Mol Breeding (2014) 33:497–501
DOI 10.1007/s11032-013-9962-5

SHORT COMMUNICATION

Genome-wide single nucleotide polymorphism discovery and validation in adzuki bean

Puji Lestari · Yang Jae Kang · Kwang-Soo Han · Jae-Gyun Gwag · Jung-Kyung Moon · Yong Hwan Kim · Yeong-Ho Lee · Suk-Ha Lee

Received: 20 May 2013 / Accepted: 18 September 2013 / Published online: 15 November 2013

Abstract Adzuki bean, also known as red bean (*Vigna angularis*), with $2n = 22$ chromosomes, is an important legume crop in East Asian countries, including China, Japan, and Korea. For single nucleotide polymorphism (SNP) discovery, we used *Vigna* accessions, *V. angularis* IT213134 and its wild relative *V. nakashimae* IT178530, because of the lack of DNA sequence polymorphism in the cultivated species. Short read sequences of IT213134 and IT178530 of approximately 37 billion and 35 billion bp were produced using the Illumina HiSeq 2000 system to a sequencing depth of 61.5× and 57.7×, respectively. After de novo assembly was carried out with trimmed HiSeq reads from IT213134, 98,441 contigs of various sizes were produced with N50 of 13,755 bp. Using Burrows–Wheeler Aligner software, trimmed short reads of *V. nakashimae* IT178530 were successfully mapped to IT213134 contigs. All sequence variations at the whole-genome level were examined between the two *Vigna* species. Of the 1,565,699 SNPs, 59.4 % were transitions and 40.6 % were transversions. A total of 213,758 SNPs, consisting of 122,327 non-synonymous and 91,431 synonymous SNPs, were identified in coding sequences. For SNP validation, 96 SNPs in the genic region were chosen from among IT213134 contigs longer than 10 kb. Of these 96 SNPs, 88 were confirmed by Sanger sequencing of 10 adzuki bean genotypes from various geographic origins as well as IT213134 and its wild relative IT178530. These genome-wide SNP markers will enrich the existing *Vigna* resources and, specifically,

Puji Lestari and Yang Jae Kang have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11032-013-9962-5) contains supplementary material, which is available to authorized users.

P. Lestari · K.-S. Han · Y.-H. Lee · S.-H. Lee (✉)
Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-921, Korea
e-mail: sukhalee@snu.ac.kr

P. Lestari
Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development, Jl. Tentara Pelajar No. 3A, Bogor 16111, Indonesia

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

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

Y. H. Kim
Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET), Anyang-si, Korea

could be of value for constructing a genetic map and evaluating the genetic diversity of adzuki bean.

Keywords Adzuki bean · De novo assembly · Resequencing · Single nucleotide polymorphism

Results and discussion

Adzuki bean (*Vigna angularis*), a self-pollinating diploid legume with $2n = 22$ chromosomes, has been an important pulse crop in East Asia for thousands of years. Adzuki bean is particularly popular in China, Japan, and Korea, but it is also consumed in Himalaya, Nepal, and India (Tomooka et al. 2002; Lumpkin and McClary 1994). Most East and Southeast Asian people use adzuki beans to make desserts because of their natural sweetness and nutritional value. In Korea, adzuki beans are the second most important legume crop after soybeans, and they are used in traditional Korean dishes such as adzuki porridge, adzuki rice cake, adzuki noodle, adzuki jelly, and shaved ice with adzuki bean. Adzuki beans are believed to have been domesticated from the wild species, *V. angularis* var. *nipponensis*. The site of domestication is still unknown, but it is reasonable to postulate that this took place in China, Japan, and Korea, where native adzuki beans grow widely. On the other hand, *V. nakashimae*, the other type of wild relative of *V. angularis*, can be found across east Asia and can be hybridized with *V. angularis*; it was suggested to be a good genetic resource for adzuki breeding.

Improvement of the adzuki bean crop has been focused on resistance to biotic and abiotic stresses, local adaptation, and seed yield. However, because of the relatively low polymorphism rate in *V. angularis*, a genetic map was constructed using a population interspecific between *V. angularis* and its wild relative, *V. nakashimae* (Kaga et al. 1996). This map was established with random amplification of polymorphic DNA and restriction fragment length polymorphism markers that were not amenable to high-throughput analysis. Identification of a large number of single nucleotide polymorphisms (SNPs), including indels, at the whole-genome level could be an important initial step in fine mapping of useful traits and in the application of a marker-assisted breeding strategy (Lai et al. 2012; Choudhary et al. 2012; Shirasawa et al. 2010).

Two *Vigna* accessions, IT213134 (*V. angularis*) and IT178530 (*V. nakashimae*), were selected for this study. IT213134 was selected from a cross between Dagokjosaeng, a Japanese cultivar, and SA8413-2, a Korean breeding line, and became a recommended adzuki bean cultivar, Kyeongwonpat, in Korea (Moon et al. 2003). *V. nakashimae* (IT178530) was collected at Yesan-Gun, Chungnam Province in Korea in 1993, and self-pollinated several times to reduce heterozygosity. *Vigna* accessions were provided by the National Agrobiodiversity Center (http://www.genebank.go.kr), National Academy of Agricultural Science, Korea. Genomic DNA from both genotypes was sequenced using the Illumina HiSeq 2000 system to produce the draft genome sequences and to develop genome-wide SNP/indel markers in the two species.

Whole-genome sequencing of IT213134 and IT178530 was performed using a shotgun, paired-end library of 500-bp insert size. A total of 37 billion bp of 100-bp paired-end reads were produced from *V. angularis* (IT213134) with a sequencing depth of 61.5×, and 35 billion bp of similar reads were obtained from *V. nakashimae* (IT178530), providing a sequencing depth of 57.7× (Supplementary Table 1). After evaluation of the quality of the Illumina HiSeq raw reads with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), de novo assembly was performed for Kyeongwonpat IT213134 using ABySS software with read quality cut-off of Q20 (Simpson et al. 2009), and was deposited as AUGG00000000. A total of 98,441 contigs were longer than 200 bp. The length of N50 was 13,755 bp, and the total length of IT213134 contigs over 200 bp was approximately 466 Mb, which covered 76.1 % of the adzuki genome of 612 Mb; this was estimated using flow cytometry (Table 1 and Supplementary Table 2). The flow cytometric analysis of nuclear DNA isolated from leaf tissue was conducted by the Benaro Research Institute (Seattle, WA, USA). The reads that were used in the assembly were re-aligned to the contigs over 10 kb to address the quality of assembly, and 71.33 % of reads were mapped while 59.96 % of reads were mapped with expected insert size (Supplementary Table 1). Genes were predicted for these contigs, without transcriptome support, by ab initio gene prediction using Geneid with gene prediction parameters of the common bean (Blanco and Abril 2009). A total of 44,185 genes was predicted from the contigs, and 11,508 were

Table 1 Assembly of IT213134 using ABySS software

	Size (bp)	Number of contigs[a]
N90	2,181	40,869
N70	7,494	18,739
N50	13,755	9,651
N30	22,188	4,259
N10	38,336	945
Total number of contigs	465,855,922	98,441
Total number of contigs over 10 kb	289,655,896	14,524

[a] Only contigs longer than 200 bp were included

Pfam-annotated using InterProScan5 software (Finn et al. 2010; Quevillon et al. 2005).

The raw reads from the wild relative, IT178530, were mapped to the IT213134 contigs that were over 10 kb in length using Burrows–Wheeler Aligner software applying reads = quality cut-off of Q20 (Supplementary File 1, Assembly.zip). The proportion of the mapped reads with expected insert size is half of the total mapped reads that would result from the structural variation including deletion and insertion between two distinct species (Supplementary table 1). All possible sequence variants, including single nucleotide substitutions and indels, were retrieved from the genomic regions with range of mapped reads of depth 5–100 and with quality score over 100 using the SAMTools program (Li et al. 2009; Li and Durbin 2009). Approximately 1.5 million SNPs between IT213134 and IT178530 were predicted (Supplementary File 2, SNP_context.zip). The proportions of transitions and transversions were 59.4 and 40.6 %, respectively. Of a total of 929,432 transitions, a similar proportion of bi-allelic types were detected, i.e., 463,560 for A/G and 465,872 for C/T. In 636,267 transversion mutations, each of four possibilities for base substitution (T/G, G/C, A/T, A/C) represented approximately 10 % of all SNPs (Table 2 and Supplementary Fig. 1A). Notably, the number of G/C substitutions was slightly lower than those of other substitutions; however, the number of non-synonymous changes due to the G/C substitutions was higher than for other transversion substitutions. Whole-genome next-generation sequencing (NGS) enabled the detection of indels of various sizes in IT213134 and IT178530. A total of 182,682 indels, ranging from 1 to 110 bp, were identified. IT178530 had 88,361 insertions, with a length of 489,488 bp, and 94,321 deletions, representing 305,600 bp. Single-nucleotide indels were the most abundant type of indel (54 %) among indels ranging in size from 1 to 6 bp (Supplementary Fig. 1B).

A total of 213,758 SNPs were detected in coding regions, corresponding to 34,467 genes, and the number of non-synonymous SNPs was 122,327. These non-synonymous SNPs in coding regions may directly represent functional changes in the corresponding genes, and can be used as functional markers if the corresponding gene affects a phenotypic trait. Using gene annotation with Pfam, we could retrieve the 17 nucleotide-binding site–leucine-rich repeat (NBS–LRR) genes that have been widely reported as resistance genes (PF00931 for the NBS domain, PF00560 and PF07725 for the LRR domain) (Kang

Table 2 Statistics for SNPs between IT213134 and IT178530

	Number of SNP	Coding region	Synonymous change	Non-synonymous change
Transition				
A/G	463,560	68,388	33,914	34,474
C/T	465,872	68,469	33,747	34,722
Total	929,432	136,857	67,661	69,196
Transversion				
T/G	161,835	19,701	6,058	13,643
G/C	149,942	19,472	4,477	14,995
A/T	162,354	18,293	7,334	10,959
A/C	162,136	19,435	5,901	13,534
Total	636,267	76,901	23,770	53,131
Total	1,565,699	213,758	91,431	122,327

Table 3 Number of synonymous/non-synonymous SNPs in putative NBS-LRR coding sequences

Gene name	Pfam annotation[a]	Number of synonymous changes	Number of non-synonymous changes
k_81_11	PF00560, PF00931, PF13504	8	10
k_167_2	PF00560, PF00560, PF00560, PF01582, PF00931, PF13504	2	16
k_2630_6	PF00931, PF00560, PF00560	5	2
k_2630_7	PF00560, PF00931, PF05659	9	3
k_4369_2	PF00481, PF00931, PF00931, PF00560, PF00560	8	17
k_6486_5	PF00931, PF01582, PF07725, PF13855	3	8
k_7046_12	PF00931, PF12799, PF00560	15	5
k_7696_4	PF00560, PF00560, PF00931	0	3
k_8117_1	PF00560, PF00560, PF00931	0	0
k_10755_1	PF00931, PF00560	9	16
k_10755_3	PF00560, PF00931	18	12
k_11220_5	PF00560, PF00931	3	7
k_11615_1	PF00931, PF00560	6	11
k_12465_1	PF00560, PF00931, PF01582, PF12799	0	0
k_13590_4	PF00560, PF00560, PF00931	0	0
k_13681_3	PF05659, PF00931, PF00560	0	0
k_13894_3	PF01582, PF00560, PF00931	0	1

[a] PF00931 for NBS domain, PF00560 and PF07725 for LRR domain

et al. 2012). The wild-type alleles with non-synonymous changes within these genes may be good resources for improving disease resistance of adzuki bean cultivars; for example, the non-synonymous changed LRR domains have the possibility to confer different recognition specificity against pathogens (Table 3, Supplementary Fig. 2).

The results of in silico SNP discovery in adzuki bean were validated with a subset of 96 candidate SNPs, randomly selected from 96 different contigs. The putative SNPs for validation were positioned within genic regions, including 1 kb each of upstream and downstream sequence. Primer3 (http://frodo.wi.mit.edu/) was used to design primers for targeting putative SNPs within 1-kb regions between IT213134 and IT178530 (Supplementary Table 3). In addition to IT213134 and IT178530, 10 adzuki bean accessions of various geographic origins were selected for SNP validation (Supplementary Table 4). The candidate SNPs were validated by polymerase chain reaction (PCR) amplification followed by Sanger sequencing (Supplementary Table 5). The target DNA of all 96 randomly selected contig sequences was successfully amplified by PCR after melting temperatures were optimized by gradient PCR. Of the 96 candidate SNPs, 88 (92 %) were validated (Supplementary Table 5), which is a fairly high SNP validation rate. It is interesting to note that only two of 88 validated SNPs showed polymorphism among the 10 adzuki bean genotypes, indicating the remarkably low allelic diversity of *V. angularis*. A single base 'A' deletion was found at SNP_36 in IT236588. Only one accession, IT236564, which had the same base as *V. nakashimae* IT178530, had a different single nucleotide 'A' at the SNP_79 position, where the other nine *V. angularis* accessions had a 'G'. The low allelic diversity among *V. angularis* accessions suggested the need for introgression of desirable traits from wild species, such as insect resistance in *V. nakashimae*, into the adzuki cultivar. Therefore, populations developed from the cross between *V. angularis* and its wild relative *V. nakashimae* may be good resources in an adzuki bean breeding program (Kaga et al. 1996).

In this study, we were able to develop large numbers of SNPs and indels between *V. angularis* and *V. nakashimae* using the Illumina HiSeq system and a shotgun, paired-end library of 500-bp insert size, although a reference genome sequence is not available for adzuki bean. NGS-based de novo assembly of a single genotype and one additional resequencing were quite effective for developing the large number of SNP/indel markers. The large amount of SNP marker data will be helpful for developing a high-density genetic map for an adzuki bean breeding program. Moreover, the draft adzuki bean genome sequence generated in this study will serve as a valuable genomic resource for comparative analysis with other legume crops.

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

Blanco E, Abril JF (2009) Computational gene annotation in new genome assemblies using GeneID. Methods Mol Biol 537:243–261

Choudhary S, Gaur R, Gupta S (2012) EST-derived genic molecular markers: development and utilization for generating an advanced transcript map of chickpea. Theor Appl Genet 124(8):1449–1462

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Res 38(Database issue):D211–D222

Kaga A, Ohnishi M, Ishii T, Kamijima O (1996) A genetic linkage map of azuki bean constructed with molecular and morphological markers using an interspecific population (*Vigna angularis* × *V. nakashimae*). Theor Appl Genet 93:658–663

Kang YJ, Kim KH, Shim S, Yoon MY, Sun S, Kim MY, Van K, Lee SH (2012) Genome-wide mapping of NBS–LRR genes and their association with disease resistance in soybean. BMC Plant Biol 12(1):139

Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, Hayden MJ, Forrest KL, Fleury D, Baumann U, Zander M, Mason AS, Batley J, Edwards D (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant Biotechnol J 10(6):743–749

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079

Lumpkin TA, McClary DC (1994) Adzuki bean: botany, production and uses. CAB International, Wallingford

Moon JK, Lee YH, Park KY, Yun HT, Kim SL, Ryu YH (2003) A new medium large and red seed coated adzuki bean cultivar 'Gyeongwon' with early maturing and lodging tolerance. RDA J Crop Res 4:200–204

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120

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(6):381–391

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123

Tomooka N, Vaughan D, Moss H, Maxted N (2002) The Asian *Vigna*: genus *Vigna* subgenus *ceratotropis* genetic resources. Kluwer, Dordrecht