ARTICLE

Genotype profiling, population structure, and seed size traits association analyses using five polymorphic SSR markers in soybean [*Glycine max* (L.) Merr.] genotypes available in the Philippines

Anna Mae M. de los Reyes*[1], Maura Mercedes L. Santos[1], Villamor A. Ladia Jr.[1], Rodel G. Maghirang[1], Elmer E. Enicola[1], Eureka Teresa M. Ocampo[2]

[1]Institute of Plant Breeding (IPB), College of Agriculture and Food Science (CAFS), University of the Philippines, Los Baños (UPLB), College, Laguna
[2]Institute of Crop Science (ICropS), College of Agriculture and Food Science (CAFS), University of the Philippines, Los Baños (UPLB), College, Laguna

ABSTRACT

Seed size is a fundamental attribute considered in soybean production with relevance in yield components and intended use in food processing. Analyses of seed size-relevant quantitative traits using molecular markers provide useful information to aid in more efficient breeding programs. This research aims to perform genetic and population structure analyses in a large soybean population comprised of 109 individuals using five polymorphic seed weight-associated SSR markers from the BARCSOYSSR database (soybase.org), and to examine the association of these markers with relevant traits. Low to high observed heterozygosity (0.073–0.835) and high polymorphism information content (PIC; 0.733–0.846) were calculated through allele frequency-based analyses. Genotypes were grouped into four clusters based on NJ dendrogram, whereas, two subpopulations (K=2) were estimated via STRUCTURE. Majority of genotypes from the NOMIARC collection grouped under cluster III; consistent with the collection belonging to one of the inferred subpopulations. Genotypes from IPB and IITA collections showed no distinct groupings. Finally, marker-trait association analysis using GLM determined BARCSOYSSR_08_1032 to be significantly associated with single-seed weight, length, and width; and BARCSOYSSR_08_1246 with seed length. MLM approach did not produce any significant associations. The two markers may be forwarded for recommendation for application on a more diverse population in terms of number and background. Furthermore, assessment of markers across chromosomes is deemed to yield more information. Conclusively, the preliminary study was able to provide insight in the performance of the SSR markers. If achieved, finding consistently useful markers would pave the way to application in marker-assisted selection in soybean breeding.

*Corresponding author
Email Address: amdelosreyes4@up.edu.ph
Date received: June 15, 2022
Date revised: September 27, 2022
Date accepted: November 20, 2022

INTRODUCTION

Soybean [*Glycine max* (L.) Merrill] is a widely in-demand oilseed crop with an estimated worldwide production of 366.23 million metric

KEYWORDS

marker-trait association, molecular marker screening, population structure, simple sequence repeats (SSRs), soybean seed size

tons in 2020/21 (USDA 2022). The legume species is an economic source of high-quality protein for human and animal consumption, and is also progressively being utilized in manufacturing of industrial products (Gaonkar and Rosentrater 2019). With its existing value as one of the world's primary crops, there is growing interest for more efficient production and, therefore, breeding of soybean (Pagano and Miransari 2016; Pagano et al. 2020). This need is even greater for domestic production in the Philippines with approximately 2,000 to 3,000 tons production per year, accounting for ~1% of the food and feed demands in the country (Galang 2020).

One of the vital qualities targeted in soybean improvement via breeding is seed size. The trait is typically associated with seed quality and crop yield (Smith and Camper 1975; Burton 1987; Ge et al. 2016) and is also of consideration for food uses; whereby, small and large-seeded varieties are preferentially cultivated and utilized for different preparations. For instance, large-seeded cultivars are best for making tofu, soymilk and miso while cultivars with smaller seeds are opted for sprouts (Chen 2004). These market demands have significant influence in the direction of soybean breeding objectives.

The traditional method of phenotype screening, however, is made challenging by substantial genotype by environment (GxE) interactions (Chaves et al. 2017; Mwiinga et al. 2020), not to mention the time and field labor it entails. The use of DNA or molecular markers allows the study of genetic factors linked to quantitative traits such as seed size (Teng et al. 2009); and overall, offers improved efficiency and precision to plant breeding through marker-assisted selection (MAS) (Collard and Mackill 2008).

DNA markers, notably, single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs) have been largely used in analyses of soybean quantitative trait loci (QTL) for their abundance and higher rate of polymorphism. (Teng et al. 2009; Hao et al. 2012; Lee et al. 2015; Zhang et al. 2016; Liu et al. 2018; Li et al. 2019). For this study, selected BARCSOYSSR_1.0 SSR markers from the USDA-ARS Soybean Genetics and Genomics Database, known as SoyBase (www.soybase.org) (Song et al. 2010), were used for analyzing 109 soybean genotypes from different collections. Thus far, the initial set of markers employed for screening were from chromosome Gm08 (Linkage Group A2) genetic map. Polymorphic markers scoring data were then used for genetic, population structure, and association analyses with seed size traits with the objective of identifying marker(s) with potential for application in MAS for soybean breeding.

MATERIALS AND METHODS

Plant materials and field design

A total of 120 soybean genotypes from the International Institute of Tropical Agriculture (IITA); Northern Mindanao Integrated Agricultural Research Center (NOMIARC); and the Institute of Plant Breeding (IPB), Los Baños, Laguna were initially grown at the IPB Demonstration field on pre-dry season October 2019. The latter collection included germplasms sourced from the local collections, Legumes Section and National Plant Genetic Resources Laboratory (NPGRL) of IPB, and one from World Vegetable (AGS 374). Following a randomized design, three replicate blocks were prepared. In every block, one meter row was allotted per entry with 0.5 m distance between consecutive rows. Ten seeds per entry were sown with 0.1 m distance between respective hills. Organic fertilization (vermicompost) was applied as topdress during planting at a rate of 3 tons/ha and hand weeding was performed as necessary. Seed pods were harvested late January to early February 2020. After accounting of failed growth or germinations, 109 genotypes (Supplementary Table 1) were forwarded for molecular and phenotypic analyses.

Phenotype measurements and ANOVA analysis

Single-seed weight (SSW) and 100-seed weight (100-SW) measurements were obtained using an analytical balance while length (SL), width (SW) and thickness (ST) were measured with a digital caliper. With the seed held with its hilum in vertical position, the top to bottom distance was considered the length. The distance from the hilum façade to the opposite side was determined as width. Seed thickness was measured as the extent between the lateral sides flanking the hilum. SSW, SL, SW and ST data were collected from 10 randomly selected seeds for each genotype per replicate. Average values of the 10 replicates for the four traits were calculated. 100-SW was obtained once for every genotype per replicate. The General Linear Model (GLM) was applied to perform ANOVA for unbalanced data and to obtain Pearson correlation coefficients of the quantitative traits through SAS® [Release 3.8 (Enterprise Edition), Copyright © 2012-2018, SAS Institute Inc., Cary, NC, USA] software.

DNA extraction

Collection of fresh leaf samples commenced when seedlings were at three-weeks old. In order to accommodate the bulk of samples, a rapid method of DNA extraction by Mamidala and Nanna (2009) was used with modifications. Three hundred milligrams of pooled leaf samples from three random plants were prepared per genotype and ground with one milliliter extraction buffer [0.1 M Tris-HCl, pH 8.0; 0.02 M EDTA, pH 8.0; 1.5 M NaCl; 8% (w/v) SDS (sodium dodecyl sulfate); and 0.2% β-mercaptoethanol]. The mixture was equally divided between two fresh tubes and added with potassium acetate (5M; 200 µL) prior to centrifugation at 12,000 rpm for 10 min at 4°C. Fifty microliters of the supernatant were transferred on to new microcentrifuge tubes pre-dispensed with 200 µL sterilized ultrapure water. Extracted DNAs were kept at –20°C.

SSR markers screening

Candidate SSR markers were identified using SoyBase, a public database for soybean genome-related information developed by the USDA-ARS SoyBase and Legume Clade Database group at the Iowa State University. The initial batch of seed weight-associated SSR markers obtained through the SoyBase-provided genetic map comprised of 17 markers at soybean chromosome Gm08 (Song et al., 2010; Supplementary Table 2). These markers were screened for polymorphism using 20 soybean genotypes from the IPB Legumes Section and local collections, and from advanced yield trials of high yielding varieties/breeding lines according to the National Seed Industry Council (NSIC) (data not published); results yielded five polymorphic markers that were forwarded to use for replicated analysis of the 109 soybean genotypes from different collections. Primer details are presented in Table 1.

PCR reaction mixtures contained 2X MyTaq™ (Bioline), 0.4 µM of each primer, 0.1 mg/mL bovine serum albumin, 1.25 U Taq polymerase, 1µL (approximately 50 ng/µL) DNA, and sterile nanopure water to bring to 10 µL volume. Cycling conditions were set at an initial denaturation at 95°C for 1 min, followed by 35 cycles of denaturation at 95°C for 15 s, annealing temperature for 15 s, and extension at 72°C for 10 s, and a final extension step at 72°C for three minutes. PCR products were separated via running on 8% (w/v) polyacrylamide gel in 1X TBE (Tris-borate-EDTA) at 100 V for 90 min. Gels were post-stained with GelRed® (Invitrogen, USA) prior to viewing using Bio-Rad Gel Doc™ XR+ system (Bio-Rad Laboratories, Inc., Hercules, CA, USA).

Genetic and population structure analyses

Band scoring was performed for the polymorphic markers. The number of alleles were determined by direct count and polymorphism information content (PIC) was calculated using

Table 1: Amplification information for polymorphic SSR markers (Song et al. 2010).

Marker		Sequence (5' → 3')	Motif	Expected product size (bp)	T_a (°C)
BARCSOYSSR_08_1494	F	GCGAATGGTTTTTGCTGGAAAGTA	(TAT)22	191	53
	R	GCGCAACGACATTTTCACGAAGTT			
BARCSOYSSR_08_1557	F	GCGGGACTTAGGAGTCTAGTATTA	(TA)25	188	53
	R	GGGAACAGGGATTTAGGACGAGAA			
BARCSOYSSR_08_1032	F	GCGTTCTATGATACCATGTTAAAAGAATGAA	(AT)33	295	52
	R	GCGAAGGAAGAGGAAAGGTTTAAAAGAG			
BARCSOYSSR_08_0458	F	GCGTGGTTTTTCGCTGGATATA	(AT)34	173	50
	R	GCGCATTTCGTAACATATTTTTCAC			
BARCSOYSSR_08_1246	F	GCGAAAGGTCGAGAAAATGAAATG	(AT)24	246	52
	R	GCGTTTGCCTTGTTGGTGACTTG			

F/R – Forward/Reverse; bp – base pairs; T_a – Annealing temperature.

Figure 1: Seeds of representative soybean genotypes from IPB, IITA and NOMIARC collections.

the Gene-Calc computer software (Bińkowski and Miks 2018). Other allele frequency-based parameters Shannon Diversity Index (I), observed (H_o) and expected heterozygosity (H_e) were obtained employing the Genetic Analysis in Excel (GenAlEx *v*. 6.503) software (Peakall and Smouse 2006). The analysis of molecular variance (AMOVA) among and within the soybean clusters was also computed through the latter software with 1000 permutations.

Genetic distances among the soybean genotypes were analyzed using the Dice Similarity Coefficient while clustering was obtained using the neighbor-joining (NJ) method through the Numerical Taxonomy and Multivariate Analysis System (NTSYS) 2.1 software (Rohlf 2002).

Population structure was assessed using the Bayesian model-based STRUCTURE 2.3.4 software program (Pritchard et al. 2000). Parameters for the project were set with a burn-in period of 10,000 and 10,000 Markov chain Monte Carlo (MCMC) replications. Hypothetical value of K (clusters/groups) was set to 1 to 10 with five repeat runs for each value of K. The true value of K was estimated based on the ad hoc statistic Delta K (ΔK) method (Evanno et al. 2005) using the web-based STRUCTURE HARVESTER program (Earl and vonHoldt 2012). Bar plot based on the determined K-value was reconstructed using Excel to align genotypes with clustering results. Results from the determination of the K value was used to obtain the population structure (Q) matrix for subsequent marker-trait association analyses.

Marker-trait association analysis

Marker-trait association analysis was performed using the small number of markers identified as polymorphic in the population under study. Association between SSR markers and seed size-

relevant traits was studied using general linear model (GLM) and mixed linear model (MLM) statistical methods performed with TASSEL5.2.77 (Bradbury et al. 2007). For the MLM (Q + K) procedure, a kinship matrix (K) obtained with Microsatellite Analyzer (MSA) 4.05 (Dieringer et al. 2003) was used in addition to the population structure (Q) matrix used to accomplish GLM analysis. From a significance level of 0.05, a new threshold of 0.01 after Bonferroni correction was used to identify significant associations between SSR markers and seed size traits for both statistical models.

RESULTS AND DISCUSSION

Evaluation of phenotypic variation

Summary statistics of seed size traits measurements are presented in Table 2. Analysis of variance of each of the seed traits single-seed weight, 100-seed weight, length, width and thickness among the genotypes had significant variation at P<0.0001 (Supplementary Table 3a). Correlation analysis showed that SSW is highly correlated with 100-SW, SL, SW and ST at 0.0001 significance level (Supplementary Table 3b).

Table 2: Analysis of seed size traits of 109 soybean genotypes.

Trait	Min	Max	Mean	SD	CV(%)
SSW	83.20	271.00	172.80	28.33	16.39
100-SW	6.45	28.31	16.14	2.52	15.59
SL	6.32	8.71	7.56	0.42	5.61
SW	5.28	7.42	6.40	0.36	5.58
ST	3.81	7.95	5.20	0.40	7.65

SSW in mg; 100-SW in g; SL, SW, and ST in mm.

Assessment of genetic parameters

Samples of PCR amplification of the five SSR markers are shown in Figure 2. Screening of polymorphism of the markers across the 109 soybean genotypes revealed a total of 42 alleles with the number of alleles for each marker ranging from five to 11 and the size of alleles ranging from 173 to 620 bp (Table 3). Major allele frequency ranged from 0.583 to 0.881 with an average of 0.697; this is lower than the reported mean value of 0.8244 by Bisen et al. (2015), which used 23 SSRs for analysis of 38 soybean genotypes in India. On the contrary, our obtained frequency is higher than the 0.467 value obtained by Chauhan et al. (2017) for locus Satt493 from chromosome 8, and the average of 0.54 from 39 SSRs used on 78 soybean varieties in the same study.

Higher observed heterozygosity compared to expected was obtained from three markers; BARCSOYSSR_08_1494, 1557, and 1246. The higher observed heterozygosity may be attributed to soybean genotypes, with, for example, homozygous alleles, that came from different collections or backgrounds but had been used for interbreeding. It may be considered that many of the POP series and some IPB varieties used in this study are crosses from different lines. On the other hand, the use of markers BARCSOYSSR_08_1032 and 0458, returned lower observed heterozygosity, which can be inferred to soybeans occurring as self-pollinating species (Orf, 2008). Mean observed heterozygosity (0.407) was higher than 0.0469 presented by Bisen et al. (2015).

The markers, overall, had relatively high PIC with a range from 0.733 to 0.846 and average of 0.820, indicating high marker informativeness. This finding is higher than the PIC values obtained by Chauhan et al. (2017) and the mean PIC (0.1998) reported for Satt493 (0.678) and higher than mean (0.59) PIC

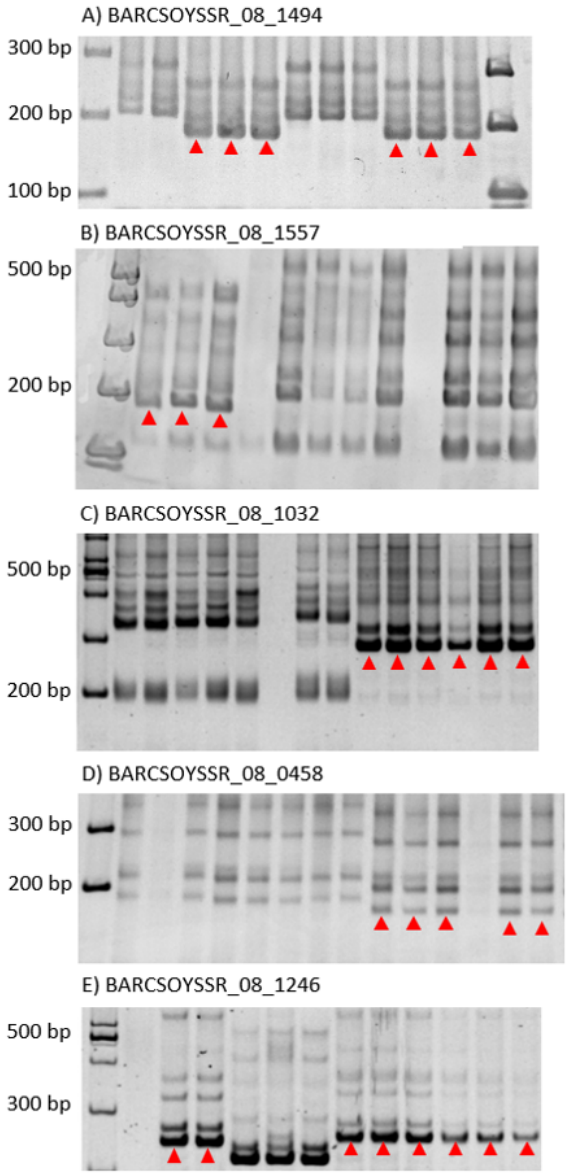

Figure 2: Electropherograms showing representative alleles of the five SSR markers.

from Bisen et al. (2015). Again, the two other genetic studies utilized more SSR markers and lower number of soybean genotypes – factors which could contribute to the variances in measurements obtained.

Clustering and Population Structure

Using distance-based approach, NJ clustering analysis assembled the 109 soybean genotypes into four major clusters: I, II, III and IV (Figure 3A), which included 11, 15, 67 and 16 genotypes, respectively. The genetic similarity coefficients observed were relatively high with the lowest at 0.54 between cluster I and the other three clusters; a coefficient of 1.00 was even observed between/among several genotypes from clusters II, III and IV. Genotypes from the IPB and IITA (TGx series) collections were distributed in the four clusters while 53 out of the 60 (88.33%) NOMIARC genotypes (POP series) analyzed were grouped together under cluster III. The distribution of IPB and IITA genotypes suggests genetic similarity or common alleles while the clustering of the bulk of NOMIARC genotypes give the indication that these are more distinct from the other two collections regardless of the overall high genetic similarity seen in the matrix.

Table 3: Genetic analysis of 109 soybean genotypes using five seed size-associated SSR markers.

SSR Marker	Allele Size Range (bp)	Na	I	Major Allele Frequency	H_o	H_e	PIC
BARCSOYSSR_08_1494	191-290	5	0.662	0.624	0.679	0.469	0.733
BARCSOYSSR_08_1557	188-510	9	0.365	0.881	0.220	0.210	0.842
BARCSOYSSR_08_1032	295-600	8	0.565	0.748	0.229	0.377	0.846
BARCSOYSSR_08_0458	173-340	9	0.647	0.651	0.073	0.454	0.837
BARCSOYSSR_08_1246	220-620	11	0.679	0.583	0.835	0.486	0.840
Mean		8.4	0.584	0.697	0.407	0.399	0.820

Na – Number of alleles; I – Shannon Diversity Index; H_o – Observed Heterozygosity;
H_e – Expected Heterozygosity PIC – Polymorphism Information Content

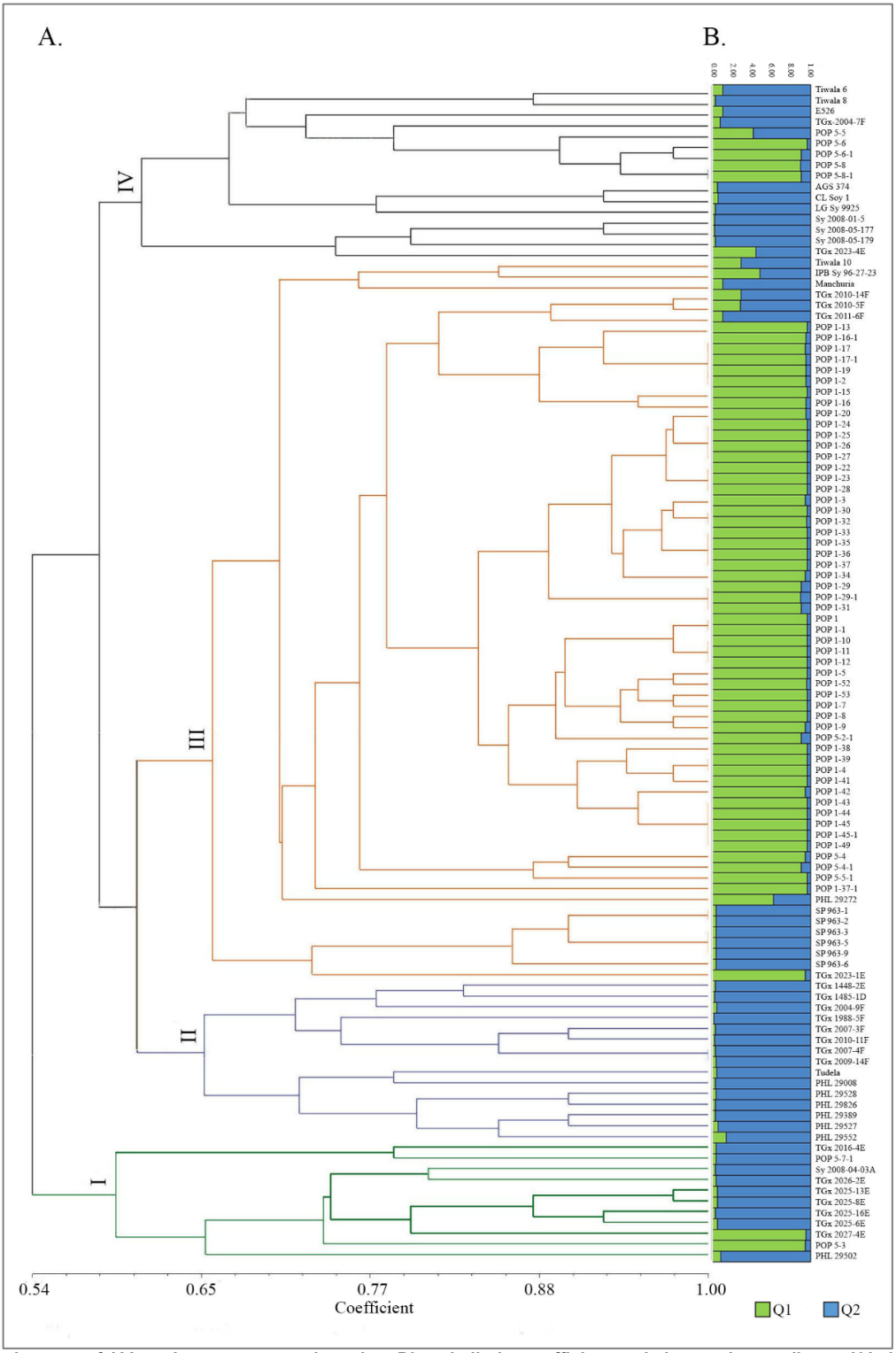

Figure 3: (A) Dendrogram of 109 soybean genotypes based on Dice similarity coefficient and clustered according to NJ algorithm; and (B) estimated population structure showing fractional memberships in each of the two subpopulations (K=2), Q1 and Q2.

Table 4: Analysis of molecular variance (AMOVA) among and within two soybean populations (K=2) inferred through population structure analysis.

Source	df	SS	MS	Est. Var.	%	*P*
Between populations	1	68.263	68.263	1.245	48%	0.001
Within populations	102	145.131	1.356	1.356	52%	
Total	108	213.394		2.602	100%	

df – degrees of freedom; SS – the sums of squares; MS – the mean sums of squares; Est. Var. – the estimated variance.

For model-based approach, the genetic structure was determined using STRUCTURE ver. 2.3.4 software. The highest delta *K* value was observed when K=2 based on the delta K method of Evanno et al. (2005) (Figure 4). Hence, two subpopulations (K) were inferred. Subpopulations were denoted as Q1 and Q2, comprised of 61 and 48 genotypes, respectively. Classification of genotypes between Q1 and Q2 were based on the degree of membership for respective subpopulations based on the Q matrices for K=2 runs. Fifty-eight out of 60 NOMIARC genotypes gathered in the Q1 subpopulation along with two genotypes from IITA (TGx 2023-1E and TGx 2027-4E) and one from IPB collection (PHL29272). On the other hand, the Q2 subpopulation consisted of the genotypes from the IPB and IITA collections (Figure 3B) along with two other NOMIARC genotypes (POP 5-5 and POP 5-7-1). To some extent, results from STRUCTURE appear consistent with the clustering via NJ tree wherein the Q1 subpopulation, which mainly consisted of NOMIARC genotypes, was largely parallel with cluster III comprised of the majority of NOMIARC genotypes (Figure 3A and 3B).

Analysis of molecular variance (AMOVA) showed a relatively moderate estimated variance between the two (K=2) identified populations (*P*-value 0.001) (Table 4), It is deduced then that the first subpopulation (Q1) composing of almost entirely NOMIARC genotypes have moderate differentiation to the other subpopulation (Q2) consisting mostly of genotypes from IPB and IITA collections.

Marker-trait association

The covariate Q generated from STRUCTURE was used for general (GLM) and mixed linear model (MLM) analyses for assessment of marker-trait associations. In the GLM test, significant associations were observed between BARCSOYSSR_08_1032 and single seed weight (0.0004), length (0.0004) and width (0.0003), and between BARCSOYSSSR_08_1246 and seed length (0.0046) (Supplementary Table 4). Based on this model, BARCSOYSSR_08_1032 emerges with the potential for use in MAS. However, in MLM analysis, no significant associations were observed between any of the markers and seed size traits (Supplementary Table 5). These results were examined at a significance level of α=0.05, with the new threshold set at 0.01 after Bonferroni correction to reduce the chances of obtaining false positives. Preferably, significant associations were also observed in MLM as it considers additional kinship data in comparison to GLM, and therefore, offers more thorough associations (Mikic et al. 2016). In the comparison of different statistical models for association mapping in soybean and maize, Kaler and colleagues (2020) reported that GLM was among the models which yielded large numbers of significant QTLs while MLM was among the complex models that did not return any significant markers.

Observation for potential environment effects on seed weight

Quantitative traits, such as seed size, are complex traits wherein expressed phenotypic variations are affected by genetic background, reactivity to environmental conditions and by genotype-environment interactions (GxE) (Mackay 2001; Mitchell-Olds 2013). Available data on 100-SW from Bukidnon, Isabela and IPB, Los Baños at wet and/or dry seasons (Supplementary Table 6) were gathered and were found to have highly significant variation (*P*-value 0.000 at 0.05 significance level) within genotypes from the different environments via ANOVA (Supplementary Table 7). These findings suggest that there may be environmental effects on the seed trait. Teng and colleagues (2009) have previously reported that quantitative traits loci (QTL) of soybean seed weight with genetic effects and QTL x environment (QE) interactions can possibly cause changes in the trait during seed development.

In summary, assessment of genetic parameters revealed a wide range of heterozygosity and high marker informativeness of the five SSR markers. Distance and model-based clustering showed NOMIARC genotypes consistently grouped together unlike those from IPB and IITA collections. Through association analysis via GLM, BARCSOYSSR_08_1032 was found to be associated to single-seed weight, length and width and BARCSOYSSR_08_1246 to seed length. Even though the markers were only associated to several or only one of the seed size traits, results forward candidate markers to pursue for further researches – using larger populations with multiple genetic backgrounds – as the seed traits measured in the study were significantly correlated to each other. It would also be preferable to find markers with significant association with seed size traits in both GLM and MLM analyses as the latter offers greater statistical power. As such, employing a more comprehensive set of markers encompassing soybean genome is vital for subsequent marker screening. It is also recommended that future investigations consider performing association analyses in separate environments given that the determination of links between markers and seed size traits proves to be complex as gene effects and GxE interactions are involved in the phenotypic expression. This study paved the commencement of the wide and intricate scope of marker screening for association with the quantitative seed size traits; ultimately, for the practical application of MAS in the improvement of soybean breeding in the country.

ACKNOWLEDGMENT

The research was accomplished through the DOST-PCAARRD project “Soybean Variety Development for Large Seed Size, Higher Yields and Enhanced Functional Properties” under the Soybean Program with fund code: N9-159-21. We thank Mr. Lizk Gil P. Villegas for his assistance during DNA extraction activities. We also acknowledge Ms. Lorena V. Duna and team from DA-NOMIARC, Bukidnon, and Ms. Vanessa Joy F. Calderon and team from DA-CVRC (Department of Agriculture – Cagayan Valley Integrated Agricultural Research Center), Isabela, for the soybean 100-seed weight data from soybean project reports.

CONFLICT OF INTEREST

The authors have no conflict of interests to declare.

AUTHOR CONTRIBUTIONS

AMDLR and MMLS were the project researchers. VAL and RGM provided plant materials and facilitated the conduct of field activities. EEE rendered administrative role for soybean program implementation. ETMO was the study leader.

REFERENCES

Bińkowski J, Miks S. Gene-Calc [Computer software]. 2018. Available from: www.gene-calc.pl

Bisen A, Khare D, Nair P, Tripathi N. SSR analysis of 38 genotypes of soybean (Glycine Max (L.) Merr.) genetic diversity in India. Physiol Mol Biol Plants 2015; 21(1):109-115.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 2007; 23(19):2633-2635.

Burton JW. Quantitative genetics: results relevant to soybean breeding. In: Wilcox JR, ed. Soybeans: Improvement, Production and Uses. 2nd Edition. Madison, WI: Agron. Monogr. 16. ASA, CSSA, and SSSA, 1987:211-247.

Chauhan DK, Bhat JA, Thakur AK, Hussain Z, Satyavathi CT. Understanding genetic relationship and population structure of Indian soybean varieties using microsatellite markers. Proc Natl Acad Sci India Sect B Biol Sci 2017; 88(3):1091-1100.

Chaves MVA, Silva NS, Jorge GL, Silveira IC, Medeiros LA, Hamawaki RL, Hamawaki OT, Nogueira APO, Hamawaki CDL. Genotype x environment interaction and stability of soybean cultivars for vegetative-stage characters. Genet Mol Res 2017; 16(3):gmr16039795.

Chen P. Developing high quality identity-preserved soybean for the specialty soyfood market. In: Production and Marketing of Identity-Preserved Soybean. MO, USA: American Soybean Association 2004; 23–31.

Collard BCY, Mackill DJ. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philo Trans R Soc Lond B Biol Sci 2008; 363(1491):557-572.

Dieringer D, Schlötterer C. MICROSATELLITE ANALYSER (MSA): a platform independent analysis tool for large microsatellite data sets. Mol Ecol Notes 2003; 3(1):167-169.

Earl DA, vonHoldt BM. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genet Resour 2012; 4:359-361.

Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol Ecol 2005; 14(8):2611–20.

Galang VMP. Soybean industry targeted for improved production, logistics — DOST. BusinessWorld, 2020. Retrieved September 21, 2022 from https://www.bworldonline.com/editors-picks/2020/01/13/272880/soybean-industry-targeted-for-improved-production-logistics-dost/

Gaonkar V, Rosentrater KA. Soybean. In: Pan Z, Zhang R, Zicari, S, eds. Integrated Processing Technologies for Food and Agricultural By-Products. Academic Press, 2019:73-104.

Ge L, Yu J, Wang H, Luth D, Bai G, Wang K, Chen R. Increasing seed size and quality by manipulating *BIG SEEDS1* in legume species. Proc Natl Acad Sci USA 2016; 113(44):12414–12419.

Hao DR, Cheng H, Yin ZT, Cui SY, Zhang D, Wang H, Yu DY. Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. Theor Appl Genet 2012; 124:447-458.

Kaler AS, Gillman JD, Beissinger T, Purcell LC. Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. Front Plant Sci 2020; 10:1794.

Lee S, Freewalt KR, McHale LK, Song Q, Jun TH, Michel AP, Dorrance AE, Rouf Mian MA. A high-resolution genetic linkage map of soybean based on 357 recombinant inbred lines genotyped with BARCSoySNP6K. Mol Breeding 2015; 35:58.

Li J, Zhao J, Li Y, Gao Y, Hua S, Nadeem M, Sun G, Zhang W, Hou J, Wang X, Qiu L. Identification of a novel seed size associated locus SW9-1 in soybean. Crop J 2019; 7(4):548-559.

Liu D, Yan Y, Fujita Y, Xu D. Identification and validation of QTLs for 100-seed weight using chromosome segment substitution lines in soybean. Breed Sci 2018; 68(4):442-448.

Mackay TF. Q&A: Genetic analysis of quantitative traits. J Biol 2001; 8:23.

Mamidala P, Nanna RS. A simple and rapid method for DNA extraction from leaves of tomato, tobacco and rape seed. J Phytol 2009; 1(6):388-390.

Mikic S, Kondic-Spika A, Brbaklic L, Stanisavljevic D, Trkulja D, Tomicic M, Nastasic A, Kobiljski B, Prodanovic S, Momirovic GS. Multiple marker-traits associations for maize agronomic traits. Chilean J Agric Res 2016; 76(3):300-306.

Mitchell-Olds T. Selection on QTL and complex traits in complex environments. Mol Ecol 2013; 22(13):3427-3429.

Mwiinga B, Sibiya, J, Kondwakwenda A, Musvosvi C, Chigeza G. Genotype x environment interaction analysis of soybean (Glycine max (L.) Merrill) grain yield across production environments in Southern Africa. Field Crops Res 2020; 256:107922.

Orf JH. 2 - Breeding, genetics and production of soybeans. In: Johnson LA, White PJ, Galloway R, eds. Soybeans. AOCS Press, 2008:33-65.

Pagano MC, Miransari M, Corrêa EJA, Duarte NF, Yelikbayev BK. Genomic research favoring higher soybean production. Curr Genomics 2020; 21(7):481-490.

Pagano MC, Miransari M. The importance of soybean production worldwide. In: Miransari M, ed. Abiotic and Biotic Stresses in Soybean Production. Academic Press, 2016:1-26.

Peakall R, Smouse PE. GENALEX 6: Genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes 2006; 6(1):288-295.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multi-locus genotype data. Genetics 2000; 155(2):945-959.

Rohlf FJ. NTSYS-Pc: Numerical taxonomy and multivariate analysis system. Version 2.1. 2002. New York: Exeter Publishing Led.

Smith TJ, Camper H. Effects of seed size on soybean performance. Agron J 1975; 67(5):681-684.

Song QJ, Jia G. Zhu Y, Grant D, Nelson R, Hwang EY, Hyten DL, Cregan PB. Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1.0) in soybean. Crop Sci 2010; 50(5):1950-1960.

Teng W. Han Y, Du Y, Sun D, Zhang Z, Qiu L, Sun G, Li W. QTL analyses of seed weight during the development of soybean (*Glycine max* L. Merr.). Heredity (Edinb) 2009; 102(4):372-380.

USDA. World Agricultural Supply and Demand Estimates. 2022. Retrieved January 24, 2022 from https://usda.library.cornell.edu/concern/publications/3t945q76s?locale=en

Zhang J, Song Q, Cregan PB, Jiang GL. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycine max). Theor Appl Genet 2016; 129:117-130.

SUPPLEMENTARY DATA

Supplementary Table 1: Soybean genotypes used for SSR markers screening and their clustering and subpopulation classifications.

No	Collection	Germplasm	Genotype	Description	Cluster	Subpopulation
1	IPB	Sy 194156	Tiwala 6	National checks	IV	Q2
2	IPB	Sy 194157	Tiwala 8		IV	Q2
3	IPB	Sy 194158	Tiwala 10		III	Q2
4	IPB	Sy 194085	IPB Sy 96-27-23	IPB line; selected based on good flavor of seeds.	III	Q2
5	IPB	Sy 194159	Manchuria	Farmer's variety; from Mindanao collection. Pedigree unknown but is accounted to be originally from Manchuria, China.	III	Q2
6	IPB	Sy 194086	SP 963-1	IPB lines; selections from cross between Tiwala 8 and LX18a, a lipoxygenase-null soybean. High-yielding and has lipoxygenase-null quality.	III	Q2
7	IPB	Sy 194087	SP 963-2		III	Q2
8	IPB	Sy 194088	SP 963-3		III	Q2
9	IPB	Sy 194089	SP 963-5		III	Q2
10	IPB	Sy 194090	SP 963-6		III	Q2
11	IPB	Sy 194092	SP 963-9		III	Q2
12	IPB	Sy 194079	Sy 2008-01-05	New IPB variety releases	IV	Q2
13	IPB	Sy 194080	Sy 2008-04-03A		I	Q2
14	IPB	Sy 194081	Sy 2008-05-177		IV	Q2
15	IPB	Sy 194082	Sy 2008-05-179		IV	Q2
16	IPB	Sy 194084	E526	Cross between 147-PHL X IPB-Sy-96-27-23-2020-05	IV	Q2
17	IPB	Sy 194083	AGS 374	IPB-developed variety; high-yielding and large-seeded compared to checks	IV	Q2
18	IPB	Sy 194160	CLSoy 1	Introduction from Thailand and was adopted by Central Luzon State University (CLSU) in Nueva Ecija, Philippines.	IV	Q2
19	IPB	Sy 194162	LG Sy 9925	Local collection	IV	Q2
20	IPB	Sy 194160	Tudela	Local collection	II	Q2
21	IPB	Sy 194072	PHL 29008	National Plant Genetics Resources Laboratory (NPGRL)	II	Q2
22	IPB	Sy 194070	PHL 29272		III	Q1
23	IPB	Sy 194073	PHL 29389		II	Q2
24	IPB	Sy 194074	PHL 29502		I	Q2
25	IPB	Sy 194069	PHL 29527		II	Q2
26	IPB	Sy 194076	PHL 29528		II	Q2
27	IPB	Sy 194075	PHL 29552		II	Q2
28	IPB	Sy 194078	PHL 29826		II	Q2
29	IITA	Sy 194049	TGx 1448-2E	International Institute of Tropical Agriculture (IITA)	II	Q2
30	IITA	Sy 194040	TGx 1485-1D		II	Q2
31	IITA	Sy 194039	TGx 1988-5F		II	Q2
32	IITA	Sy 194026	TGx 2004-7F		IV	Q2
33	IITA	Sy 194025	TGx 2004-9F		II	Q2
34	IITA	Sy 194030	TGx 2007-3F		II	Q2
35	IITA	Sy 194028	TGx 2007-4F		II	Q2
36	IITA	Sy 194024	TGx 2009-14F		II	Q2
37	IITA	Sy 194043	TGx 2010-11F		II	Q2
38	IITA	Sy 194027	TGx 2010-14F		III	Q2
39	IITA	Sy 194029	TGx 2010-5F		III	Q2
40	IITA	Sy 194042	TGx 2011-6F		III	Q2
41	IITA	Sy 194050	TGx 2016-4E		I	Q2
42	IITA	Sy 194037	TGx 2023-1E		III	Q1
43	IITA	Sy 194033	TGx 2023-4E		IV	Q2
44	IITA	Sy 194044	TGx 2025-13E		I	Q2
45	IITA	Sy 194032	TGx 2025-16E		I	Q2
46	IITA	Sy 194046	TGx 2025-6E		I	Q2
47	IITA	Sy 194045	TGx 2025-8E		I	Q2
48	IITA	Sy 194056	TGx 2026-2E		I	Q2
49	IITA	Sy 194034	TGx 2027-4E		I	Q1
50	NOMIARC	Sy 194093	POP 1	Northern Mindanao Integrated Agricultural Research Center (NOMIARC)	III	Q1
51	NOMIARC	Sy 194094	POP 1-1		III	Q1
52	NOMIARC	Sy 194102	POP 1-10		III	Q1
53	NOMIARC	Sy 194103	POP 1-11		III	Q1
54	NOMIARC	Sy 194104	POP 1-12		III	Q1

No	Collection	Germplasm	Genotype	Description	Cluster	Subpopulation
55	NOMIARC	Sy 194105	POP 1-13		III	Q1
56	NOMIARC	Sy 194106	POP 1-15		III	Q1
57	NOMIARC	Sy 194107	POP 1-16		III	Q1
58	NOMIARC	Sy 194108	POP 1-16-1		III	Q1
59	NOMIARC	Sy 194109	POP 1-17		III	Q1
60	NOMIARC	Sy 194110	POP 1-17-1		III	Q1
61	NOMIARC	Sy 194111	POP 1-19		III	Q1
62	NOMIARC	Sy 194095	POP 1-2		III	Q1
63	NOMIARC	Sy 194112	POP 1-20		III	Q1
64	NOMIARC	Sy 194113	POP 1-22		III	Q1
65	NOMIARC	Sy 194114	POP 1-23		III	Q1
66	NOMIARC	Sy 194115	POP 1-24		III	Q1
67	NOMIARC	Sy 194116	POP 1-25		III	Q1
68	NOMIARC	Sy 194117	POP 1-26		III	Q1
69	NOMIARC	Sy 194118	POP 1-27		III	Q1
70	NOMIARC	Sy 194119	POP 1-28		III	Q1
71	NOMIARC	Sy 194121	POP 1-29		III	Q1
72	NOMIARC	Sy 194122	POP 1-29-1		III	Q1
73	NOMIARC	Sy 194096	POP 1-3		III	Q1
74	NOMIARC	Sy 194123	POP 1-30		III	Q1
75	NOMIARC	Sy 194124	POP 1-31		III	Q1
76	NOMIARC	Sy 194125	POP 1-32		III	Q1
77	NOMIARC	Sy 194126	POP 1-33		III	Q1
78	NOMIARC	Sy 194127	POP 1-34		III	Q1
79	NOMIARC	Sy 194129	POP 1-35		III	Q1
80	NOMIARC	Sy 194130	POP 1-36		III	Q1
81	NOMIARC	Sy 194131	POP 1-37		III	Q1
82	NOMIARC	Sy 194132	POP 1-37-1		III	Q1
83	NOMIARC	Sy 194133	POP 1-38		III	Q1
84	NOMIARC	Sy 194134	POP 1-39		III	Q1
85	NOMIARC	Sy 194097	POP 1-4		III	Q1
86	NOMIARC	Sy 194135	POP 1-41		III	Q1
87	NOMIARC	Sy 194136	POP 1-42		III	Q1
88	NOMIARC	Sy 194137	POP 1-43		III	Q1
89	NOMIARC	Sy 194138	POP 1-44		III	Q1
90	NOMIARC	Sy 194139	POP 1-45		III	Q1
91	NOMIARC	Sy 194140	POP 1-45-1		III	Q1
92	NOMIARC	Sy 194141	POP 1-49		III	Q1
93	NOMIARC	Sy 194098	POP 1-5		III	Q1
94	NOMIARC	Sy 194142	POP 1-52		III	Q1
95	NOMIARC	Sy 194143	POP 1-53		III	Q1
96	NOMIARC	Sy 194099	POP 1-7		III	Q1
97	NOMIARC	Sy 194100	POP 1-8		III	Q1
98	NOMIARC	Sy 194101	POP 1-9		III	Q1
99	NOMIARC	Sy 194144	POP 5-2-1		III	Q1
100	NOMIARC	Sy 194145	POP 5-3		I	Q1
101	NOMIARC	Sy 194146	POP 5-4		III	Q1
102	NOMIARC	Sy 194147	POP 5-4-1		III	Q1
103	NOMIARC	Sy 194148	POP 5-5		IV	Q2
104	NOMIARC	Sy 194149	POP 5-5-1		III	Q1
105	NOMIARC	Sy 194150	POP 5-6		IV	Q1
106	NOMIARC	Sy 194151	POP 5-6-1		IV	Q1
107	NOMIARC	Sy 194153	POP 5-7-1		I	Q2
108	NOMIARC	Sy 194154	POP 5-8		IV	Q1
109	NOMIARC	Sy 194155	POP 5-8-1		IV	Q1

Supplementary Table 2: Seventeen seed weight-associated simple sequence repeat (SSR) markers used for polymorphic screening (Song et al., 2010).

Primer		Sequence (5' → 3')	Product size (bp)	Annealing Temperature (°C)
BARCSOYSSR_08_0458	F	GCGTGGTTTTTCGCTGGATATA	173	50
	R	GCGCATTTCGTAACATATTTTTCAC		
BARCSOYSSR_08_0909	F	GCGATACACGTGACGAAGAGAATCATGC	186	48
	R	GCGATAACATAATGGATTGACATAAAT		
BARCSOYSSR_08_1142	F	GCGGGACGCAAAATTGGATTTAGT	274	55
	R	GCGCCGAATAAAACGTGAGAACTG		
BARCSOYSSR_08_0849	F	GCGCAATTTGACTATTTTTAGCTGTTG	281	48
	R	GCGCGATATAAGATGATTTTTATTGAT		
BARCSOYSSR_08_0579	F	CAACCTGTATTCCACAAAAAATCTCACC	267	52
	R	GCGCCCCAATTTGACTATAAATAAAAGT		
BARCSOYSSR_08_0506	F	GCGTAGCAACAAAGCAATCTACAG	259	54
	R	GCGTCCCATTTTATTCCACACTATGTAAT		
BARCSOYSSR_08_0460	F	GCGGTTTTGCAAGATGTGATGAGT	274	51
	R	GCGCGTACGCAAAATTTATATTCA		
BARCSOYSSR_08_1361	F	GACCATTATGCTTTTTCTCTTT	258	45
	R	CTTTATCCTCCTCCATTAGTTCTA		
BARCSOYSSR_08_1164	F	GCGGCCTCCTCAAAATAATAATAAT	269	50
	R	GCGGTTAGGAATAGAAGATAGTTACATT		
BARCSOYSSR_08_1032	F	GCGTTCTATGATACCATGTTAAAAGAATGAA	295	52
	R	GCGAAGGAAGAGGAAAGGTTTAAAAGAG		
BARCSOYSSR_08_1480	F	GTTCTAGTTCTTCTTTTCACTTG	260	44
	R	TTGTCATCAAATATCATCCATTT		
BARCSOYSSR_08_1494	F	GCGAATGGTTTTTGCTGGAAAGTA	191	53
	R	GCGCAACGACATTTTCACGAAGTT		
BARCSOYSSR_08_1557	F	GCGGGACTTAGGAGTCTAGTATTA	188	53
	R	GGGAACAGGGATTTAGGACGAGAA		
BARCSOYSSR_08_1245	F	GCGTGAGCAATACATTTTCTGTAAAAAAAAA	293	51
	R	GCGATAGCATATTAAATGAATCTCTGACTTT		
BARCSOYSSR_08_1216	F	GCGCGTCCTTTATTTAATTTAATATGAA	283	48
	R	GCGTGGCTTTGCTAATAATGAATGAT		
BARCSOYSSR_08_1637	F	CGGATTGTGTCCTTTGTTGTTATTAT	252	50
	R	ACCTCGCACACAATTTGAGTC		
BARCSOYSSR_08_1246	F	GCGAAAGGTCGAGAAAATGAAATG	246	52
	R	GCGTTTGCCTTGTTGGTGACTTG		

F – forward primer; R – reverse primer.

Supplementary Table 3a: ANOVA test for seed size traits of different soybean genotypes.

Trait	Source	Df	Type III SS	Mean Square	F Value	Pr > F
Single-seed weight	Gen	108	173845.7553	1609.6829	4.54	<.0001
100-seed weight	Gen	108	1392.9698	12.8979	5.87	<.0001
Seed length	Gen	108	40.8008	0.3778	5.33	<.0001
Seed width	Gen	108	27.2811	0.2526	4.44	<.0001
Seed thickness	Gen	108	30.4952	0.2824	3.12	<.0001

df – degrees of freedom; SS – sum of squares

Supplementary Table 3b: Pearson correlation coefficients of seed size characters of 109 soybean genotypes as evaluated by SAS®.

	Single-seed weight	100-seed weight	Seed length	Seed width	Seed thickness
Single-seed weight		0.840* <.0001	0.650* <.0001	0.759* <.0001	0.658* <.0001
100-seed weight			0.639* <.0001	0.732* <.0001	0.635* <.0001
Seed length				0.707* <.0001	0.454* <.0001
Seed width					0.706* <.0001
Seed thickness					

*Statistically significant (*P*<0.0001); Prob > |r| under H0: Rho=0.

Supplementary Table 4: *P*-values obtained from General Linear Model (GLM) association analysis via TASSEL software.

Seed trait	BARCSOYSSR_08_1494		BARCSOYSSR_08_1557		BARCSOYSSR_08_1032		BARCSOYSSR_08_0458		BARCSOYSSR_08_1246	
	F	P	F	P	F	P	F	P	F	P
Weight	0.0061	0.9378	0.0856	0.7705	13.3378	0.0004*	2.6920	0.1038	3.0890	0.0817
100-seed weight	0.0139	0.9063	0.3860	0.5357	6.0440	0.0156	5.8895	0.0169	3.0790	0.0822
Length	0.3260	0.5692	4.7321	0.0318	13.3023	0.0004*	0.6734	0.4137	8.3929	0.0046*
Width	0.0414	0.8392	0.8898	0.3477	14.3686	0.0003*	6.6206	0.0115	1.4840	0.2259
Thickness	0.4219	0.5174	4.1402	0.0444	6.6581	0.0112	6.0393	0.0156	0.1591	0.6908

*Significant after adjusted α = 0.01.

Supplementary Table 5: *P*-values obtained from Mixed Linear Model (MLM) association analysis via TASSEL software.

Seed trait	BARCSOYSSR_08_1494		BARCSOYSSR_08_1557		BARCSOYSSR_08_1032		BARCSOYSSR_08_0458		BARCSOYSSR_08_1246	
	F	P	F	P	F	P	F	P	F	P
Weight	0.1149	0.7353	0.0021	0.9639	2.5189	0.1155	0.2510	0.6174	1.8460	0.1771
100-seed weight	0.0456	0.8313	0.0644	0.8001	1.5004	0.2233	2.6103	0.1091	2.0047	0.1597
Length	0.1530	0.6965	1.8782	0.1734	0.7155	0.3995	0.2013	0.6546	3.2465	0.0744
Width	0.0642	0.8005	0.1350	0.7140	2.2029	0.1407	1.5596	0.2145	0.5610	0.4555
Thickness	0.0067	0.9348	1.6108	0.2072	1.0405	0.3100	1.8043	0.1821	0.0008	0.9769

Adjusted α = 0.01.

Supplementary Table 6: Soybean 100-Seed weight data from different environments in the Philippines.

Genotype	Average 100-Seed Weight (g)				Mean	Standard Deviation
	Bukidnon DS (2019)	Bukidnon WS (2019)	Isabela DS (2018-19)	IPB, Los Baños (DS 2019-20)		
(1) Tiwala6	14.25	13.53	11.88	15.34	13.75	1.26
(2) Tiwala8	15.25	14.47	12.51	15.74	14.49	1.23
(3) Tiwala10	14.50	12.75	11.40	14.21	13.21	1.24
(4) IPBSy96-27-23	21.25	21.40	-	19.38	20.68	0.92
(5) Manchuria	15.12	17.43	15.45	15.89	15.97	0.88
(6) SP963-1	18.25	16.07	-	17.57	17.30	0.91
(7) SP963-2	15.75	14.40	-	15.52	15.22	0.59
(8) SP963-3	17.75	15.47	-	15.92	16.38	0.99
(9) SP963-5	17.25	15.45	-	15.44	16.05	0.85
(10) SP963-6	18.00	14.90	-	15.34	16.08	1.37
(11) SP963-9	17.25	15.60	-	15.31	16.05	0.85
(12) Sy2008-01-05	14.25	14.72	13.09	14.84	14.23	0.69
(13) Sy2008-04-3A	14.25	14.18	12.58	13.56	13.64	0.67
(14) Sy2008-05-177	14.50	16.10	12.61	15.27	14.62	1.29
(15) Sy2008-05-179	15.25	16.18	13.09	14.80	14.83	1.12
(16) E526	15.72	16.93	15.96	13.94	15.64	1.08
(17) AGS374	21.00	22.83	-	15.82	19.88	2.97
(18) CLSOY 1	14.25	12.28	11.00	13.24	12.69	1.20
(19) LGSY	14.00	13.55	-	15.16	14.24	0.68
(20) Tudela	-	-	10.88	11.40	11.14	0.26

(-) Data not available;

Supplementary Table 7: ANOVA test (α=0.05) for means of 100-Seed Weight of 20 genotypes from Bukidnon, Isabela and IPB, Los Baños.

Source of Variation	SS	df	MS	F	*P*-value	F crit
Between Genotypes	285.4045351	19	15.0213	7.7957	3.60×10^{-9}	1.8029
Within Genotypes	94.41624167	49	1.9269			
Total	379.8207768	68				

SS – sum of squares; df – degrees of freedom; MS – mean square