

Research Article

# Sequence Information on Simple Sequence Repeats and Single Nucleotide Polymorphisms through Transcriptome Analysis of Mungbean

Kyaw Thu Moe<sup>1</sup>, Jong-Wook Chung<sup>1,2</sup>, Young-Il Cho<sup>1</sup>, Jung-Kyung Moon<sup>2</sup>, Ja-Hwan Ku<sup>2</sup>, Jin-Kyo Jung<sup>2</sup>, Jungran Lee<sup>2</sup> and Yong-Jin Park<sup>1,3\*</sup>

<sup>1</sup>Department of Plant Resources, College of Industrial Sciences, Kongju National University, Yesan 340-702, Republic of Korea

<sup>2</sup>The Rural Development Administration (RDA), Suwon 441-100, Republic of Korea

<sup>3</sup>Legume Bio-Resource Center of Green Manure (LBRGCM), Kongju National University, Yesan 340-702, Republic of Korea

\* Corresponding author

Tel: +82 41 330 1201; Fax: +82 41 330 1209; E-mail: [yjpark@kongju.ac.kr](mailto:yjpark@kongju.ac.kr)

Available online on 18 November 2010 at [www.jipb.net](http://www.jipb.net) and [www.wileyonlinelibrary.com/journal/jipb](http://www.wileyonlinelibrary.com/journal/jipb)

doi: 10.1111/j.1744-7909.2010.01012.x

## Abstract

Mungbean (*Vigna radiata* (L.) Wilczek) is a unique species in its ability to fix atmospheric nitrogen, with early maturity, and relatively good drought resistance. We used 454 sequencing technology for transcriptome sequencing. A total of 150 159 and 142 993 reads produced 5 254 and 6 374 large contigs ( $\geq 500$  bp) with an average length of 833 and 853 for Sunhwa and Jangan, respectively. Functional annotation to known sequences yielded 41.34% and 41.74% unigenes for Jangan and Sunhwa. A higher number of simple sequence repeat (SSR) motifs was identified in Jangan (1 630) compared with that of Sunhwa (1 334). A similar SSR distribution pattern was observed in both varieties. A total of 8 249 single nucleotide polymorphisms (SNPs) and indels with 2 098 high-confidence candidates were identified in the two mungbean varieties. The average distance between individual SNPs was  $\sim 860$  bp. Our report demonstrates the utility of transcriptomic data for implementing a functional annotation and development of genetic markers. We also provide large resource sequence data for mungbean improvement programs.

Moe KT, Chung JW, Cho YI, Moon JK, Ku JH, Jung JK, Lee J, Park YJ (2011) Sequence information on simple sequence repeats and single nucleotide polymorphisms through transcriptome analysis of mungbean. *J. Integr. Plant Biol.* 53(1), 63–73.

## Introduction

Pulses are important protein resources that help meet the nutritional requirements of developing countries. Among them, mungbean (*Vigna radiata* (L.) Wilczek) is one of the most widely cultivated species throughout the southern half of Asia including India, Pakistan, Bangladesh, Sri Lanka, Laos, Cambodia, South China, and Central Asia (particularly in the rain fed areas). It is adapted to short growth duration, low water requirements, and poor soil fertility. It is a self-pollinating diploid plant with  $2n = 2x = 22$  chromosomes (Menancio et al. 1993) and a genome size of 515 Mb/1C (Parida et al. 1990).

A transcriptome is the set of all RNA molecules, including mRNA, rRNA, tRNA, and non-coding RNA, produced in one cell or a population of cells. Although the analysis of relative mRNA expression levels might be complicated by the fact that relatively small changes in mRNA expression can produce large changes in the total amount of corresponding protein present in the cell, a number of organism-specific transcriptome databases have been constructed and annotated to aid in identifying genes that are differentially expressed in distinct cell populations or subtypes. Unlike genome analysis, transcriptome analysis offers a full profile of gene function information under various conditions, and it differs with dissimilar

environments, cell types, developmental stages, and cell states (Baerson et al. 2005; Dharmawardhana et al. 2010; Gohin et al. 2010). Parchman et al. (2010) presented that transcriptome or expressed sequence tags (EST) sequencing was an efficient way to generate functional genomic level data for non-model organisms.

Rapidly expanding repositories of highly informative genomic data have generated increasing interest in methods for predicting protein function and inferring biological networks. At present, several gene annotation schemes, such as gene ontology (Ashburner et al. 2000), MultiFun (Serres and Riley 2000), and Function Catalog (FunCat) (Ruepp et al. 2004), are available online or as stand-alone configurations. The FunCat annotation scheme consists of 28 main categories covering general features such as cellular transport, metabolism, and protein activity regulation.

Relatively-complete genome sequences with next generation sequencing technologies are now available (Ferguson et al. 2010). A 454 sequencer is a large-scale parallel pyrosequencing system, using a genome sequencer (GS) FLX titanium instrument, with the ability to sequence 400–600 million base pairs per run with 400–500 base-pair read lengths. Pyrosequencing is a powerful tool that has been increasingly applied in genome and functional expression analyses. The combination of long, accurate reads and high throughput makes 454 sequencing analysis on the FLX genome sequencer ideally suitable for detailed transcriptome investigations (Jarvie and Harkins 2008). This sequencing technology is a rapid and cost effective way to sequence gene-containing portions of the genome (Wicker et al. 2006). Transcriptome profiling using 454-pyrosequencing has been increasingly performing on a variety of purposes.

Microsatellites or simple sequence repeats (SSRs) are the markers of choice for crop improvement in many species, because they are reliable and easy to score (Gupta and Varshney 2000; Park et al. 2009). The SSRs are clusters of short tandem repeated nucleotide bases distributed throughout the genome. The SSR markers are co-dominant, multi-allelic, and only require a small amount of DNA for scoring. The previous method for developing SSR markers involved constructing an SSR-enriched library, cloning, and sequencing, (e.g. Yu et al. 2009) which is costly and labor intensive. Mungbean genomic research has lagged behind the other crop species due to a lack of polymorphic DNA markers (Tangphatsornruang et al. 2009). A limited number of polymorphic SSR markers have been published for mungbean. Therefore, developing and identifying polymorphisms of the SSR motifs of mungbean is an important requirement for mungbean development. For this reason, Gwag et al. (2006, 2010) developed polymorphic SSR markers and used for genotyping of mungbean. Moreover, SSR markers are a powerful tool for diversity analysis and linkage mapping, especially for resistance gene analysis. For instance,

a core collection of resistance associated with SSR for soybean cyst nematode (Ma et al. 2006), for rice Cheng et al. (2009), Zhao et al. (2009), Cui et al. (2010), and for Cymbidium Moe et al. (2010) have been employed.

Single nucleotide polymorphisms are the most frequently found variation in DNA (Brookes 1999; Galeano et al. 2009) and are valuable markers for high-throughput genetic mapping, analysis of genetic variation and association mapping studies in crop plants (Deleu et al. 2009). The stability of SNPs and the relative fidelity of their inheritance are higher than that of other marker systems (Gupta et al. 2001). Several methods have been described for SNP detection (Ganal et al. 2009), such as high-throughput sequencing technologies (Barbazuk et al. 2007) and EcoTILLING (Barkley et al. 2008). Kadaru et al. (2006) described the EcoTILLING protocol as a rapid and cost-effective method for discovering SNPs and conducting rice (*Oryza sativa* L.) genotyping. The discovery of SNP markers based on transcribed regions has become a common application in plants because of the large number of ESTs available in databases (Deleu et al. 2009), and EST-SNPs have been successfully mined from EST databases in non-model species such as tomato (Yamamoto et al. 2005), white spruce (Pavy et al. 2006), Atlantic salmon (Hayes et al. 2007), and catfish (Wang et al. 2008). However, SNP data for mungbean is presently very rare. Our recent work has focused on the analysis of transcriptomic functions and investigation of SSR and SNP markers. This result will support the first clear understanding of the transcriptomic functions in mungbean and we also provided our resource data for the purpose of crop improvement programs.

## Results

### 454 sequencing

A summary of the 454 sequencing data for the two mungbean varieties were presented in **Tables 1** and **2**. We also provided sequence data for all contigs of both varieties as supplementary file (Supporting Information files 1 and 2). Based on the GS FLX sequencer standard procedures, the mungbean transcriptome sequencing yielded 61.71 Mb and 60.68 Mb with 411 and 424 per read by 150 159 and 142 993 reads for Sunhwa and Jangan, respectively. The two sample reads were assembled separately. Assembly by the De Novo assembler produced a lower number of completely assembled (98 716) and a higher number of partially assembled (26 838), singleton (16 161), and repeats (25) in Jangan compared with those of Sunhwa. A total of 5 254 and 6 374 large contigs ( $\geq 500$  bp), with an average length of 833 and 853 were detected for Sunhwa and Jangan, respectively. The largest contig size varied from 2 885 (Sunhwa) to 4 738 (Jangan). The total bases for all 9 977 Sunhwa contigs ( $\geq 100$  bp) was 6.16 Mb and 7.76 Mb

**Table 1.** Summary of 454 sequencing after *de novo* assembly

Sample	Total no. reads	Total no. bases	Assembled	Partial	Singleton	Repeat
Sunhwa	150 159	61 712 977	106 750	25 736	16 136	5
Jangan	142 993	60 677 419	98 716	26 838	16 161	25

**Table 2.** Summary of 454 sequencing after *de novo* assembly and sequence cleaning

Sample	Large contig (Length $\geq$ 500 bp)							All contig (Length $\geq$ 100 bp)		Singletons after sequence cleaning	Total valid unigenes (contigs+ singletons)
	Contigs	Bases	ACZ <sup>a</sup>	N50 contig size <sup>b</sup>	Largest contig size	Q40 Plus bases <sup>c</sup>	%Q40 <sup>d</sup>	Contigs	Bases		
Sunhwa	5 254	4 376 680	833	891	2 885	4 173 210	95.35	9 977	6 164 100	10 835	20 812
Jangan	6 374	5 438 223	853	911	4 738	5 174 247	95.15	12 596	7 762 019	13 176	25 772

<sup>a</sup>Average contig size; <sup>b</sup>Contig size means that half of all bases reside in contigs of this size or longer. <sup>c</sup>The number of bases called that have a quality score of 40 or greater. <sup>d</sup>The percentage of bases called that have a quality score of 40 or greater.

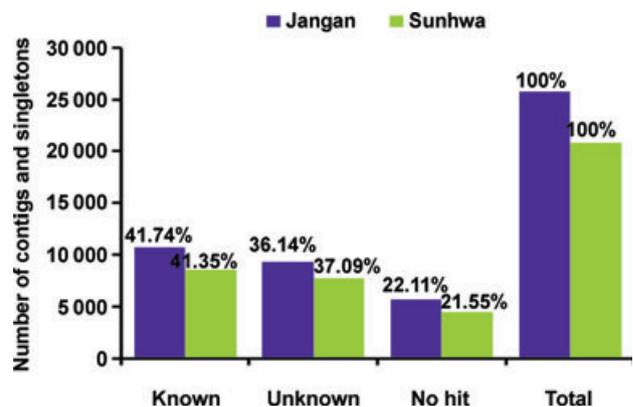
for the 12 596 Jangan contigs. All sequences resulted from the reference mapper were filtered using the SeqClean program to remove low-quality sequences and those shorter than 100 bp. The raw sequences were separately reduced, and resulted in 95.35% high-quality sequences with a quality score  $>40$  for Sunhwa and 95.15% for Jangan, respectively. Only a small portion of the raw reads were eliminated from the total reads due to a significant improvement in the entire sequencing process using the high performance 454 technology. The average length of all contigs was supported by the data from the large contigs, which were  $>500$  bp.

### FunCat schemes

The FunCat results are summarized in **Table 3**. Approximately 41% (8 606) of Sunhwa ESTs and 41.74% (10 758) of Jangan ESTs were matched to known functional sequences (**Figure 1**, **Table 3**). A cluster analysis method was performed to investigate correlations among the transcriptome profiles. Functional categories such as structural or catalytic proteins with binding function or cofactor requirements, subcellular localization, metabolism, protein fate, regulation of metabolism and protein function, and cellular transport were dominantly represented in the young-leaf mungbean transcriptome, whereas genes corresponding to energy, the cell cycle and DNA processing, transcription, protein synthesis, cellular communication/signal transduction mechanisms, cell rescue, defense and virulence, environmental interaction, systemic interaction with the environment, cell fate, systemic development, cellular component biogenesis, and organ differentiation represented a lower percentage.

Sunhwa-unigenes were categorized into 18 of 28 FunCat functional categories, whereas Jangan-unigenes were categorized into 17 categories. There was no relevant sequence

for organ differentiation in Jangan-unigenes. However, Jangan had a higher percent contribution in cell cycle and DNA processing (1.25  $>$  1.04), transcription (5.08  $>$  5.00), protein fate (9.73  $>$  9.65), proteins with binding functions or cofactor requirements (structural or catalytic) (35.99  $>$  35.10), regulation of metabolism and protein function (8.89  $>$  7.96), cellular communication/signal transduction mechanisms (1.79  $>$  1.56), environmental interaction (0.29  $>$  0.26), systemic interaction with the environment (0.57  $>$  0.47), and cell fate than did Sunhwa (0.90  $>$  0.67) (**Figure 2**, **Table 3**). In contrast, Sunhwa had a greater percent contribution value than that of Jangan for metabolism (12.28  $>$  11.05), energy (1.27  $>$  1.09), protein synthesis (3.25  $>$  2.78), cellular transport, transport facilities and transport routes (6.81  $>$  6.63), cell rescue, defense, and virulence (1.17  $>$  1.03), biogenesis of cellular components (0.77  $>$  0.69), organ differentiation (0.01  $>$  0), and subcellular localization (12.62  $>$  12.12) (**Figure 2**, **Table 3**). However,

**Figure 1.** A comparison of functional annotation result using BLAST program against Uniprot protein database.

**Table 3.** Summary of the functional classification catalog annotation results

Functional classification catalog		Sanhwa		Jangan	
		Quantity	Percentage contribution	Quantity	Percentage contribution
Known	e-value $\leq 1.00E-5$	8 606	41.35	10 758	41.74
Unknown	Match to unknown or unclassified function gene	7 720	37.09	9 315	36.14
No Hit	e-value $> 1.00E-5$	4 486	21.55	5 699	22.11
Total		20 812		25 772	
1	Metabolism	1 810	12.28	2 104	11.05
2	Energy	187	1.27	207	1.09
10	Cell cycle and DNA processing	153	1.04	238	1.25
11	Transcription	737	5.00	967	5.08
12	Protein synthesis	479	3.25	530	2.78
14	Protein fate (folding, modification, destination)	1 423	9.65	1 852	9.73
16	Protein with binding function or cofactor requirement (structural or catalytic)	5 174	35.10	6 852	35.99
18	Regulation of metabolism and protein function	1 174	7.96	1 692	8.89
20	Cellular transport, transport facilities and transport routes	1 004	6.81	1 262	6.63
30	Cellular communication/signal transduction mechanism	230	1.56	341	1.79
32	Cell rescue, defense and virulence	173	1.17	197	1.03
34	Interaction with the environment	38	0.26	56	0.29
36	Systemic interaction with the environment	69	0.47	109	0.57
40	Cell fate	99	0.67	171	0.90
41	Development (systemic)	16	0.11	21	0.11
42	Biogenesis of cellular components	114	0.77	131	0.69
47	Organ differentiation	1	0.01	0	0
70	Subcellular localization	1 861	12.62	2 308	12.12
Total		14 742		19 038	

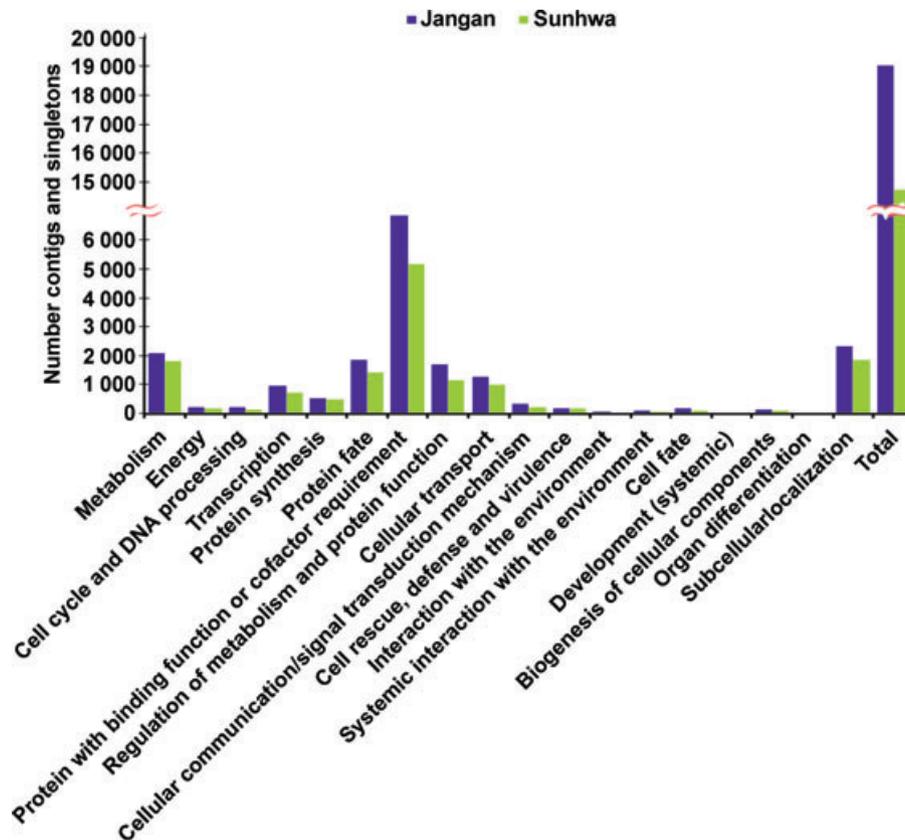
Jangan had a greater number of total bases, total contigs, quality contigs, contig size, and valid unigenes than had Sunhwa.

### SSR and SNP discovery

The SSR motifs found in Sunhwa are summarized in **Figure 3** and those for Jangan are summarized in **Figure 4**. A higher number of repeat motifs were identified in Jangan (1 630) than that of Sunhwa (1 334). Tri-nucleotide 743 (55.7%) and 915 (56.1%) type SSR motifs were most abundant, followed by di-nucleotides 448 (33.9%), 552 (33.9%), and others 143 (10.7%), and 163 (10.0%) in both Sunhwa and Jangan. Of the tri-nucleotide types, the GAA/AAG/AGA class dominated (226 and 284), whereas the GA/AG class dominated (255 and 377) in the di-nucleotide types in both Sunhwa and Jangan. The different features of the repeat motif types present in the transcriptome sequences indicated that some differences exist between these two mungbean varieties.

Other than the differences in the SSR repeat motifs, we identified variations in the SNPs in these two varieties

(**Table 4, Figures 5 and 6**). Assembly using GS Reference Mapper software revealed 8 249 SNP variations, which was supported by the 69 915 read count. Among all variations, the maximum value was an indel (1–94 nucleotide) (22.2%), with 141 (>6 nucleotide indel) included in it. It was then followed by C/T (11.53%), T/C (11.25%), G/A (10.49%), A/G (10.45%), others (5.9%), A/T (4.49%), T/A (4.16%), A/C (3.36%), C/A (3.33%), G/C (3.33%), G/T (3.31%), C/G (3.12%), T/G (2.86), and N/K (0.23%) (**Table 4**). The lowest percentage of the SNP type was N/K (0.23%). These results indicated that SNP variation could be involved, even in such undefined sequence positions. Four requirements were applied to the screening process to obtain highly confident differences for all of those variations: (i) a variation must be demonstrated by three or more non-duplicate reads; (ii) both forward and reverse reads should support the same variation; (iii) five or more reads with a quality score value greater than 20 must be presented in both of them; and (iv) the single nucleotide-indel should meet most of the reads aligned. Although these requirements undoubtedly reduced the sensitivity for detecting rare SNPs, they increased the specificity for true SNP detection by lowering the likelihood



**Figure 2.** A comparison of functional annotation result, which hits known functions.

of including false variants that arise due to sequencing errors. After screening with these criteria, 2 098 highly confident SNPs remained, which supported the total depth of 23 770. The ratio of the high-confident SNP type was similar to that of total SNPs, except indel, which distinctly reduced the percentage from 22.2% to 11.39%. It was clear that it was difficult to meet the criteria used in this screening process with the high number of indel nucleotide differences (1–94).

## Discussion

Experimentally uncharacterized genes have benefited from annotation and microarray expression analyses, which help generate biologically relevant clusters and improve functional predictions (Bainbridge et al. 2006). In this study, we compiled a transcriptome dataset for Sunhwa susceptible, and Jangan resistance varieties of mungbean using 454 GS FLX sequencing technology, and classified the data using a functional catalog scheme to create gene clusters based on functional roles. The FunCat classification scheme used attributes to assign membership to individual categories. Attribute selection and a description of the relationship of the categories were the key issues of the FunCat design (Ruepp et al. 2004). Some attributes

were well defined, computable, and highly selective, while others were only associative and descriptive (Ruepp et al. 2004). Our results showed that some of the universally expressed genes were undoubtedly housekeeping genes, whereas the molecular category of structure was overrepresented. Of the more than 20 000 valid unigenes in Sunhwa and more than 25 000 in Jangan, only 41.35% and 41.74%, respectively, could be classified with specific functions, whereas the remaining 58.64% and 58.25% unigenes were not annotated with definite functions, even though some homologous genes were obtained with the BLAST program based on nucleotide sequence similarity. Because the significance of sequence similarity depends, in part, on the length of the query sequence, many of the short sequencing reads obtained from next-generation sequencing cannot often be matched to known genes (Novaes et al. 2008). In addition, these genes that were not matched with functional annotations were probably composed of mutants from alternative splicing, novel gene products, or differentially expressed genes (Bainbridge et al. 2006). Our results demonstrated the utility of transcriptomic data to implement the functional annotation of new legume genome sequences.

A higher number of SSR motifs were observed in Jangan (1 630) than in Sunhwa (1 334), and a similar SSR distribution

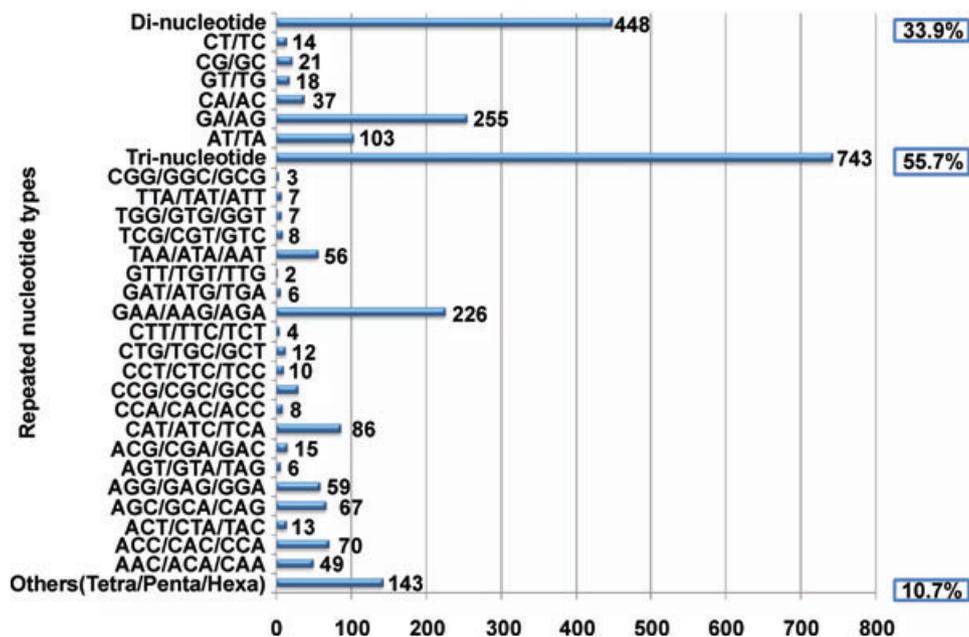


Figure 3. Distribution of simple sequence repeat (SSR) nucleotide classes among different nucleotide types found in the sequence of Sunhwa.

pattern was observed in both varieties. GAA/AAG/AGA was most concentrated among tri-nucleotide types, and GA/AG was most concentrated among di-nucleotide types. One strategy to discover SNPs is to prescreen loci for copy number and polymorphisms (Deynze et al. 2009). Transcriptome analysis provided a powerful tool for differential gene expression, mutant

splicing, SSR or SNP analysis, and functional genetics studies. However, to identify a highly confident SNP assay, SNPs with a coverage of more than 12 reads should be removed to eliminate paralogs (Matukumalli et al. 2009). In recent work, we employed four criteria to screen highly confident candidate SNPs. We discovered 8 249 SNP variations including 2 098

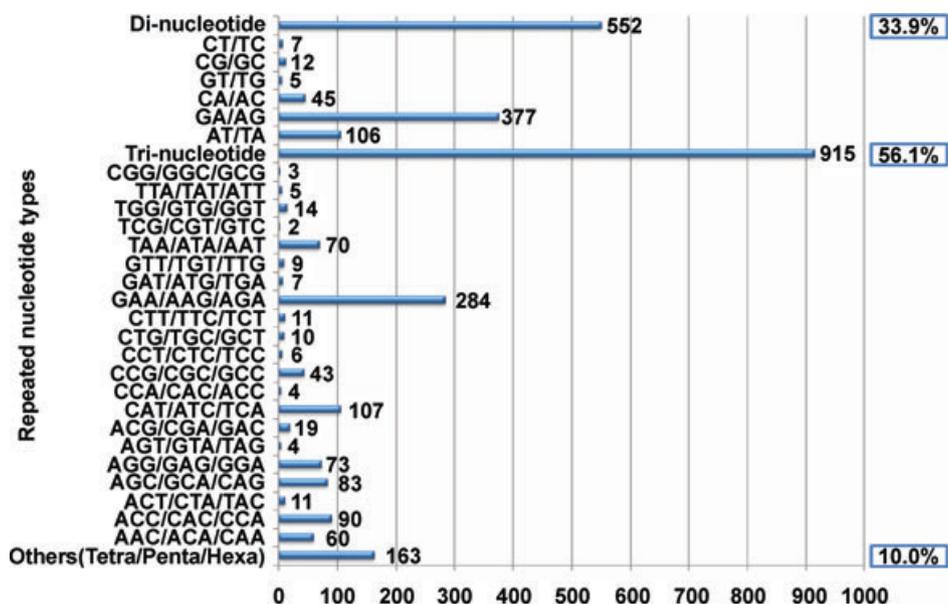


Figure 4. Distribution of simple sequence repeat (SSR) nucleotide classes among different nucleotide types found in the sequence of Jangan.

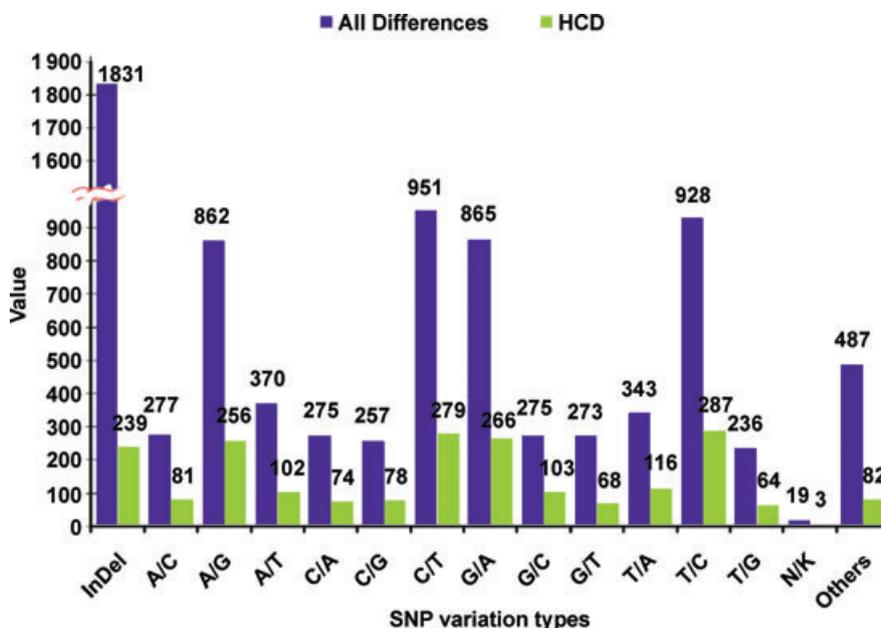
**Table 4.** Summary of single nucleotide polymorphism data from genome sequencer (GS) reference mapper and reduced for highly confident difference

SNP types	All differences		HCD		NOR for all differences		NOR for HCD	
	Number	%	Number	%	Number	%	Number	%
InDel	1 831	22.20	239	11.39	11 626	16.63	906	3.81
A/C	277	3.36	81	3.86	2 020	2.89	810	3.41
A/G	862	10.45	256	12.20	6 877	9.84	3 134	13.18
A/T	370	4.49	102	4.86	2 842	4.06	1 029	4.33
C/A	275	3.33	74	3.53	2 163	3.09	934	3.93
C/G	257	3.12	78	3.72	1 966	2.81	986	4.15
C/T	951	11.53	279	13.30	7 458	10.67	4 314	18.15
G/A	865	10.49	266	12.68	6 801	9.73	3 128	13.16
G/C	275	3.33	103	4.91	2 103	3.01	1 167	4.91
G/T	273	3.31	68	3.24	1 864	2.67	827	3.48
T/A	343	4.16	116	5.53	2 695	3.85	1 306	5.49
T/C	928	11.25	287	13.68	7 266	10.39	3 543	14.91
T/G	236	2.86	64	3.05	1 797	2.57	821	3.45
N/K	19	0.23	3	0.14	46	0.07	9	0.04
Others	487	5.90	82	3.91	8 391	12.00	856	3.60
<b>Total</b>	<b>8 249</b>		<b>2 098</b>		<b>69 915</b>		<b>23770</b>	

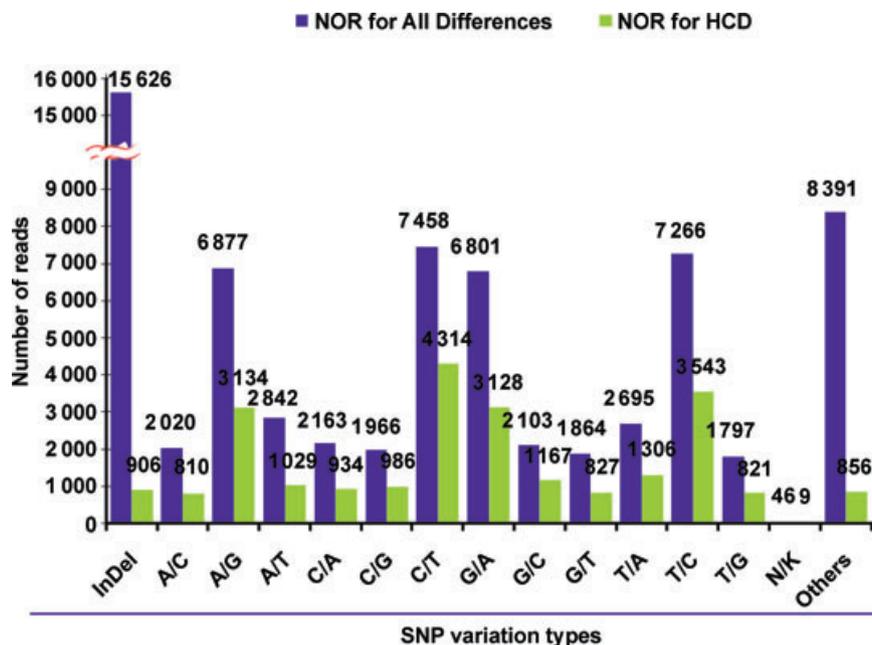
HCD, highly confident difference; K, any one of the four nucleotides (A,C,G,or T); NOR, number of reads; SNP, single nucleotide polymorphism.

high-confidence candidates. The average distance between individual SNPs was ~860 bp, whereas in the rice genome the average distance was estimated to be ~500 bp (Feltus et al. 2004), one SNP per 217 bp, and one in/del per 906 bp (Lee

et al. 2009). The success of genome-wide single nucleotide polymorphism detection analysis has resulted in a vast number of potential markers available for use in the construction of dense SNP maps (Gruber et al. 2002). There is a growing need



**Figure 5.** Number of single nucleotide polymorphism (SNP) variations for each SNP type found between Sunhwa and Jangan mungbeans varieties.



**Figure 6.** Number of reads assembled for each single nucleotide polymorphism (SNP) variation type between Sunhwa and Jangan mungbeans varieties.

to saturate the genetic map with SNPs to provide functional markers that are more amenable to high-throughput analysis, especially if the markers are located in gene-coding regions (Deleu et al. 2009). An alternative use of SNPs, for instance, SNP-based CAPS and dCAPS markers had been used in linkage mapping of mutant gene in garden pea (Li et al. 2010). Next-generation transcriptome sequencing will serve as a superior resource for developing polymorphic DNA markers, not only because of the enormous quantities of sequence data in which markers can be discovered, but also because the discovered markers are gene-based. Such markers are advantageous because they facilitate the detection of functional variation and selection in genomic scans or genetic association studies (Parchman et al. 2010). The large number of SSRs and SNPs that we detected will provide markers potentially useful for multiple applications ranging from population genetics, linkage mapping, and comparative genomics to gene-based association studies.

## Materials and Methods

### Plant materials

Two varieties of mungbean (*Vigna radiata* (L.) Wilczek) were selected from the National Institute of Horticultural and Herbal Science, Rural Development Administration: (i) Sunhwa, susceptible to stink bug (*Riptortus clavatus*) and adzuki bean weevil (*Callosobruchus chinensis*); and (ii) Jangan, resistant. The seedlings were cultivated in a glasshouse. The leaves of

young seedlings were used to extract the mRNA required for the synthesis of a cDNA library and for 454 sequencing.

### cDNA synthesis

The sequential steps of total RNA isolation, mRNA purification, cDNA synthesis, fragmentation by nebulization, and adaptor ligation were conducted prior to the 454 sequencing. Total RNA isolation was performed using a Trizol RNA isolation protocol (modified by D. Francis from Edgar Huitema) and the RNeasy Plant Mini kit (Qiagen, Valencia, CA, USA) following the manufacturer's manual. Young seedling leaves (100 mg) were placed in liquid nitrogen, ground into a powder, and subjected to total RNA extraction. Total RNA density was determined using a Biospec-Nano spectrophotometer (Shimadzu, Kyoto, Japan) and agarose gel electrophoresis. mRNAs were purified with the PolyAtract mRNA Isolation System (Promega, Madison, WI, USA). The purified products were used to synthesize the full-length cDNA using the ZAP-cDNA Synthesis kit (Stratagene, Santa Clara, CA, USA). Then the cDNA was fragmented by nebulization for library construction.

### Library preparation

A single-stranded template DNA library was generated to ensure the quality of the cDNA. The cDNA was fragmented by nebulization using an Agilent 2100 bioanalyzer (Waldbronn, Germany) with a mean fragment size of about 600 bp. Approximately 1 µg cDNA was used to generate a library for genome sequencing with an FLX Titanium analyzer (Roche, Mannheim,

Germany). The cDNA fragment ends were polished (blunted), and two short adapters were ligated to each end according to standard procedures (Margulies et al. 2005). The adapters provided priming sequences for amplification and sequencing of the sample library fragments as well as a “sequencing key,” which was a short sequence of four nucleotides used by the system software for base calling. The sequencing key also released the unbound strand of each fragment (with 5-adaptor A) following repair of any nicks in the double-stranded library. The quality of the single-stranded template DNA fragment library was assessed using the 2100 bioanalyzer, and the library was quantitated, including a functional quantitation to determine the optimal amount of the library to use as input for emulsion-based clonal amplification.

#### 454 sequencing

Single “effective” copies of template species from the DNA library to be sequenced were hybridized to DNA capture beads. The immobilized library was then re-suspended in an amplification solution, and the mixture was emulsified, followed by PCR amplification. After amplification, the DNA-carrying beads were recovered from the emulsion and enriched. The second strands of the amplification products were melted away, leaving the amplified single-stranded DNA library bound to the beads. The sequencing primer was then annealed to the immobilized amplified DNA templates. After amplification, a single DNA-carrying bead was placed into each well of a PicoTiterPlate (PTP) device. Simultaneous sequencing with multiple samples on a single PTP (4 region gasket) was used. The PTP was then inserted into the FLX genome titanium sequencer for pyrosequencing (Ronaghi 2001; Elahi and Ronaghi 2004), and sequencing reagents were sequentially flowed over the plate. Information from the PTP wells was captured simultaneously by a camera, and the images were processed in real time by an onboard computer. Multiplex identifiers (MIDs) were used to specifically tag unique samples in a GS FLX Titanium sequencing run, which were recognized by the GS FLX data analysis software after the sequencing run and provided high confidence in assigning an individual sequencing read to the correct sample.

After sequencing, sequence assembly was performed using the GS De Novo Assembler software to produce contigs and singletons. All sequence data were confirmed with references using GS Reference Mapper software. The resulting sequences were trimmed using SeqClean and the Lucy program.

#### FunCat annotation

All contigs and singletons (referred to as unigenes) resulting from 454 sequencing were analyzed using BLAST, to search the protein database for “non-redundant (NR)” and “Uniprot,”

respectively, and to obtain gene annotation accession numbers of related functions based on sequence similarity using the arbitrary expectation value of E-5. Functional assignment of the unigenes was classified by FunCat scheme version 2.1 on the Munich information center for protein sequences website (Ruepp et al. 2004).

#### SSR motifs and SNP variation

All sequences from the 454 sequencing were used to search for SSR motifs with the ARGOS 1.46 program at the default setting. The parameters were designed for identifying perfect di-, tri-, tetra-, penta-, and hexa-nucleotide motif types. The genome-wide SNP variation analysis used the new technology of the 454 genotyping assay. We detected the SNP and insertion/deletion (indel) data by aligning individual reads yielded by the sequencer, using the GS Reference Mapper software (Roche). This software automatically computes the alignment of reads from amplicon-based samples against a reference sequence and can detect low-frequency (<1%) variants. For a sequence difference to be declared a true polymorphism, at least two individual reads aligning to the consensus must have the variant allele and at least two others must have the consensus allele (Novaes et al. 2008). High-confidence variations were screened from all variations using the following criteria; a variation must be demonstrated by three or more non-duplicate reads, and both forward and reverse reads must support the same variation. Five or more reads with a quality score value greater than 20 must be present in both of them, and the single nucleotide-indel must meet most of the reads aligned.

#### Acknowledgements

This work was carried out with the support of the “Cooperative Research Program for Agriculture Science & Technology Development (Project No. 200908FHT020609001)” Rural Development Administration (RDA), Republic of Korea.

Received 1 Jul. 2010 Accepted 26 Oct. 2010

#### References

- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29.
- Baerson SR, Moreiras AS, Bonjoch NP, Schulz M, Kagan IA, Agarwal AK, Reigosa MJ, Duke SO (2005) Detoxification and

- transcriptome response in *Arabidopsis* seedlings exposed to the allelochemical benzoxazolin-2(3H)-one. *J. Biol. Chem.* **280**, 21867–21881.
- Bainbridge MN, Warren R, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, Mardis ER, Sadar MD, Siddiqui AS, Marra MA, Jones SJM** (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genom.* **7**, 246.
- Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS** (2007) SNP discovery via 454 transcriptome sequencing. *Plant J.* **51**, 910–918.
- Barkley NA, Wang ML, Gillaspie AG, Dean RE, Pederson GA, Jenkins TM** (2008) Discovering and verifying DNA polymorphisms in a mungbean [*V. radiata* (L.) R. Wilczek] collection by EcoTILLING and sequencing. *BMC Res. Notes* **1**, 28.
- Brookes AJ** (1999) The essence of SNPs. *Gene* **234**, 177–186.
- Cheng Y, Cho YI, Chung JW, Ma KH, Park YJ** (2009). Analysis of genetic diversity and population structure of rice cultivars from Africa, Asia, Europe, South America and Oceania using SSR markers. *Korean J. Crop Sci.* **54**, 441–451.
- Cui H, Moe KT, Chung JW, Cho YI, Lee GA, Park YJ** (2010). Genetic diversity and population structure of rice accessions from South Asia using SSR markers. *Korean J. Breed. Sci.* **42**, 11–12.
- Deleu W, Esteras C, Roig C, González-To M, Fernández-Silva I, Gonzalez-Ibeas D, Blanca J, Aranda MA, Arús P, Nuez F, Monforte AJ, Picó MB, Garcia-Mas J** (2009). A set of EST-SNPs for map saturation and cultivar identification in melon. *BMC Plant Biol.* **9**, 90
- Deynze AV, Stoffel K, Lee M, Wilkins TA, Kozik A, Cantrell RG, Yu JZ, Russel J, Kohel RJ, David M, Stelly DM** (2009) Sampling nucleotide diversity in cotton. *BMC Plant Biol.* **9**, 125.
- Dharmawardhana P, Brunner A, Strauss S** (2010). Genome-wide transcriptome analysis of the transition from primary to secondary stem development in *Populus trichocarpa*. *BMC Genom.* **11**, 150.
- Elahi E, Ronaghi M** (2004) Pyrosequencing: a tool for DNA sequencing analysis. *Methods Mol. Biol.* **255**, 211–219.
- Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH** (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Res.* **14**, 1812–1819.
- Ferguson BJ, Indrasumunar A, Hayashi S, Lin MH, Lin YH, Reid DE, Gresshoff PM** (2010) Molecular analysis of legume nodule development and autoregulation. *J. Integr. Plant Biol.* **52**, 61–76.
- Galeano CH, Gomez M, Rodriguez LM, Blair MW** (2009) CEL I nuclease digestion for SNP discovery and marker development in common bean (*Phaseolus vulgaris* L.). *Crop Sci.* **49**, 381–394.
- Ganal MW, Altmann T, Roder MS** (2009) SNP identification in crop plants. *Curr. Opin. Plant Biol.* **12**, 211–217.
- Gohin M, Bobe J, Chesnel F** (2010) Comparative transcriptomic analysis of follicle-enclosed oocyte maturational and developmental competence acquisition in two non-mammalian vertebrates. *BMC Genom.* **11**, 18.
- Gruber JD, Colligan PB, Wolford JK** (2002) Estimation of single nucleotide polymorphism allele frequency in DNA pools by using Pyrosequencing. *Hum. Genet.* **110**, 395–401.
- Gupta PK, Roy JK, Prasad M** (2001). Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr. Sci.* **80**, 4.
- Gupta PK, Varshney R** (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* **113**, 163–185.
- 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 mung bean, *Vigna radiata* (L.). *Mol. Ecol. Notes* **6**, 1132–1134.
- Gwag JG, Dixit A, Park YJ, Kyung-Ho Ma, Kwon SJ, Cho GT, Lee GA, Lee SY, Kang HK, Lee SH** (2010) Assessment of genetic diversity and population structure in mungbean (*Vigna radiata* L.). *Genes Genom.* **32**, 299–308.
- Hayes B, Lærdahl JK, Lien S, Moen T, Berg P, Hindar K, Davidson WS, Koop BF, Adzhubei A, Høyheim B** (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture* **265**, 82–90.
- Jarvie T, Harkins T** (2008) Transcriptome sequencing with the Genome Sequencer FLX system. *Nature Methods*, **5** Application note, Roche, 454 Sequencing.
- Kadaru SB, Yadav AS, Fjellstrom RG, Oard JH** (2006) Alternative ecotilling protocol for rapid, cost-effective single-nucleotide polymorphism discovery and genotyping in rice (*Oryza sativa* L.). *Plant Mol. Biol. Rep.* **24**, 3–22.
- Lee GA, Koh HJ, Chung HK, Dixit A, Chung JW, Ma KH, Lee SK, Lee JR, Lee GS, Gwag JG, Kim TS, Park YJ** (2009) Development of SNP-based CAPS and dCAPS markers in eight different genes involved in starch biosynthesis in rice. *Mol. Breeding* **24**, 93–101.
- Li X, Zhuang LL, Ambrose M, Rameau C, Hu XH, Yang J, Luo D** (2010) Genetic analysis of *ele* mutants and comparative mapping of *ele1* locus in the control of organ internal asymmetry in garden pea. *J. Integr. Plant Biol.* **52**, 528–535.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JK, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Eugene W, Myers EW, Nickerson E, Nobile JR** (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.
- Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, Connell JO, Moore SS, Smith TPL, Sonstegard TS, Tassell CPV** (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* **4**, e5350.

- Ma YS, Wang WH, Wang LX, Ma FM, Wang PW, Chang RZ, Qiu LJ** (2006) Genetic diversity of soybean and the establishment of a core collection focused on resistance to soybean cyst nematode. *J. Integr. Plant Biol.* **48**, 722–731.
- 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, Zhao W, Song HS, Kim YH, Chung JW, Cho Yi, Park PH, Park HS, Chae SC, Park YJ** (2010) Development of SSR markers to study diversity in the genus *Cymbidium*. *Biochem. Syst. Ecol.* **38**, 585–594.
- Novaes E, Drost D, Farmerie WG, Pappas GJ, Grattapaglia D, Ronald R, Sederoff RR, Kirst** (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genom.* **9**, 312.
- Parchman TL, Geist KS, Grahn JA, Craig W, Benkman CW, Buerkle CA** (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genom.* **11**, 180.
- Parida A, Raina SN, Narayan RKJ** (1990) Quantitative DNA variation between and within chromosome complements of *Vigna species* (Fabaceae). *Genetica* **82**, 125–133.
- Park YJ, Lee JK, Kim NS** (2009) Simple sequence repeat polymorphisms (SSRPs) for evaluation of molecular diversity and germplasm classification of minor crops. *Molecules* **14**, 4546–4569.
- Pavy N, Parsons LS, Paule C, MacKay J, Bousquet J** (2006) Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. *BMC Genom.* **7**, 174.
- Ronaghi M** (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res.* **11**, 3–11.
- Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkötter M, Werner Mewes H** (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* **18**, 5539–5545.
- Serres M, Riley HM** (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genom.* **5**, 205–222.
- Tangphatsornruang S, Somta P, Uthapaisanwong 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.
- Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridhivej B, Peatman E, Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N** (2008) 454 sequencing put to the test using the complex genome of barley. *BMC Genom.* **7**, 275.
- Yamamoto N, Tsugane T, Watanabe M, Yano K, Maeda F, Kuwata C, Toriki M, Ban Y, Nishimura S, Shibata D** (2005) Expressed sequence tags from the laboratory-grown miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cultivars. *Gene* **356**, 127–134.
- Yu JW, Dixit A, Ma KH, Chung JW, Park YJ** (2009) A study on relative abundance, composition and length variation of microsatellites in eighteen underutilized crop species. *Gen. Resour. Crop Evol.* **56**, 237–246.
- Zhao W, Chung JW, Ma KH, Kim TS, Kim SM, Shin DI, Kim CH, Koo HM, Park YJ** (2009) Analysis of genetic diversity and population structure of rice cultivars from Korea, China and Japan using SSR markers. *Genes Genom.* **31**, 283–292.

(Co-Editor: Kang Chong)

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

Supplementary file 1. All contigs for Sunhwa (fna file)

Supplementary file 2. All contigs for Jangan (fna file)

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.