Characterization of 337 exon-based single nucleotide polymorphisms (SNPs) unique to the Indonesian soybean varieties

I M Tasma*, D Satyawan and H Rijzaani

Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development (ICABIOGRAD), Indonesian Agency for Agricultural Research and Development, Ministry of Agriculture, Jalan Tentara Pelajar No. 3A, Bogor 16111, West Java, Indonesia

*E-mail: imade.tasma@gmail.com

Abstract. Genome resequencing of five Indonesian soybean varieties resulted in a total of 2,597,286 single nucleotide polymorphisms (SNPs), 257,598 insertions, and 202,157 deletions. Out of those SNPs, only 95,207 (2.15%) were located in the protein-coding region (exon). The objective of this study was to characterize 337 exon-based SNPs unique to the Indonesian soybean varieties. The study was conducted by taking SNP samples located in the exons using criteria of gene fragments containing the SNPs that were sequenced at least five times within each of the soybean varieties. Out of 95,154 gene-based SNPs detected, only 337 SNPs met the criteria. Each of the soybean varieties was genotyped with the 337 SNP loci, and the genotypic data were scored and analyzed. Results showed that 59 SNPs were common to all five soybean genotypes. A total of 43, 41, 25, 32 and 28 SNPs loci were unique to soybean genotype Davros, Grobogan, Malabar, Tambora and B3293, respectively. These unique SNPs can function as DNA fingerprints for each variety. Out of 59 common SNPs, 24 SNPs were mutations that change the amino acid sequence of the encoded proteins. These genes with amino acid change may have high economic values such as those controlling soybean adaptation in tropical climate, photoperiod insensitivity, disease and insect resistance genes. Expression analyses of the genes with amino acid change showed variation in the expression pattern across different soybean tissues. Functional genomic analysis is necessary to isolate genes useful for breeding purposes.

Keywords: soybean, SNP, genome variation, gene expression, DNA fingerprinting.

1. Introduction

One of the DNA markers recently very popular and which is commonly used, especially in high-throughput marker analysis is single nucleotide polymorphism (SNP) [1,2]. SNP is a DNA variation within the genome that occurs when a single nucleotide changes to another type of nucleotide at the same position. With the rapid development of the genome sequencing technology, SNP markers become more popular due to abundant availability in the genome and can be used in high-throughput genotyping system. SNP marker is biallelic, almost limitless in the genome, easy to be automated and SNP data is easily combined from one laboratory to another making the marker easier to be applied among different laboratories [1–3].

Other types of sequence-based DNA markers include insertion (addition of a base) and deletion (missing of a base) at a specific location within the genome, and both markers are termed as insertion and deletion (indel) [4]. However, the frequency of indel occurrence in the genome is much lower than that of SNPs. SNP and indel are of high-value DNA markers with rapid development in recent years. SNP markers are the basic materials for SNP chip development to be used in a high-throughput genotyping system. The high-throughput genotyping system facilitates gene discovery and quantitative trait loci (QTL) tagging of economically important traits (e.g. yield, seed size, oil content, pest and disease resistances) of agriculturally important crop species such as soybean.

The discovery and detection of superior alleles, genes and QTLs of important traits can be expedited by using high-throughput sequencing and genotyping technologies [5,6]. One of the high-throughput sequencing technology is next generation sequencing (NGS) HiSeq platform that can result in billions of bases (300–600 Gb) at one run of the platform [7] of high-quality sequence due to high sequencing accuracy [6]. A large amount of sequence data resulted by NGS with lower sequencing cost significantly facilitates the improvement of genomic studies of important crop species [8,9]. Supported by modern bioinformatics the important DNA variants useful for breeding can be easily detected to expedite breeding programs.

With the availability of reference genome sequence of various important crop species, the NGS technology becomes very powerful to identify the genomic variation of a crop species through resequencing of various genotypes of the species member to discover superior genes and QTLs together with a large number of DNA markers applicable for plant breeding programs.

Using NGS technology, genotypic characterization of Plant Genetic Resources (PGR) collection can be done in a more comprehensively manner at genome level, and hence the superior gene and QTL discoveries become more efficient, more accurate, and faster. Management of PGR collection in the genebank will be more efficient as all of the materials stored have been identified as genetically distinct.

Resequencing of individual genotypes of a plant species can be easily done with the availability of the reference genome sequence of that particular species. The availability of the reference genome sequence expedites analysis of the resequencing data. Alignment of the resequencing data with those of the reference genome sequences resulted in millions of genomic variations, such as SNPs, indels, and simple sequence repeats (SSRs) [1,10]. These are the main source of DNA markers for breeding purposes [1,6].

The soybean reference genome sequence derived from cultivar Williams 82 has been available since 2010, and the sequence can be accessed by public [11]. The size of soybean genome is about 1.1 Gb predicted to contain 46,430 protein-coding genes. The gene content is about 70% more compared to that of the model plant *Arabidopsis thaliana*. The availability of the soybean reference genome sequence facilitates the identification of the genetic basis of many economically important traits through resequencing studies of many soybean genotypes to support the acceleration of national soybean cultivar development.

The objective of this study was to characterize 337 exon-based SNPs unique to the Indonesian soybean varieties. Characterization included the discovery of exon-based SNPs obtained in all genotypes sequenced (common SNPs) and the ones that existed in only one specific genotype (unique SNPs). The expression pattern of the genes in which the SNPs changed the amino acid sequences was also studied at various soybean tissues.

2. Materials and methods

2.1. Genetic materials

Genetic materials used in this study were five Indonesian soybean varieties, i.e. Grobogan, Tambora, Davros, Malabar and B3293. Grobogan is a superior variety of high productivity with large seed size. Tambora is a superior variety with medium seed size that was introduced from the Philippines. B3293 demonstrated some tolerance to aluminum toxicity and acid soils. Davros is a high productivity soybean variety, and many Indonesian soybean varieties inherit genomes from Davros. Malabar is a

shading-tolerant soybean variety. The selected five varieties are genetically distant based on phylogenetic studies using SSR markers [10,12–14] and are good sources to be used for detecting sequence-based DNA variations such as SNPs and indels.

2.2. DNA isolation, genomic library construction and genome sequencing using NGS Hiseq

Genomic DNAs of the five varieties were isolated using CTAB buffer by following the method of Michiels et al. [15] with a small modification [16]. Genomic DNA libraries were prepared following the method of Tasma et al. [17] and then sequenced in NGS Hiseq platform by using the Illumina reagents, kits, and protocols as previously reported [14,17].

2.3. SNP and indel detection

The alignment of the final sequences derived from five Indonesian genotypes with that of the soybean reference genome sequence Williams 82 [11] was done by using Bowtie2 software [19] followed by SNP identification by using mpileup of Samtools [20]. Effect prediction of SNPs was done using snpEff software [21].

2.4. Exon-based DNA mutation characterization

SNPs identified were then filtered using the following criteria: (i) SNP was located within an exon, (ii) DNA fragment containing the SNP was sequenced at least 5 times (5 times genome coverage) in each of the five soybean genotypes. Among the 95,154 SNPs identified, only 337 SNPs met the criteria. Each soybean genotype (Davros, Grobogan, Malabar, Tambora and B3293) was genotyped with 337 SNP markers, scored and analyzed. The genes containing SNP loci common to all five soybean varieties that changed amino acid composition were further characterized for their expression patterns.

3. Results and discussion

3.1. Common and unique SNPs distributed across five soybean genotypes

Table 1. Common and unique SNPs found in five Indonesian soybean varieties (Davros, Grobogan, Malabar, Tambora and B3293).

Genotype	Number
Unique genic SNP[a]	
Davros	23
Grobogan	41
Malabar	25
Tambora	32
B3293	28
Common genic SNP[b]	59
Total SNPs	208

[a] SNP derived from exon that is unique to a particular soybean genotype analyzed.
[b] SNP derived from exon that belonged to all soybean genotypes analyzed.

Among 95,154 SNPs located within the exon, 337 SNPs were found to have at least five sequence copies in each of the five soybean genotypes, but the identified SNPs were different from the ones in Williams 82 variety. Among the 337 SNPs, 59 SNP loci were found to demonstrate the same

genotypes in all of the five soybean genotypes (i.e. Davros, Grobogan, Malabar, Tambora and B3293), that were different from the ones in the reference genome Williams 82 (Table 1, Figure 1). The 59 SNP loci were called common SNPs.

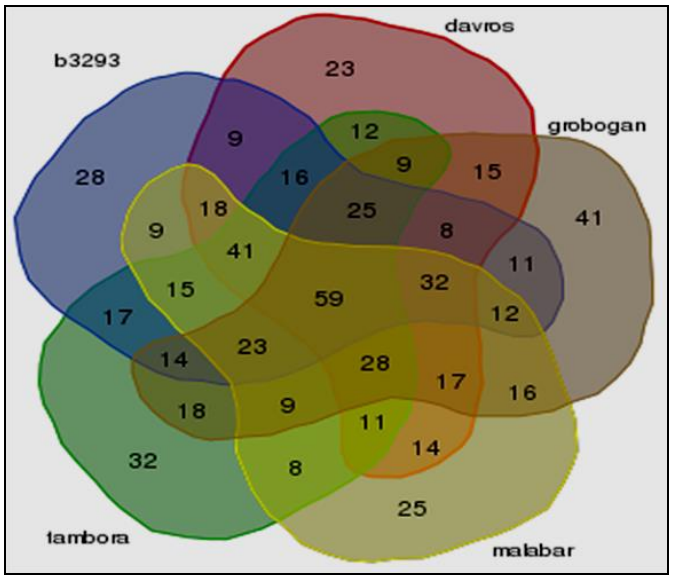

Figure 1. Venn diagram showing the unique and common SNP distributed across the five Indonesian soybean genotypes (Davros, Grobogan, Malabar, Tambora and B3293). Shown in the picture is 59 SNP loci common to all five soybean varieties. Other SNP loci were common to four, three, or two soybean genotypes. The remaining SNP loci were unique to a particular soybean genotype.

Hundreds of unique SNP loci were found only within a particular soybean genotype. A total of 23 SNP loci were observed to be unique in Davros, 41 loci were found to be unique in Grobogan, 25 loci in Malabar, 32 unique loci were discovered in Tambora, and 28 unique SNP loci were found only in the genome of B32933 (Table 2, Figure 2).

These unique SNP loci can be used as fingerprints for that specific genotype and, therefore, they can function as a unique identity of the soybean genotype. Such kind of DNA identity is very important to protect the soybean genotypes from the use by other parties without permission from the germplasm owners. The remaining SNP loci (129 SNPs) were not classified as unique nor common SNPs as they were found in two, three, or four soybean genotypes (Figure 1).

3.2. Exon-based gene mutation characteristics of common SNP loci

Among the 59 SNP loci common to all five soybean varieties but different from the ones in the reference genome Williams 82, a total of 24 SNP loci were SNPs mutations that caused amino acid change of the proteins encoded by the respective genes where the mutated SNPs were located (non-synonymous SNP mutation) (Table 2). Interestingly, among the mutated genes and the proteins they encoded, few might control important phenotypes, such as the ones for tropical adaptation, photoperiod insensitivity, heat/high temperature tolerance, or resistance to tropical soybean diseases and insect pests.

Table 2. Mutated genes that changed their amino acid composition obtained in all five soybean varieties (Davros, Grobogan, Malabar, Tambora and B3293).

Gene designation	Type of protein encoded
Glyma01g42650	Winged-helix DNA-binding transcription factor family protein
Glyma01g44270	RNA-binding (RRM/RBD/RNP motifs) family protein
Glyma03g04810	NB-ARC domain-containing disease resistance protein
Glyma06g40790	Target of AVRB operation1
Glyma06g44770	Glycosyl hydrolase superfamily protein
Glyma07g07530	Receptor-like protein kinase 4
Glyma08g00420	F-box family protein
Glyma08g00420	F-box family protein
Glyma08g01040	DCD (Development and Cell Death) domain protein
Glyma08g01040	DCD (Development and Cell Death) domain protein
Glyma08g24530	Septum site-determining protein (MIND)
Glyma08g26150	0 (the gene product has not been determined)
Glyma09g02860	Myosin heavy chain-related
Glyma11g35320	Phospholipase C 2
Glyma13g26070	Annexin 8
Glyma14g04690	Receptor-like protein 27
Glyma14g04690	Receptor-like protein 27
Glyma16g05600	C2H2 and C2HC zinc fingers superfamily protein
Glyma18g02200	Ethylene insensitive three family protein
Glyma18g46080	Disease resistance protein (TIR-NBS-LRR class) family
Glyma20g06680	Dynamin-related protein 3A
Glyma20g34200	Tetratricopeptide repeat (TPR)-like superfamily protein
Glyma20g34200	Tetratricopeptide repeat (TPR)-like superfamily protein
Glyma20g36630	Ankyrin repeat family protein

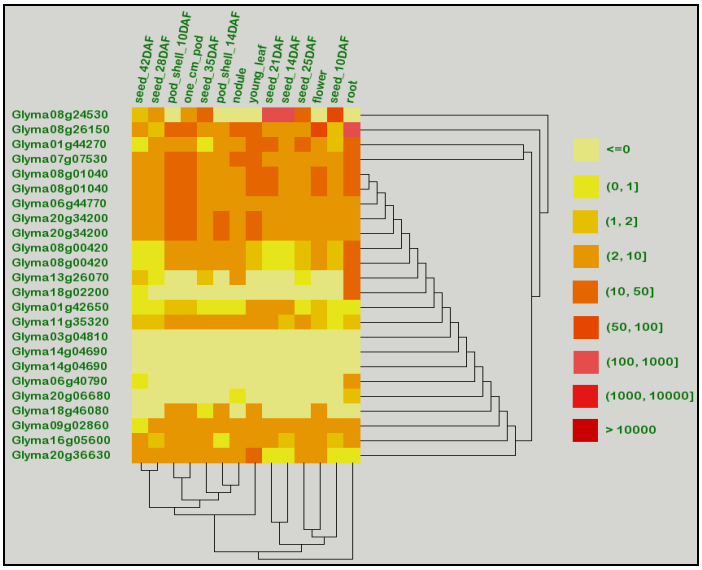

Figure 2. Expression pattern of the 24 mutated genes that were found to be common to all of the five Indonesian soybean varieties expressed in various soybean tissues. DAF = days after flowering.

The twenty-four genes described in Table 2 were differentially expressed across different soybean tissues (Figure 2). Some genes were highly expressed in leaf, root, flower, pod, or seed; while several other genes were very weakly expressed in all tissues (Figure 2). This indicates that most genes identified in this study were tissue-specific.

4. Conclusions

Out of 95,154 SNPs located within the exon, 337 SNPs were found to have at least five sequence copies in each of the five soybean genotypes but were different from the ones in the variety Williams 82. Among the 337 SNPs, 59 SNPs showed capability to differentiate the five soybean varieties compared to the reference variety (Williams 82), and some SNP alleles were also found in only one variety that is potentially used for DNA fingerprinting purposes for that particular variety. Out of 59 common SNPs, 24 were mutations that changed amino acid composition of the encoded proteins. Among the genes with amino acid sequence changes some might be genes of high economic values such as those controlling the soybean adaptation in tropical climate and disease and insect resistance genes. The genes with amino acid change showed variation in the expression pattern across different soybean tissues.

5. Acknowledgement

This study was funded by the 2014–2016 Indonesian National Budget for the Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development, IAARD.

6. Authors contribution

IMT is the main contributor, responsible for and designed the research, and wrote the manuscripts. DS is member contributor, prepared the genome library and bioinformatics analysis. HR is member contributor, conducted the genome sequencing (NGS) and sequence data analysis.

7. References

[1] Tasma I M 2014 Single nucleotide polymorphism (SNP) sebagai marka DNA masa depan *War. Biog.* **10** 7–10
[2] Thomson M J 2014 High-throughput SNP genotyping to accelerate crop improvement *Plant Breed. Biotechnol.* **2** 195–212
[3] Leonforte A, Sudheesh S, Cogan N O, Salisbury P A, Nicolas M E, Materne M, Forster J W and Kaur S 2013 SNP marker discovery, linkage map construction and identification of QTLs for enhanced salinity tolerance in field pea (*Pisum sativum* L.) *BMC Plant Biol.* **13** 161–74
[4] Väli Ü, Brandström M, Johansson M and Ellegren H 2008 Insertion-deletion polymorphisms (indels) as genetic markers in natural populations *BMC Genet.* **9** 1–8
[5] Schuster S C 2008 Next-generation sequencing transforms today's biology *Nat. Methods* **5** 16–8
[6] Zhang J, Chiodini R, Badr A and Zhang G 2011 The impact of next-generation sequencing on genomics *J. Genet. Genomics* **38** 95–109
[7] Pettersson E, Lundeberg J and Ahmadian A 2009 Generations of sequencing technologies *Genomics* **93** 105–11
[8] Voelkerding K V, Dames S A and Durtschi J D 2009 Next-generation sequencing: from basic research to diagnostics *Clin. Chem.* **55** 641–58
[9] Metzker M L 2010 Sequencing technologies the next generation *Nat. Rev. Genet.* **11** 31–46
[10] Tasma I M 2016 Resekuensing genom, metode baru karakterisasi variasi SDG tanaman secara komprehensif mendukung akselerasi pemuliaan tanaman *War. Biog.* **12** 2–6
[11] Schmutz J et al 2010 Genome sequence of the palaeopolyploid soybean *Nature* **463** 178–83
[12] Santoso T J, Utami D W and Septiningsih E M 2006 Analisis sidik jari DNA plasma nutfah kedelai menggunakan markah SSR *J. AgroBiogen* **2** 1–7
[13] Tasma I M and Warsun A 2008 Development and Characterization of F_2 population for molecular mapping of aluminum-toxicity tolerant QTL in soybean **4** 1–8

[14] Satyawan D, Rijzaani H and Tasma I M 2014 Characterization of genomic variation in Indonesian soybean (*Glycine max*) varieties using next-generation sequencing *Plant Genet. Resour.* **12** S109–13
[15] Michiels A, Van Den Ende W, Tucker M, Van Riet L and Van Laere A 2003 Extraction of high-quality genomic DNA from latex-containing plants *Anal. Biochem.* **315** 85–9
[16] Satyawan D and Tasma I M 2011 Genetic diversity analysis of elite *Jatropha curcas* (L.) genotypes using randomly amplified polymorphic DNA markers. *Karnataka J. Agric. Sci.* **22** 293–5
[17] Tasma I M, Satyawan D and Rijzaani H 2015 Pembentukan pustaka genom, resekuensing, dan identifikasi SNP berdasarkan sekuen genom total genotipe kedelai Indonesia *J. AgroBiogen* **11** 7–16
[18] Tasma I M, Satyawan D and Rijzaani H 2015 Pembentukan pustaka genom, resekuensing, dan identifikasi SNP berdasarkan sekuen genom total genotipe kedelai Indonesia (Genomic library construction, resequencing, and SNP identification based on whole-genome sequences of Indonesian soybean genotype) *J. AgroBiogen* **11** 7–16
[19] Langmead B and Salzberg S L 2012 Fast gapped-read alignment with Bowtie 2 *Nat. Methods* **9** 357–9
[20] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R 2009 The sequence alignment/map format and SAMtools *Bioinformatics* **25** 2078–9
[21] Cingolani P, Platts A, Wang L L, Coon M, Nguyen T, Wang L, Land S J, Lu X and Ruden D M 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3 *Fly (Austin)* **6** 80–92