TÜBİTAK

Turkish Journal of Agriculture and Forestry

http://journals.tubitak.gov.tr/agriculture/

Research Article

Turk J Agric For
(2015) 39: 459-469

doi:10.3906/tar-1409-3

Transcriptome characterization and large-scale identification of SSR/SNP markers in symbiotic nitrogen fixation crop faba bean (*Vicia faba* L.)

Sundan SURESH[1,2], Tae-Sung KIM[3], Sebastin RAVEENDAR[1], Joon-Hyeong CHO[4], Jung Yoon YI[1], Myung Chul LEE[1], Sok-Young LEE[1], Hyung-Jin BAEK[1], Gyu-Taek CHO[1], Jong-Wook CHUNG[1,*]

[1]National Agrobiodiversity Center, National Academy of Agricultural Science, Rural Development Administration, Jeonju, Republic of Korea

[2]Department of Botany, Directorate of Distance Education, Madurai Kamaraj University, Palkalai Nagar, Madurai, Tamil Nadu, India

[3]Department of Plant Resources, College of Industrial Science, Kongju National University, Yesan, Republic of Korea

[4]Department of Biological and Environmental Science, Dongguk University, Seoul, Republic of Korea

Received: 01.09.2014 • **Accepted:** 21.05.2015 • **Published Online:** 12.06.2015 • **Printed:** 30.06.2015

Abstract: We used 454 sequencing technology for faba bean transcriptome, which yielded 29.61 Mb sequence data. A total of 81,333 raw sequence reads were obtained and assembled by de novo sequence assembly. The contig distributions in three nonmutually exclusive gene ontology classifications and clusters of orthologous gene classes were assigned, which were also used to identify genes related to nitrogen fixation. Furthermore, a set of 1729 reads with simple sequence repeat (SSR) motifs were identified. Subsets of 55 SSR primer pairs were selected to validate SSR marker assay. Fifty-five primer pairs were used to amplify from one or more of the template genotypes, the single nucleotide polymorphism (SNP) types manifested high confidence differences, and 1946 SNPs, 145 insertion-deletions, and 110 variants were observed with more than one nucleotide among the detected SNPs. This study provides large-scale identification of the faba bean transcriptome resources for functional genomics studies and the development of molecular markers, which will certainly lead to crop improvement.

Key words: 454 pyrosequencing, gene ontology, molecular marker, transcriptome analysis

1. Introduction

Legumes (Fabaceae) constitute the third largest family of flowering plants and are vitally important for agriculture and the environment. Because of their symbiotic nitrogen fixation ability, legumes provide a substantial fraction of nitrogen and reduce the requirements for chemical fertilizers. The legumes are a highly diversified plant family, composed of nearly 18,000 species distributed in almost all terrestrial habitats (Doyle and Luckow, 2003). Numerous studies have been conducted in legumes, particularly focusing on their genome organization and evolution (Young et al., 2003; Zhu et al., 2005). Due to the lack of genetic and genomic resources, improvements in many food legume species have been hindered globally. In the early 1990s, *Medicago truncatula* Gaertn. and *Lotus japonicus* L. were selected as candidate model legumes due to their relatively small genome size and short life cycles (Young et al., 2005; Cannon et al., 2006).

Whole-genome sequencing has been undertaken in these model legume species, which delivers ways to identify putative orthologous gene sequence resources in other crop legume species (Varshney et al., 2009). Moreover, a draft genome sequence was completed with warm-season food legume soybean (*Glycine max* (L.) Merr.), which provides additional insights into comparative genomics within the family Fabaceae (Varshney et al., 2009). However, due to absence of genomic information for many crop legume species, effective utilization of genomic resources using molecular markers is being delayed. Therefore, there is a need to develop efficient molecular markers to enable identification of orthologous genes through genome synteny analysis (Varshney et al., 2009).

Vicia faba is the economically most important species in the genus *Vicia*. The world production of faba beans in 2010 was 4.3×10^6 t from 2.55×10^6 ha, which is relatively small when compared with soybean (262×10^6 t) and pea (10×10^6 t). Faba bean is not only important as a human dietary protein but is also used as animal feed around the world (Duc et al., 2010). Faba bean is diploid with 2n = 2x = 12 chromosomes (Sjödin, 1971) and partially cross-pollinated (4%–84%) (Bond, 1983), and it possesses the largest genomes among crop legumes

* Correspondence: jwchung73@korea.kr

(~13,000 Mb), which makes conservation of this genotype resource more expensive and difficult (Duc et al., 2010). However, despite its long cultivation history and economic importance, reports about faba bean diversity are limited, and few molecular markers are available (Pozarkova et al., 2002; Terzopoulos and Bebeli, 2008; Zeid et al., 2009). Developing more reliable and efficient molecular markers will facilitate diversity assessments and promote faba bean breeding. Owing to recent advances in sequencing technology, it becomes possible to generate large datasets rapidly with less time and labor.

Transcriptome analysis using next-generation sequencing is a powerful platform for scrutinizing such complex molecular mechanisms (Hafner et al., 2008). Transcriptomics offers a full profile for extracting gene functional information under various conditions. Unlike the genome, the transcriptome differs according to environment, cell type, developmental stage, and cell state (Gohin et al., 2010). Thus, transcriptomics provides important information on the temporal and spatial regulation of gene expression, assists in illuminating gene function and gene structure, and helps to determine the underlying molecular mechanisms involved in various biological processes.

The transcriptome itself is a precious resource for discovering and identifying polymorphic molecular markers, such as simple sequence repeats (SSRs), single nucleotide polymorphisms (SNPs), and insertion-deletions (InDels). Since the development of map-based strategies, which are employed to locate quantitative trait loci, molecular markers have played an increasingly important role in identifying genes that have serious impacts on agronomic traits (Michelmore et al., 1991) and have helped to clarify genetic diversity and population structures (Ma et al., 2010). Large-scale transcriptome sequencing is gaining momentum, particularly with economically important crops. Candidate genes for enzymes involved in biosynthesis of cyanogenic glucosides have been investigated using 454 pyrosequencing (Zagrobelny et al., 2009). Pyrosequencing is a useful tool for discovering novel transcripts, such as genes of unknown function, sequences with high-quality base discrepancies, and alternative splice variants (Cheung et al., 2008). Clustering is usually employed to investigate differential expression, promoter activity, or expression patterns and to denote genes that are categorized based on similar function or expression pattern. Several gene annotation schemes are available online or as stand-alone applications, including Function Catalogue (Ruepp et al., 2004), Gene Ontology (GO; Ashburner et al., 2000; Riley, 1993), and MultiFun (Serres and Riley, 2000).

Many reports have been published recently on massively parallel transcriptome sequencing (Wicker et al., 2006) and genome assembly using draft genomes of model organisms (Weber et al., 2007; Kristiansson et al., 2009). Furthermore, de novo assembly of transcriptome data has also been reported for organisms with no prior genomic resource development (Meyer et al., 2009). Whole-genome sequencing strategy is still required for some plant species in order to access their genomic resources, and careful decisions must be made to formulate efficient sequencing strategies, particularly for plant species which has large genome (Wheat, 2010).

Expressed sequence tags (ESTs) are a rapid and cost-effective method to analyze transcribed portions of the genome. EST sequencing was reported for gene discovery, especially for finding gene family structure, for expression analyses, and for determination of phylogenetic relationships using molecular markers such as SSRs and SNPs (Weber et al., 2007). Hence, transcriptome sequencing analysis provides the opportunity to discover molecular markers and identify differentially expressed genes in faba bean, where only a limited number of SNPs and SSRs are available (Hafner et al., 2008; Terzopoulos and Bebeli, 2008; Kaur et al., 2012). This study was aimed to define transcriptome generation, de novo assembly, and gene annotation using cDNA samples from faba bean. Moreover, clustering and annotation were carried out to generate a unigene set which was used to discover SSR and SNP markers.

2. Materials and methods

2.1. Plant material

Faba bean seeds were collected from the National Agrobiodiversity Center, Rural Development Administration, Republic of Korea. Seeds were germinated and grown in a greenhouse, and leaves from single young seedlings were used to extract the mRNA required to synthesize a cDNA library and for 454 sequencing.

2.2. cDNA preparation

Total RNA was extracted from *V. faba* leaves using an RNeasy Plant Mini Kit (QIAGEN, Valencia, CA, USA). All of the RNA extraction, mRNA isolation, and cDNA synthesis was performed according to the manufacturer's instructions and a previously published report (Chung et al., 2013). Finally, cDNA was fragmented by nebulization for library construction.

2.3. Library preparation

Approximately 1 µg of cDNA was used to generate a DNA library for sequencing with the Genome Sequencer (GS)-FLX Titanium System (Roche, Mannheim, Germany). The cDNA fragment ends were polished (blunted) and two different short adapters were ligated according to manufacturer instructions (Roche Diagnostics). The adapters, along with a four-nucleotide short sequencing

key for base calling, provided priming of the sequences for both the amplification and sequencing of the sample library fragments. Following repair of any nicks in the double-stranded library, the unbound strand of each fragment was released with 5-Adaptor A. Finally, the quality of this single-stranded template DNA library was assessed using a 2100 BioAnalyzer (Agilent, Waldbronn, Germany). The library was quantified to determine the optimal amount of the library needed as input for emulsion-based clonal amplification.

2.4. 454 pyrosequencing

Single effective copies of templates from the DNA library were used for 454 pyrosequencing. The templates from the DNA library were immobilized with DNA capture beads and emulsified with amplification solution followed by polymerase chain reaction (PCR) amplification. The DNA carrying beads were recovered from the emulsion and enriched after amplification. The amplified products were melted and 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. The PTP was then inserted into the FLX Genome Titanium Sequencer for pyrosequencing (Ronaghi, 2001; Elahi and Ronaghi, 2004). Multiplex identifiers were used to specifically tag unique samples in a GS FLX Titanium sequencing run, which were recognized by the GS data analysis software after the sequencing run and provided high confidence for assigning individual sequencing reads to the correct sample.

2.5. Functional analysis of the transcriptome

De novo sequence assembly was performed to identify all contiguous sequences (contigs) and singletons using GS De Novo Assembler software (newbler v 2.5.3), a tool for processing larger, more complex genomes and transcriptomes. The resulting sequences were trimmed using the SeqClean (http://sourceforge.net/projects/seqclean/) and Lucy (http://sourceforge.net/projects/seqclean/) programs. Next, BLAST was used to search the mRNA sequences of all contigs and singletons (i.e. unigenes) against the NCBI nonredundant and UniProt databases, respectively, using an arbitrary expectation value of E^{-5}. The aim of this procedure was to obtain gene accession numbers and associated annotation information based on sequence similarity. Unigenes were then functionally classified using FunCat (version 2.1), available at the Munich Information Center for Protein Sequences (MIPS) website (http://mips.gsf.de/projects/funcat), to evaluate potential gene functions expressed in *V. faba*. GO terms were also assigned to the set of unigenes that showed hits against the *Arabidopsis thaliana* database using the "Gene Ontology at TAIR" tool. Additionally, a BLASTx search against the 10 TAIR version databases (http://www.arabidopsis.org/) was performed with an E-value threshold of $<10^{-5}$ To annotate the function of the faba bean unigenes more specifically, we performed cluster of orthologous group (COG) analysis wherein we BLASTed faba bean unigenes against the COG database (cutoff, E^{-5}).

2.6. Discovery of SSRs and SNPs, and validation of SSR markers

The faba unigene set used to mine SSR and SNP motifs was identified using the ARGOS pipeline (version 1.46) with default settings to survey the molecular markers present in the *V. faba* accessions (Kim, 2004). Parameters were designed for identifying perfect di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of six repeats. To validate the faba SSRs, primers flanking the SSR sequences were designed as follows: length range, 18–23 nucleotides with 21 as optimum; PCR product size range, 100–400 bp; optimum annealing temperature, 55 °C; and GC content, 40%–60% with 50% as optimum. Faba bean genomic DNA was extracted from 18 diverse faba bean accession samples (Accession No. IT228628) for SSR marker validation using a DNeasy Plant Mini Kit (QIAGEN) according to the manufacturer's instructions. Fresh leaf tissue from each accession was used for each extraction and ground well using liquid nitrogen. DNA was resuspended in 100 µL of water, and dilutions were made to 10 ng/µL followed by storage at either –20 °C or –80 °C. Randomly selected SSR primer pairs were validated experimentally, and forward primers were synthesized by adding the M13 sequence to enable fluorescent tail addition through the PCR amplification process (Riley, 1993). PCR conditions included a hot-start at 95 °C for 10 min; followed by 10 cycles at 94 °C for 30 s, 60–50 °C for 30 s, and 72 °C for 30 s; followed by 25 cycles at 94 °C for 30 s, 50 °C for 30 s, and 72 °C for 30 s; and with a final elongation step of 72 °C for 10 min. PCR products were separated and visualized using the QIAxcel Gel Electrophoresis System (QIAGEN). The amplification intensity for individual markers was determined on an ABI Prism 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) according to manufacturers' instructions, after adding the ABI GeneScan LIZ500 size standard and amplification product sizes determined using GeneMapper v3.7 software (Applied Biosystems).

Analyses of genome-wide SNP variations were conducted by 454 genotype assay. We detected the SNP and InDel data by aligning individual reads yielded by the sequencer, using GS Reference Mapper software. This software automatically computes the read alignments from amplicon-based samples against a reference sequence and detects low-frequency (1%) variants. At least two individual reads aligning to the consensus must have the variant allele to be declared truly polymorphic (Novaes et al., 2008). High confidence variation was screened from all

variations using the following criteria: a variation must be demonstrated by three or more nonduplicate reads, and both forward and reverse reads must support the same variation. Five or more reads with a quality score value of >20 must be present in both, and the single nucleotide InDel must meet most of the reads aligned.

3. Results

3.1. 454 sequencing

A summary of 454 sequencing data and the following sequence assembly analyses for *V. faba* are presented in Table 1. The *V. faba* transcriptome sequencing by using the 454 GS FLX platform yielded 81,333 raw sequencing reads (29.61 Mb). Raw data from the 454 sequencing run were submitted to the National Center for Biotechnology Information (NCBI) Short Read Archive and can be retrieved as accession SRP043650. The sample reads were assembled separately using the De Novo Assembler, which produced 50,632 completely assembled, 3985 partially assembled, 25,231 singletons, and 18 repeats (Table 1). The overall sequence assembly resulted in 1379 isogroups and the largest contigs count in the isogroup was 20. Conversely, the largest isotig count in the isogroup was 12 and the number of isogroups assembled by one isotig was 1269. This was analogous to an individual transcript of isotigs assembled with 1379 isogroups, with an average count of contigs in the isotigs of 1.257 (Table 1). The largest contig count in the isotigs was 7. Similarly, the number of isotigs assembled by one contig was 1272, and the total number of bases in the isotigs was 1,016,846, with an average isotig size of 663.738 (Table 1). Moreover, after primary trimming little singletons were removed using the SeqClean and Lucy programs, ensuing the rules of minimum length requirement (100 bp). All sequences resulting from the SeqClean program to remove the low-quality sequences were separately reduced. SeqClean resulted in 23,263 sequences and 23,124 sequences after obtaining results from Lucy. The total number of valid unigenes after quality filtering was 26,497 (Table 1).

Table 1. Summary of *Vicia faba* 454 pyrosequencing data.

454 pyrosequencing term	
Reads	
Number of raw sequencing reads (n)	81,333
Bases of raw sequencing reads (bp)	29,618,632
Assembled	50,632
Partial	3985
Singleton	25,231
Repeat	18
Read by GS De Novo	671
Isogroups	
Number of isogroups	1379
Average contig count	1.257
Largest contig count	20
Number with one contig	1266
Largest isotig count	12
Number with one isotig	1269
Isotigs	
Number of isotigs	1532
Bases of isotigs	1,016,846
Average contig count	1.257180
Largest contig count	7
Number with one contig	1272
Number of bases	1,016,846
Average isotig size	663.738
N50 isotig size	695
Largest isotig size	3389
Singleton	
After SeqClean (minimum length)	23,263
After Lucky (minimum length)	23,124
Total number of unigenes	26,497

3.2. Functional classification (FunCat) of expressed gene sequences

The large numbers of unigenes were assigned to a wide range of GO categories and COG classifications. Different numbers and percentages of annotated unigenes were distributed under different GO categories, with 28% (4642) of unigenes for biological processes, 36% (5911) of unigenes for cellular components, and 36% (5886) of unigenes for molecular functions, which indicated the high accuracy of the annotation. The GO annotation analyses provided the faba unigenes with a diverse set of putative biological functions in each GO category. Among the ones that were classified, in the biological process category, the most abundant GO function or term was metabolic processes (34%), followed by responses to stimuli (25%) (Figure 1A). GO classification of the biological process category was also used to identify genes related to stress responses. In the cellular component category, a cell part was the most abundant (73%), followed by organelle (19%) (Figure 1B). Catalytic activity was the most abundant category (50%) in the molecular function category (Figure 2), followed by binding (35%). The three groups (biological processes, cellular components, and molecular functions) are not mutually exclusive; therefore, some contigs were assigned GOs in more than one category.

All unigenes were subjected to searches against the COG database to predict and classify their functions. Overall, 951 of the 5993 sequences (16%) showing

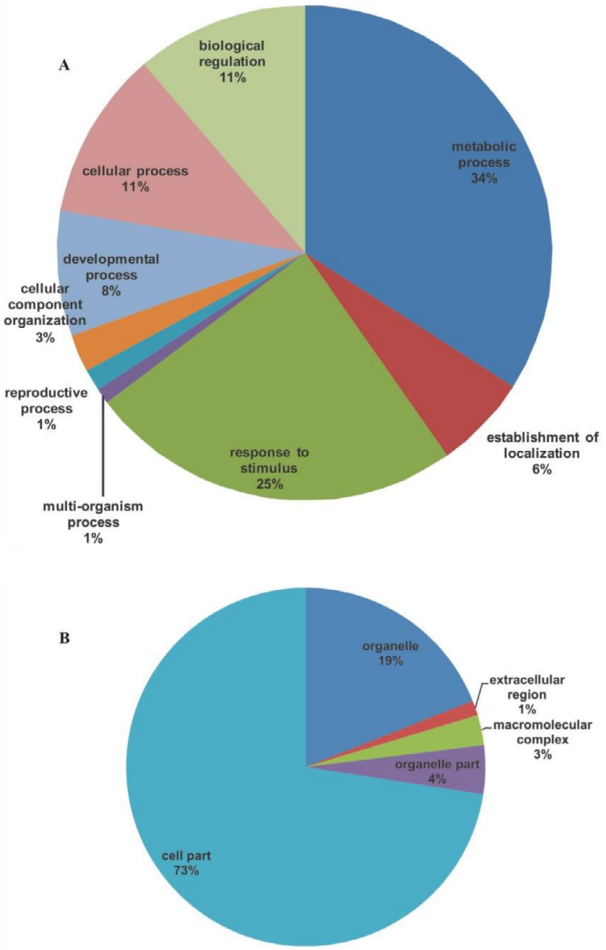

Figure 1. Gene ontology (GO) annotation results from faba bean consensus sequences in the biological process category **(A)**. Gene ontology (GO) annotation results from the faba bean consensus sequences in the cellular processes category **(B)**.

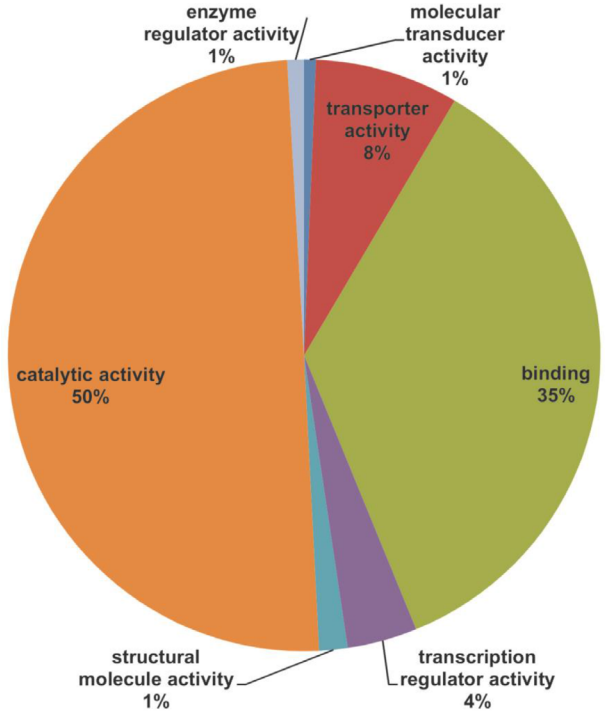

Figure 2. Gene ontology (GO) annotation results from faba bean consensus sequences in the molecular processes category.

nonredundant hits were assigned to COG classifications. COG-annotated putative proteins were functionally classified into 23 molecular families, such as information storage, processing, cellular processes, and metabolism (Figure 3). The R class for general predictions of function (158; 16.61%) represented the largest group, followed by the J class for translation, ribosomal structure, and biogenesis (124; 13.04%) and the O class for posttranslational modification and protein turnover chaperones (117; 12.30%). COG results also indicate the expression of symbiotic-related genes and genes related to various biological mechanisms (Table S1; on the journal's website).

3.3. Discovery of SSR and SNP markers

SSRs are one of the most popular marker systems, consisting of varying numbers of tandemly repeated di-, tri-, or tetranucleotide DNA motifs. To identify SSR markers, we used the ARGOS program with default settings for the *V. faba* unigene collections (Table 2). In total, 1729 potential SSR motifs were identified (Table S2; on the journal's website), and the majority belonged to trinucleotide (67.49%) and dinucleotide (19.15%) repeats. All other types of SSRs such as tetra-, penta-, and hexanucleotide motifs were relatively low in frequency (13.3%), and the majority of trinucleotide SSRs had the GAA/AAG/AGA motif, followed by those with the TGG/CGT/GGT motif and those with the CTT/TTC/TCT motif (Table 2). The GA/AG/, AT/TA, and GT/TG motifs were identified among the dinucleotide SSRs (Table 2). In total, 4856 SNP types were detected (Figure 4) (Table S3; on the journal's website). These variants contained 4120 SNPs, 500 InDels, and 236 variants involving more than one nucleotide. During this process, three criteria were used to screen out all but the highest confidence differences in sequences, although these requirements reduced sensitivity for detecting rare SNPs. This specific screening process revealed that 2201 variants were present in at least 10% of the reads aligned at the polymorphic locus, which was confirmed by 24,198 reads. The high confidence differences were composed of 1946 SNPs, 145 InDels, and 110 variants involving more than one nucleotide (Figure 4; Table S3). Within the detected SNP transition, 59.89% were much more common than the transversion manifestation at 40.10% (Table S3). A similar number of C/T transitions (1282) and a similar percentage

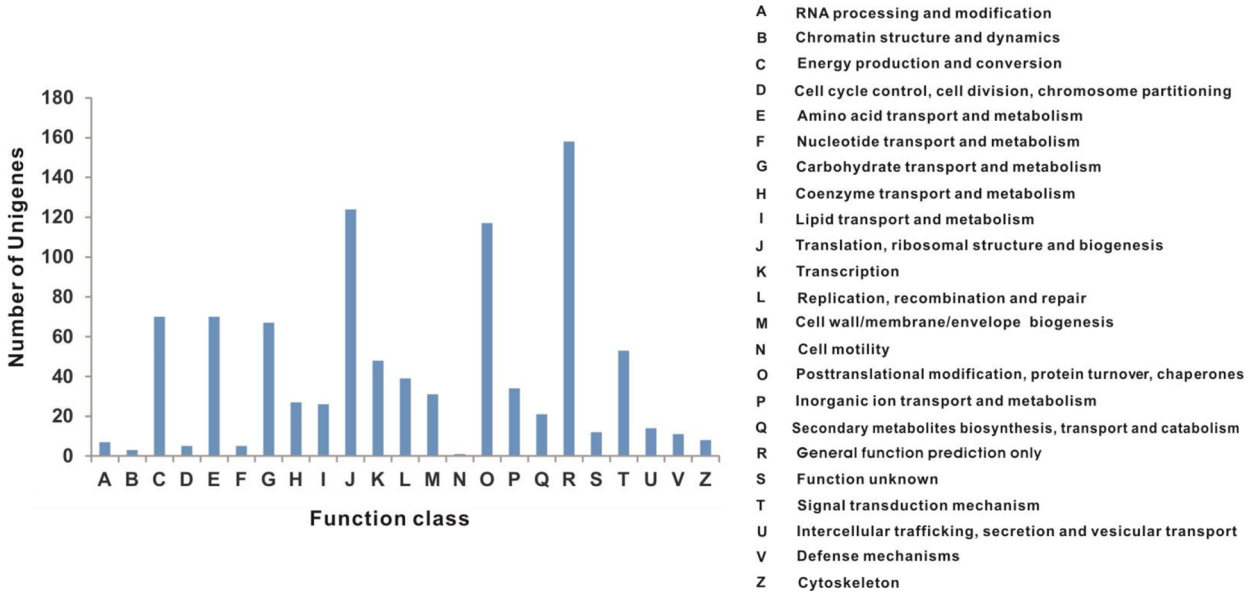

Figure 3. Cluster of orthologous groups (COG) classification. Overall, 951 of the 5993 sequences showing nonredundant hits were assigned to 23 COG classifications.

of the other four transversion types (A/T, A/C, G/T, and C/G) were found (Figure 4; Table 3). A set of SNPs could be accurately located with respect to putative initiation and termination codons.

3.4. Validation of the SSR assay

A subset of SSR primer pairs was selected to validate marker assay performance. In total 1729 potential SSR loci were identified, and among the identified loci we selected 440 SSR-containing sequences, which were deposited in GenBank (GenBank accession numbers: KC218573 to KC218812 & KF658462 to KF658661). Among 440 primers sequences, only 240 primer pairs were synthesized for validation and only 55 primer pairs were successfully amplified from one or more template accessions, of which 53 (96.36%) revealed polymorphisms between eight *V. faba* accessions.

4. Discussion

Limited genomic information is available for the major food and feed legume faba bean cultivars as well as for other legume species such as chickpea, field pea, and lentil. The 454 GS-FLX Titanium next-generation sequencing technology provides a rapid and efficient method for enriching genomic resources by generating large numbers of ESTs with individual read lengths of up to 500 bp. This technology has been used previously to perform de novo bacterial genome sequencing, whole-genome shotgun sequencing, metagenomic studies, transcriptome characterization, and small RNA sequence determination (Margulies et al., 2005).

The present study describes the wide-ranging characterization of faba bean transcriptome. Large-scale transcript sequence data were generated mainly to identify genes and develop gene-based markers to accelerate basic and applied genomics research with faba bean cultivars. We used 454 sequencing technology to sequence the faba bean transcriptome, which yielded 29.61 Mb sequence reads. A total of 81,333 raw sequence reads were obtained and assembled by the De Novo Assembler. The de novo sequence assembly resulted in 50,632 completely assembled contigs, 3985 partially assembled contigs, 25,231 singletons, and 18 repeat sequences.

The complete unigene set was analyzed against the draft genomes of model legume species *Medicago truncatula* and *Arabidopsis thaliana* to identify unique matches. A total of 26,497 faba bean unigenes were subsequently annotated from GenBank. Some recent studies indicated that short reads from 454 GS FLX pyrosequencing can be assembled and used effectively to characterize the gene space in various organisms (Barbazuk et al., 2007; Vera et al., 2008; Meyer et al., 2009; Parchman et al., 2010).

In our study, the most abundant GO function in the biological process category was metabolic processes (34%), followed by responses to stimuli (25%) (Figure 1A.). Catalytic activity was the most abundant category (50%) in the molecular processes category (Figure 2), followed by binding (35%). In the cellular component category, a cell part was the most abundant (73%), followed by organelle (19%) (Figure 1B.). In a root transcriptome study, Barrero et al. (2011) reported that under biological

Table 2. The number of di-, tri-, and other (tetra-, penta-, hexa-) nucleotide repeats identified in the *Vicia faba* unigene dataset.

Dinucleotide repeats	Number of di-SSRs
CT/TC	51
CG/GC	3
GT/TG	35
CA/AC	31
GA/AG	133
AT/TA	77
Total	330
%	19.15
Trinucleotide repeats	**Number of tri-SSRs**
CGG/GGC/GCG	10
TTA/TAT/ATT	13
TGG/GTG/GGT	162
TCG/CGT/GTC	2
TAA/ATA/AAT	11
GTT/TGT/TTG	59
GAT/ATG/TGA	73
GAA/AAG/AGA	175
CTT/TTC/TCT	110
CTG/TGC/GCT	21
CCT/CTC/TCC	54
CCG/CGC/GCC	11
CCA/CAC/ACC	61
CAT/ATC/TCA	102
ACG/CGA/GAC	17
AGT/GTA/TAG	0
AGG/GAG/GGA	30
AGC/GCA/CAG	44
ACT/CTA/TAC	1
ACC/CAC/CCA	122
AAC/ACA/CAA	85
Total	1163
%	67.49
Others (tetra, penta, and hexa)	236
%	13.3

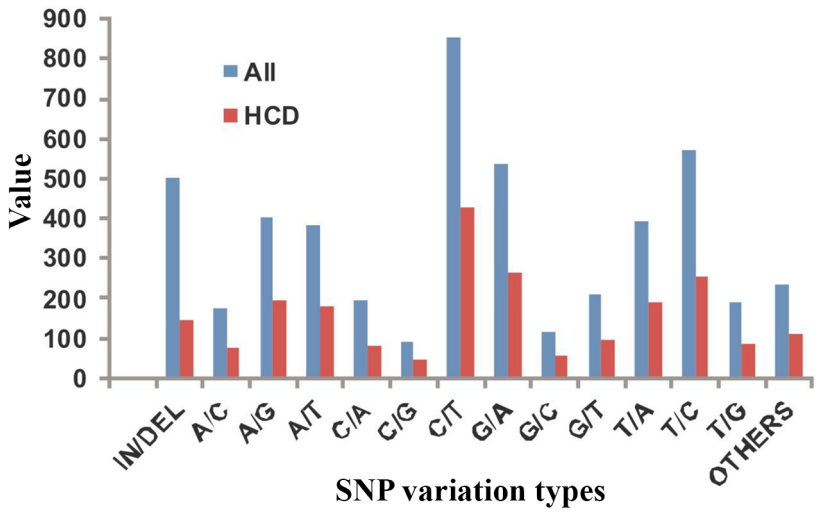

Figure 4. Summary of all variation types and high confidence variation type of single nucleotide polymorphisms (SNPs).

Table 3. Single nucleotide polymorphism (SNP) statistics. Types and numbers of transitions and transversions are shown for putative high-quality SNPs identified in faba bean.

SNPs	Number	SNPs	Number
Transition		Transversion	
AG	600	AT	567
CT	1282	GT	304
		CG	136
		AC	253
Total	1882	Total	1260

process GO functions, the largest percent of transcripts in the *E. fischeriana* root were 'Metabolic process' (23.2%) and 'Response to stimulus' (13.4%), indicating that a large range of metabolic activities occur in the root, which is essential for plant growth and development. Under molecular functions GO, the two most abundant transcript categories were 'binding' and 'catalytic activity', accounting for 33.2% and 11.8%, respectively. The distribution of faba bean unigenes followed similar tendencies to those previously reported in Arabidopsis, melon, and lentil transcriptomes (Gonzalez-Ibeas et al., 2007; Gan et al., 2011; Kaur et al., 2011).

In this study, a large number of unigenes were assigned to a wide range of COG classifications, suggesting that the assembled unigenes represented a wide diversity of transcripts in the faba bean genome (Figure 3). Among COG classifications, general function prediction (158; 16.61%) represented the largest group, followed by translation, ribosomal structure, and biogenesis (13.04%) and by posttranslational modification and protein turnover chaperones (12.30%). The COG classification of this study has similar tendencies to the previously reported *Hevea brasiliensis* transcriptome (Li et al., 2012); among the 24 COG categories, the cluster for general function prediction (959; 17.25%) was the largest group, followed by posttranslational modification, protein turnover, and chaperones (485; 8.72%).

SSR-based marker systems have increasingly become popular for population genetic analyses and genetic mapping studies (Luikart et al., 2003). However, only a few microsatellites were available for faba bean until recently (Pozarkova et al., 2002). Genetic diversity studies within this genus have been mainly performed with random amplification of polymorphic DNA or amplified fragment

length polymorphism, which are less reliable than sequence-based cosegregation markers (Zeid et al., 2009). In our study, the majority of SSRs showed trinucleotide (67.61%) and dinucleotide (19.08%) repeats. All other types of SSRs such as tetra-, penta-, and hexanucleotide motifs were relatively low in frequency (13.3%), and the majority of the trinucleotide SSRs showed the GAA/AAG/AGA motif, followed by those with the TGG/CGT/GGT motif and the CTT/TTC/TCT motif. Among the dinucleotide SSRs, the GA/AG/, AT/TA, and GT/TG motifs were identified.

Blair et al. (2009) reported gene-based microsatellites in common bean; the lack of GC microsatellites has been observed within the bean genome**,** while AT-rich microsatellites were not expected to be found in gene sequences neither as dinucleotides nor trinucleotides. (Blair et al., 2008). There were only a few ACn-based microsatellites and it was surprising that enhancement of this motif generated about the same number of markers as AGn- or GAn-based probes (Gaitan-Solis et al., 2002; Hanai et al., 2007). Among the trinucleotide motifs, AAG (23), ACC (12), AGC (12), AGG (16), and ATC (12) microsatellites are the most common and their frequency might be used for triplet codons for amino acid incorporation in polypeptides. Moreover, open reading frames are known to have a higher GC percentage when compared to nontranslated regions, which favor trinucleotide motifs such as ACC, AGC, and AGG (Li et al., 2004). Similarly in this study, we found that trinucleotide motifs were most common and their frequency was higher, which shows the majority being located in the open reading frame of the original mRNA transcripts represented by the cDNA sequences.

The present study mainly focused on 55 cDNA-SSR primer pairs that were successfully amplified from one or more template genotypes, of which 53 (96.36%) revealed polymorphisms between eight *V. faba* accessions. The results provide a valuable tool and are useful for population genetic analyses and genetic mapping studies with faba bean germplasm, as well as for dissecting the genetic control of important agronomic traits. Kaur et al. (2011) reported SSR marker validation with lentil genotypes; a total of 192 primer pairs were selected for validation, of which 166 primer pairs were successfully amplified and 51 revealed polymorphism between 12 *L. culinaris* genotypes. Kaur et al. (2012) also reported validation of SSR markers with field pea and faba bean; a total of 96 EST-SSR primer pairs from field pea and faba bean were selected for validation with field pea, of which 86 primer pairs successfully amplified one or more template genotypes and 40 (46.5%) revealed polymorphism between 5 field pea genotypes. Moreover, 81 faba bean primer pairs amplified only 24 (29.6%) and thus detected polymorphism between cultivated *V. faba* genotypes (Icarus and Ascot).

In our study, the SNP types of high confidence difference were composed of 1946 SNPs, 145 InDels, and 110 variants involving more than one nucleotide (Figure 4). Within the detected SNPs, transitions (59.89%) were much more common than transversions (40.10%). A similar number of C/T transitions (1282) and a similar percentage of the four transversion types (A/T, A/C, G/T, and C/G) were also found (Table 3). Blanca et al. (2011) reported SNP discovery in *Cucurbita pepo*; a total of 19,980 SNPs and 1174 InDels were distributed. Within the detected SNPs, transitions (68%) were much more common than transversions (32%). A similar number of A/G and C/T transitions and similar percentages of the four transversion types (A/T, A/C, G/T, C/G) were also reported. Similarly, in the present study, highly informative SSRs and SNPs with high polymorphism were successfully identified.

In conclusion, we sequenced and characterized the faba bean transcriptome using 454 GS FLX Titanium sequencing technology. Transcriptomes that resulted in the identification of a large number of informative SSRs and SNPs with high polymorphism provide a set of functional markers, which constitute a resource for mapping and marker-assisted breeding in faba bean. Moreover, these data can be utilized for either basic or applied research with faba bean, particularly in structural and functional genomics.

Acknowledgment

This study was carried out with the support of "Research Program for Agricultural Science & Technology Development (Project No. PJ008623)", National Academy of Agricultural Science, Rural Development Administration, Republic of Korea.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. (2000). Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29.

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

Barrero RA, Chapman B, Yang YF, Moolhuijzen P, Keeble-Gagnere G, Zhang N, Tang Q, Bellgard MI, Qiu DY (2011). De novo assembly of Euphorbia fischeriana root transcriptome identifies prostratin pathway related genes. BMC Genomics 12: 600.

Blair M, Buendia HF, Giraldo M, Metais I, Peltier D (2008). Characterization of AT-rich microsatellites in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 118: 91–103.

Blair MW, Torres MM, Giraldo MC, Pedraza F (2009). Development and diversity of Andean-derived, gene-based microsatellites for common bean (*Phaseolus vulgaris* L.). BMC Plant Biol 9: 100.

Blanca J, Canizares J, Roig C, Ziarsolo P, Nuez F, Pico B (2011). Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). BMC Genomics 12: 104.

Bond DA (1983). Pollination. In: Hebblethwaite PD, editor. The Faba Bean (*Vicia faba* L.). London, UK: Butterworths, pp. 77–101.

Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Schiex T et al. (2006). Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. P Natl Acad Sci USA 103: 14959–14964.

Cheung F, Win J, Lang JM, Hamilton J, Vuong H, Leach JE, Kamoun S, Levesque CA, Tisserat N, Buell CR (2008). Analysis of the *Pythium ultimum* transcriptome using Sanger and Pyrosequencing approaches. BMC Genomics 9: 542.

Chung JW, Kim TS, Suresh S, Lee SY, Cho GT (2013). Development of 65 novel polymorphic cDNA-SSR markers in common vetch (*Vicia sativa* subsp. *sativa*) using next generation sequencing. Molecules 18: 8376–8392.

Doyle JJ, Luckow MA (2003). The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol 131: 900–910.

Duc G, Bao SY, Baum M, Redden B, Sadiki M, Suso MJ, Vishniakova M, Zong XX (2010). Diversity maintenance and use of *Vicia faba* L. genetic resources. Field Crop Res 115: 270–278.

Elahi E, Ronaghi M (2004). Pyrosequencing: a tool for DNA sequencing analysis. Methods Mol Biol 255: 211–219.

Gaitan-Solis E, Duque MC, Edwards KJ, Tohme J (2002). Microsatellite repeats in common bean (*Phaseolus vulgaris*): isolation, characterization, and cross-species amplification in *Phaseolus* ssp. Crop Sci 42: 2128–2136.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477: 419–423.

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 Genomics 11: 18.

Gonzalez-Ibeas D, Blanca J, Roig C, Gonzalez-To M, Pico B, Truniger V, Gomez P, Deleu W, Cano-Delgado A, Arus P et al. (2007). MELOGEN: an EST database for melon functional genomics. BMC Genomics 8: 306.

Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, Holoch D, Lim C, Tuschl T (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods 44: 3–12.

Hanai LR, de Campos T, Camargo LEA, Benchimol LL, de Souza AP, Melotto M, Carbonell SAM, Chioratto AF, Consoli L, Formighieri EF et al. (2007). Development, characterization, and comparative analysis of polymorphism at common bean SSR loci isolated from genic and genomic sources. Genome 50: 266–277.

Kaur S, Cogan NOI, Pembleton LW, Shinozuka M, Savin KW, Materne M, Forster JW (2011). Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. BMC Genomics 12: 265.

Kaur S, Pembleton LW, Cogan NOI, Savin KW, Leonforte T, Paull J, Materne M, Forster JW (2012). Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. BMC Genomics 13: 104.

Kim D (2004). Developing one step program (SSR manager) for rapid identification of clones with SSRs and primer designing. Thesis, Seoul National University, Seoul, Republic of Korea.

Kristiansson E, Asker N, Forlin L, Larsson DGJ (2009). Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. BMC Genomics 10: 345.

Li DJ, Deng Z, Qin B, Liu XH, Men ZH (2012). De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). BMC Genomics 13: 192.

Li YC, Korol AB, Fahima T, Nevo E (2004). Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21: 991–1007.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003). The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4: 981–994.

Ma KH, Kim KH, Dixit A, Chung IM, Gwag JG, Kim TS, Park YJ (2010). Assessment of genetic diversity and relationships among *Coix lacryma-jobi* accessions using microsatellite markers. Biol Plantarum 54: 272–278.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.

Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV (2009). Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics 10: 219.

Michelmore RW, Paran I, Kesseli RV (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. P Natl Acad Sci USA 88: 9828–9832.

Novaes E, Drost D, Farmerie W, Pappas G Jr, Grattapaglia D, Sederoff R, Kirst M (2008). High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9: 132.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11: 180.

Pozarkova D, Koblizkova A, Roman B, Torres AM, Lucretti S, Lysak M, Dolezel J, Macas J (2002). Development and characterization of microsatellite markers from chromosome 1-specific DNA libraries of Vicia faba. Biol Plantarum 45: 337–345.

Riley M (1993). Functions of the gene-products of *Escherichia coli*. Microbiol Rev 57: 862–952.

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, Munsterkotter M et al. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 32: 5539–5545.

Serres MH, Riley M (2000). MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. Microb Comp Genomics 5: 205–222.

Sjödin J (1971). Induced morphological variation in *Vicia faba* L. Hereditas 67: 155–179.

Terzopoulos PJ, Bebeli PJ (2008). Genetic diversity analysis of Mediterranean faba bean (*Vicia faba* L.) with ISSR markers. Field Crop Res 108: 39–44.

Varshney RK, Close TJ, Singh NK, Hoisington DA, Cook DR (2009). Orphan legume crops enter the genomics era! Curr Opin Plant Biol 12: 202–210.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17: 1636–1647.

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007). Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. Plant Physiol 144: 32–42.

Wheat CW (2010). Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica 138: 433–451.

Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N (2006). 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7: 245.

Young ND, Cannon SB, Sato S, Kim D, Cook DR, Town CD, Roe BA, Tabata S (2005). Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiol 137: 1174–1181.

Young ND, Mudge J, Ellis TH (2003). Legume genomes: more than peas in a pod. Curr Opin Plant Biol 6: 199–204.

Zagrobelny M, Scheibye-Alsing K, Jensen NB, Moller BL, Gorodkin J, Bak S (2009). 454 pyrosequencing based transcriptome analysis of *Zygaena filipendulae* with focus on genes involved in biosynthesis of cyanogenic glucosides. BMC Genomics 10: 574.

Zeid M, Mitchell S, Link W, Carter M, Nawar A, Fulton T, Kresovich S (2009). Simple sequence repeats (SSRs) in faba bean: new loci from *Orobanche*-resistant cultivar 'Giza 402'. Plant Breeding 128: 149–155.

Zhu H, Choi HK, Cook DR, Shoemaker RC (2005). Bridging model and crop legumes through comparative genomics. Plant Physiol 137: 1189–1196.