research was supported by a grant from the Next Generation BioGreen 21 Program (Code No. PJ00811701), Rural Development Administration, Republic of Korea.

References

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51: 910–918

Chaitieng B, Kaga A, Tomooka N, Isemura T, Kuroda Y, Vaughan DA (2006) Development of a black gram [*Vigna mungo* (L.) Hepper] linkage map and its comparison with an azuki bean [*Vigna angularis* (Willd.) Ohwi and Ohashi] linkage map. Theor Appl Genet 113:1261–1269

Chankaew S, Somta P, Sorajjapinun W, Srinives P (2011) Quantitative trait loci mapping of Cercospora leaf spot resistance in mungbean, *Vigna radiata* (L.) Wilczek. Mol Breed 28:255–264

Childs KL, Hamilton JP, Zhu W, Cheung F, Wu H, Rabinowicz PD, Town CD, Buell CR, Chan AP (2007) The TIGR plant transcript assemblies database. Nucleic Acids Res 35(Database issue): D846–D851

Choudhary S, Gaur R, Gupta S, Bhatia S (2012) EST-derived genic molecular markers: development and utilization for generating an advanced transcript map of chickpea. Theor Appl Genet. doi: 10.1007/s00122-012-1800-3

Datta S, Gupta S (2009) Use of genomic resources in improvement of *Vigna* species. J Food Legum 22:1–10

Deschamps S, Rota ML, Ratashak JP, Biddle P, Thureen D, Farmer A, Luck S, Beatty M, Nagasawa N, Michael L (2010) Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina genome analyzer. Plant Genome 3:53–68

Fatokun CA, Danesh D, Young ND, Stewart EL (1993) Molecular taxonomic relationships in the genus *Vigna* based on RFLP analysis. Theor Appl Genet 86:97–104

Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K (2011) Crop genome sequencing: lessons and rationales. Trends Plant Sci 16:77–88

Gelvin SB, Schilperoort RA (1995) Plant molecular biology manual. Kluwer Academic Publishers, Norwell

Gwag JG, Chung JW, Chung HK, Lee JH, Ma KH, Dixit A, Park YJ, Cho EG, Kim TS, Lee SH (2006) Characterization of new microsatellite markers in mungbean, *Vigna radiata* (L.). Mol Ecol Notes 6:1132–1134

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10:R32

Hong EH, Lee YH, Kim SD, Hwang YH, Moon YH, Ham YS (1983) New disease resistant and high yielding mungbean variety "Seonhwanogdu". Res Rept ORD 25:178–180

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C, Fan D, Lu Y, Weng Q, Liu K, Zhou T, Jing Y, Si L, Dong G, Huang T, Lu T, Feng Q, Qian Q, Li J, Han B (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44:32–39

Humphry ME, Konduri V, Lambrides CJ, Magner T, McIntyre CL, Aitken EAB, Liu CJ (2002) Development of a mungbean (*Vigna radiata*) RFLP linkage map and its comparison with lablab (*Lablab purpureus*) reveals a high level of colinearity between the two genomes. Theor Appl Genet 105:160–166

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8:R143

Hyten DL, Cannon SB, Song Q, Weeks N, Ficks EW, Shoemaker RC, Specht JE, Farmer AD, May GD, Cregan PB (2010a) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics 11:38

Hyten DL, Song Q, Fickus EW, Quigley CV, Lim JS, Choi IY, Hwang EY, Pastor-Corrales M, Cregan PB (2010b) High-throughput SNP discovery and assay development in common bean. BMC Genomics 11:475

Imrie BC (1996) Mungbean. In: Hyde K (ed) The new rural industries. A handbook for farmers and investors. Rural Industries Research and Development Corporation, Canberra, pp 355–360

Iqbal MJ, Mamidi S, Ahsan R, Kianian SF, Coyne CJ, Hamama AA, Narina SS, Bhardwaj HL (2012) Population structure and linkage disequilibrium in *Lupinus albus* L. germplasm and its implication for association mapping. Theor Appl Genet 125:517–530

Isemura T, Kaga A, Konishi S, Ando T, Tomooka N, Han OK, Vaughan DA (2007) Genome dissection of traits related to domestication in azuki bean (*Vigna angularis*) and comparison with other warm-season legumes. Ann Bot 100:1053–1071

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, Jiao Y, Ni P, Zhang J, Li D, Guo X, Ye K, Jian M, Wang B, Zheng H, Liang H, Zhang X, Wang S, Chen S, Li J, Fu Y, Springer NM, Yang H, Wang J, Dai J, Schnable PS, Wang J (2010) Genome-wide patterns of genetic variation among elite maize inbreds. Nat Genet 42:1027–1030

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SSM, Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identified patterns of genetic diversity and selection. Nat Genet 12:1053–1059

Lambrides CJ, Godwin ID (2007) Mungbean. In: Kole C (ed) Genome mapping and molecular breeding in plants, vol 3., Pulses, sugar and tuber cropsSpringer, Germany, pp 69–90

Lawn RJ (1979) Agronomic studies on *Vigna* spp. in southeastern Queensland I. Phenological response of cultivars to sowing date. Aust J Agric Res 30:855–870

Li H, DurbIn R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) 1000 genome project data processing subgroup. The sequence alignment/map (SAM) format and SAM tools. Bioinformatics 25:2078–2079

Lu Y, Yan J, Guimaraes CT, Taba S, Hao Z, Gao S, Chen S, Li J, Zhang S, Vivek BS, Magorokosho C, Mugo S, Makumbi D, Parentoni SN, Shah T, Rong T, Crouch JH, Xu Y (2009) Molecualr characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theor Appl Genet 120:93–115

Mammadov JA, Chen W, Ren R, Pai R, Marchione W, Yalcin F, Witsenboer H, Greene TW, Thompson SA, Kumpatla SP (2010) Development of highly polymorphic SNP markers from the complexity reduced portion of maize [*Zea mays* L.] genome for use in marker-assisted breeding. Theor Appl Genet 121:577–588

Menancio-Hautea D, Fatokun CA, Kumar L, Danesh D, Young ND (1993) Comparative genome analysis of mungbean (*Vigna radiata* L. Wilczek) and cowpea (*V. unguiculata* L. Walpers) using RFLP mapping data. Theor Appl Genet 86:797–810

Moe KT, Chung J-W, Cho Y-I, Moon J-K, Ku J-H, Jung J-K, Lee J, Park Y-J (2011) Sequence information on simple sequence repeats and single nucleotide polymorphisms through transcriptome analysis of mungbean. J Integr Plant Biol 53:63–73

Murray MG, Palmer JD, Cuellar RE, Thompson WF (1979) Deoxyribonucleic acid sequence organization in the mungbean genome. Biochem 18:5259–5266

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9:312

Peace C, Bassil N, Main D, Ficklin S, Rosyara UR, Stegmeir T, Sebolt A, Gilmore B, Lawley C, Mockler TC, Bryant DW, Wilhelm L, Iezzoni A (2012) Development and evaluation of a genome-wide 6 K SNP array for diploid sweet cherry and tetraploid sour cherry. PLoS ONE 7:e48305

Rafalski A (2002) Application of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Semagn K, Magorokosho C, Vivek BS, Makumbi D, Beyene Y, Mugo S, Prasanna BM, Warburton ML (2012) Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. BMC Genomics 13:113

Shirasawa K, Isobe S, Hirakawa H, Asamizu E, Fukuoka H, Just D, Rothan C, Sasamoto S, Fujishiro T, Kishida Y, Kohara M, Tsuruoka H, Wada T, Nakamura Y, Sato S, Tabata S (2010) SNP discovery and linkage map construction in cultivated tomato. DNA Res 17:381–391

Smartt J (1984) Gene pools in grain legumes. Econ Bot 38:24–35

Somta P, Srinives P (2007) Genome research in mungbean (*Vigna radiata* (L.) Wilczek) and blackgram (*V. mungo* (L.) Hepper). ScienceAsia 33(Suppl 1):69–74

Somta P, Musch W, Kongsamai B, Chanprame S, Nakasathien S, Toojinda T, Sorajjapinun W, Seehalak W, Tragoonrung S, Srinives P (2008) New microsatellite markers isolated from mungbean (*Vigna radiata* (L.) Wilczek). Mol Ecol Res 8:1155–1157

Tangphatsornruang S, Somta P, Uthaipaisanwong P, Chanprasert J, Sangsrakru D, Seehalak W, Sommanas W, Tragoonrung S, Srinives P (2009) Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (*Vigna radiata* (L.) Wilczek). BMC Plant Biol 9:137

Van K, Hwang E-Y, Kim MY, Kim Y-H, Cho Y-I, Cregan PB, Lee S-H (2004) Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs. Euphytica 139: 147–157

Van K, Kim DH, Shin JH, Lee S-H (2011) Genomics of plant genetic resources: past, present and future. Plant Genet Resour 9:155–158

Van K, Kang YJ, Shim SR, Lee S-H (2012) Genome-wide scan of the soybean genome using degenerate oligonucleotide primed PCR: an example for studying large complex genome structure. Genes Genomics 34:467–474

Verdcourt B (1970) Studies in the Leguminosae–Papilionoideae for the flora of tropical East Africa. IV. Kew Bull 24:507–569

Vishnu-Mittre A (1974) Palaeobotanical evidence in India. In: Hutchison J (ed) Evolutionary studies in world crops. Cambridge University Press, Cambridge, pp 3–30

Xu Y, Lu Y, Xie C, Gao S, Wan J, Prasanna BM (2012) Whole-genome strategies for marker-assisted plant breeding. Mol Breed 29:833–854

You FM, Huo N, Deal KR, Gu YQ, Luo M-C, McGuire PE, Dvorak J, Anderson OD (2011) Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. BMC Genomics 12:59

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young N, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163: 1123–1134