© NIAB 2014
ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2014) **12**(S1); S109–S113
doi:10.1017/S1479262114000380

Characterization of genomic variation in Indonesian soybean (*Glycine max*) varieties using next-generation sequencing

Dani Satyawan*, Habib Rijzaani and I. Made Tasma

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

Abstract

Soybean is an important crop in Indonesia and its consumption has consistently surpassed local production in recent times. As the average yield is relatively low, a more efficient breeding programme that utilizes the latest technological developments in DNA analysis is required. To provide a genomic data resource for future breeding programmes, in this study, whole-genome sequencing was performed for five Indonesian soybean varieties, with an average sequencing depth of 34 reads. Comparison of these sequences with the Williams 82 reference sequence revealed 3,150,869 DNA variations, which averages to one variation in every 308 bases. Comparison of these variations with known single-nucleotide polymorphisms (SNPs) in the SoyKB database revealed that approximately 29% of them were novel SNPs unique to the Indonesian cultivars. Variations found within exons totalled 95,154. Of these, 57,171 were capable of causing mutations that would modify the amino-acid composition of the encoded proteins (nonsynonymous mutations). Phylogenetic analysis using a subset of these SNP data indicated that the cultivars had genetic similarities to landraces from China and Japan, which could provide clues to the origin of soybeans that were introduced into Indonesia.

Keywords: deep sequencing; genome comparison; phylogenetic analysis; sequence variations

Introduction

Soybean (*Glycine max*) is widely cultivated and consumed in Indonesia, even though it is not a native plant of this country. The earliest written record mentioning soybean cultivation in Indonesia is dated around the 13th century (Shurtleff and Aoyagi, 2010). Another manuscript has indicated that tempeh made from soybean had been consumed in the early 17th century (Santoso and Pringgoharjono, 2013). Soybean-derived food products remain popular to this day, and the current domestic demand for soybean consistently exceeds the quantity that can be produced locally, partly because the average yield of Indonesian soybean farms is relatively low, ranking 66th worldwide in 2011 (FAOSTAT, 2011). It seems that centuries of domestication have not produced outstanding cultivars and improvements via breeding programmes are needed to produce varieties that perform well in the Indonesian climate.

Newer techniques that utilize DNA-based tools such as marker-assisted selection and genomic selection should be explored to improve the speed and efficiency of local breeding programmes. The use of next-generation sequencing can assist in designing a large number of DNA markers, which would be useful for fine-mapping, genome-wide association studies and genomic selection (Chagne *et al.*, 2012). This study was carried out with the intention to assist future soybean breeding

* Corresponding author. E-mail: d.satyawan@gmail.com

https://doi.org/10.1017/S1479262114000380 Published online by Cambridge University Press

programmes in Indonesia by characterizing the genome of local soybean cultivars. Whole-genome sequencing was performed for five selected local cultivars to (1) assess their genetic diversity and relationship with Chinese cultivars, (2) identify genetic mutations that underlie phenotypic variations, (3) identify allelic variation for the development of DNA markers for future soybean breeding activities in Indonesia.

Materials and methods

Plant materials consisted of five soybean cultivars from Indonesia, namely B3292, Davros, Grobogan, Malabar and Tambora. The five varieties were selected based on several criteria, such as the presence of useful traits, their utilization in breeding programmes, and genetic diversity according to a previous phylogenetic study using simple sequence repeat markers (Santoso *et al.*, 2006). Sequencing was performed using an Illumina Hiseq 2000 sequencing system, according to the manufacturer's instructions. Sequence data were aligned to the Williams 82 reference sequence (Schmutz *et al.*, 2010), which was downloaded from Phytozome (www.phytozome.net), using Bowtie2 (Langmead and Salzberg, 2012) followed by single-nucleotide polymorphism (SNP) calling using mpileup in SAMtools (Li *et al.*, 2009). Annotation of the locations and predicted effects of the SNPs was performed using SnpEff (Cingolani *et al.*, 2012). The resultant data were compared with sequencing data from 31 Chinese accessions (Lam *et al.*, 2010) downloaded from the SoyKB database (soykb.org). Phylogenetic analysis and tree construction were carried out using DARwin (Perrier and Jacquemoud-Collet, 2006). Tree drawing was generated in Dendroscope (Huson and Scornavacca, 2012).

Results and discussion

The average sequence coverage depth for all the loci was 34 reads, and more than 95% of the genome was sequenced at least ten times. In total, we identified 3,150,869 sequence changes, an average of one sequence change per 308 bases. Among these changes, 2,692,193 were SNPs, 257,625 were insertions, and the remaining 201,051 were deletions.

To assist future research in fine-mapping and gene identification using quantitative trait locus mapping and association studies, sequence changes in exon regions were further characterized. A total of 95,154 sequence changes were located in exons. More than half of these changes (49,926 mutations) were missense mutations, while 1535 were nonsense mutations. Table 1 categorizes the non-silent mutations according to their effect on mRNA/protein composition and lists the number of mutations of each type.

To investigate whether some of the exon sequence changes are unique to Indonesian accessions, data on 80,630 SNPs that could be mapped to the 20 soybean chromosomes were compared with SNP data obtained from resequencing 31 Chinese accessions (Lam *et al.*, 2010). There were 57,009 SNPs that matched the SNPs from the Chinese accessions, while 23,621 were unique to the five Indonesian cultivars. These mutation data could comprise a valuable resource for dissecting genetic adaptation to the tropical climate of Indonesia.

Using these mutation data, we then assessed the genetic diversity of the five cultivars compared with the Chinese accessions, which were expected to have greater diversity as they originated in the area where soybean was initially domesticated and some wild accessions were also present among these 31 accessions. A neighbour-joining tree based on 1000 bootstrap replicates was then constructed from the polymorphism data of 1400 genic SNPs that exhibited polymorphism among the Chinese and Indonesian accessions and had a sequencing depth of at least three reads in all the five Indonesian cultivars (Fig. 1). As expected, the five Indonesian accessions were clustered relatively close to each other within the cluster of cultivated accessions from China, even in the case of the Tambora cultivar, which is a recent introduction from the Philippines. The closest relative to Tambora is C16, a Taiwanese cultivar that originated from a Japanese cultivar and is also the closest relative to two other Indonesian cultivars, Malabar and Davros. Malabar is the result of a recent breeding programme that crossed superior local cultivars, while Davros was purified from landraces commonly planted in Garut District (West Java). B3293, a landrace from Kediri (East Java), belongs to a different group and was shown to be most similar to C17, a landrace from Sichuan, in Southwest China.

An unexpected grouping can be observed in the case of Grobogan. Similar to Davros, Grobogan was purified from landraces that are popular in the District of Grobogan, Central Java. It was originally thought to be a variant of Malabar, due to their similar flowering time. Nevertheless, Grobogan is most genetically similar to a landrace from Guangdong (C35), a coastal region located in Southeast China. Grobogan is clearly genetically distinct from Malabar, and its origin might be closer to the original soybean that was introduced from China to Indonesia.

It is unlikely that Indonesian soybeans are derived from a single introduction event that later spread throughout the country, as even accessions derived from traditional landraces exhibit similarities to cultivars

https://doi.org/10.1017/S1479262114000380 Published online by Cambridge University Press

https://doi.org/10.1017/S1479262114000380 Published online by Cambridge University Press

Table 1. Number and types of sequence variations detected among the five Indonesian cultivars

	B3293	Davros	Grobogan	Malabar	Tambora	
Origin in Indonesia	Kediri (East Java)	Garut (West Java)	Grobogan (Central Java)	Elite variety from multiple crosses	Introduction from the Philippines	Total
Sequenced bases	15,908,297,200	6,182,429,000	10,030,813,600	2,886,637,300	8,530,532,500	43,538,709,600
Mapped bases	14,957,923,800	5,417,232,900	8,963,009,500	2,768,735,300	7,620,598,000	39,727,499,500
No. of total variants						
SNPs	1,570,109	819,998	1,035,383	463,714	1,060,835	2,692,193
Indels	277,364	101,313	174,912	45,011	172,649	458,676
Genic SNPs						
In introns	170,984	111,666	127,781	69,237	126,899	312,509
In exons	53,752	26,042	36,521	17,740	36,459	90,395
Synonymous	24,090	11,371	15,799	7702	15,824	39,120
Nonsynonymous	29,662	14,671	20,722	10,038	20,635	51,461
Missense	29,014	14,343	20,095	9822	20,029	49,926
Start codons gained	900	583	686	370	641	1635
Start codons lost	50	18	38	19	35	105
Stop codons gained	648	328	627	216	606	1534
Stop codons lost	226	120	148	88	174	500
Splice site acceptor	256	122	176	85	192	431
Splice site donor	236	124	180	73	209	412
Genic indels						
In introns	41,571	15,848	27,045	8193	25,120	69,192
In exons	2511	809	1832	427	1877	4759
Frameshift	1494	504	1241	308	1227	2958
Codon change plus codon deletion	162	53	72	25	81	250
Codon change plus codon insertion	342	93	239	31	245	656
Codon deletion	290	98	139	37	157	456
Codon insertion	191	54	108	24	138	362

SNPs, single-nucleotide polymorphisms.

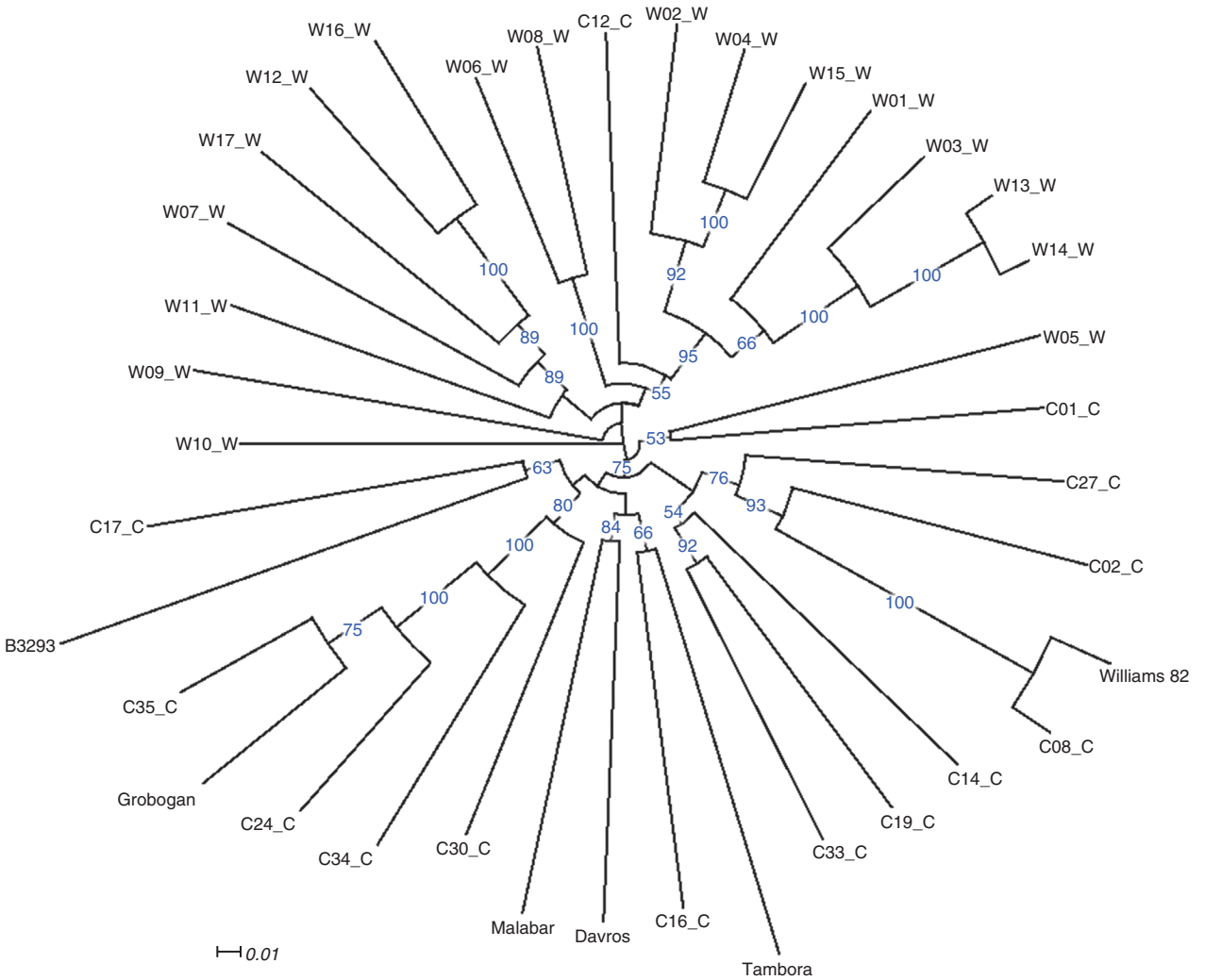

Fig. 1. Phylogenetic relationship between the five Indonesian cultivars and 31 wild and cultivated soybean accessions from China. Accessions labelled with C represent cultivated varieties, while those labelled with W represent wild varieties. Full description of the accessions can be found in the supplementary material section of Lam *et al.* (2010). Blue numbers at the nodes represent the percentage of bootstrap values for 1000 replications and values less than 50 are not displayed in the tree. Scale indicates substitutions per site.

that originated from different regions in China and Japan. Contact between the two countries had been recorded since the 7th century (Müller and Takakusu, 1896), and Khubilai Khan even sent an expedition to invade Java in the 13th century (Marshall, 1993). Such a long history of interactions could present many opportunities for different types of soybean to be introduced.

Nevertheless, the phylogenetic analysis indicated that the breeding programmes could benefit from more genetic materials. This study identified genetic variations and polymorphisms in local soybean cultivars, which could comprise a useful resource for marker development in breeding programmes for soybean in Indonesia. Nonsynonymous mutations were also catalogued to facilitate their use as potential candidates for fine-mapping of useful traits in local cultivars. These genetic data have also given us a glimpse of the overall picture of Indonesian soybean diversity, as well as serve as a starting point for investigations into the origin of Indonesian cultivars.

Acknowledgements

This study was funded by an Indonesian Government grant from the Indonesian Agency for Agricultural Research and Development – Ministry of Agriculture. The authors thank Dr Joshi Trupti for her assistance in accessing the SNP data in SoyKB.

https://doi.org/10.1017/S1479262114000380 Published online by Cambridge University Press

References

Chagne D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, Kumar S, Cestaro A, Velasco R, Main D, Rees JD, Iezzoni A, Mockler T, Wilhelm L, Van de Weg E, Gardiner SE, Bassil N and Peace C (2012) Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS One* 7: e31745.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X and Ruden DM (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* 6: 80–92.

FAOSTAT (Food and Agriculture Organization of the United Nations) (2011) FAOSTAT database. Available at: http://faostat.fao.org/

Huson DH and Scornavacca C (2012) Dendroscope 3: an interactive viewer for rooted phylogenetic trees and networks. *Systematic Biology* 61: 1061–1067.

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SS and Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* 42: 1053–1059.

Langmead B and Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Marshall R (1993) *Storm from the East: From Genghis Khan to Khubilai Khan*. Los Angeles, CA: University of California Press.

Müller FM and Takakusu J (1896) *A Record of the Buddhist Religion as Practised in India and the Malay Archipelago (A. D. 671–695)*. Oxford: Clarendon Press.

Perrier X and Jacquemoud-Collet J (2006) DARwin software. Available at: http://darwin.cirad.fr/

Santoso S and Pringgoharjono K (2013) *Stories from the Serat Centhini: Understanding the Javanese Journey of Life*. Singapore: Marshall Cavendish International.

Santoso TJ, Utami DW and Septiningsih EM (2006) Analisis sidik jari DNA plasma nutfah kedelai menggunakan markah SSR. *Jurnal Agrobiogen* 2: 1–7.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC and Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178–183.

Shurtleff W and Aoyagi A (2010) *History of Soybeans and Soyfoods in Southeast Asia (13th Century to 2010): Extensively Annotated Bibliography and Sourcebook*. Lafayette, CA: Soyinfo Center.

https://doi.org/10.1017/S1479262114000380 Published online by Cambridge University Press