Identification of Prospective Soybean Accessions for the Creation of a Genebank Core Collection Based on High Density DNA Marker Data

D Satyawan* and I M Tasma

Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development (ICABIOGRAD-IAARD), Jl. Tentara Pelajar 3A, Cimanggu, Bogor 16111, Indonesia

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

Abstract. Genebanks typically collect and conserve as many existing genetic variations as possible. However, a smaller core population that is representative of the whole collection but maintains as much diversity as possible is more practical for in-depth phenotypic characterization, genetic studies, and plant breeding. Here we describe genotype-based core collection development from USDA soybean germplasm collection, which contains 20,095 accessions that had been genotyped using 50,000 single nucleotide polymorphism DNA markers. Analysis using trimmed marker data and Core Hunter 3 software identified 382 accessions with maximum genetic distances from each other. Population structure analysis of the selected accessions indicated that the software favored admixed individuals instead of those with pure ancestor population genotypes. Allele frequencies of most markers are largely conserved in the core population, which indicates successful maintenance of alleles in the core population. Phenotypic variations for branch number and yield were largely similar to the original population, but slightly different for seed weight and maturity, where very high seed weight and high-latitude maturity groups were less prominent in the core collection. The resulting core population can be used as a base population for any institutions interested in utilizing worldwide soybean genetic diversity for breeding and genetic studies.

1. Introduction

Soybean is an economically important food crop in Indonesia with a significant demand that could not be fulfilled by domestic production [1]. Part of the reason is because the average on-farm productivity in Indonesia is relatively low at around 1.5 tons/ha, compared to 3 tons/ha achieved in major soybean-producing countries in the world [2]. Plant breeding can contribute a solution to this problem by producing higher yielding cultivars, which will improve profit per unit of land and further incentivize farmers to plant more soybeans.

Since plant breeding can be simplistically viewed as an effort to accumulate beneficial genes from multiple parental lines into a single cultivar, identification of potentially useful parental lines is an important step in plant breeding. Candidate parents can be chosen based on certain traits of interest or certain genetic backgrounds that maximize the possibility of obtaining useful alleles [3]. Consequently, a rich collection of soybean accessions that contains diverse genetic and phenotypic variations will be
very advantageous for breeding programs, as it will provide more options for genes and traits that can be incorporated in future soybean cultivars.

Although breeders can store and manage their own genetic resources that they need, genetic resource collection and management is more effectively handled by genebanks, since they specialize in large scale collection and long term maintenance of a wide range of organisms. Nevertheless, genebanks are not always equipped to perform thorough phenotypic and genotypic evaluations on their collections, since large-scale evaluations can be costly and time-consuming. As a result, they often cannot assist breeders in choosing the correct accessions that can serve their needs [4]. One of the approaches to address this problem is by developing a core collection, which is a smaller subset of the collection that still retains most of the genetic and phenotypic diversity of the whole collection but more manageable for detailed evaluations [5].

Unfortunately, core collection development still requires either phenotypic or genotypic data of the whole collection to be used as the basis for screening and selecting individual accessions [6]. This can present a chicken and egg dilemma for resource-limited genebanks or breeding programs. To solve this conundrum, we propose the use of data from USDA soybean germplasm collection, where more than 20,000 soybean accessions from around the world had been genotyped using 50,000 single nucleotide polymorphism (SNP) markers [7] and the majority of those accessions have also been characterized phenotypically [8]. However, phenotyping is deeply affected by the environment and the environment in the US could be very different from Indonesia. Thus, accessions selected based on a certain phenotype in the US may not produce the same phenotype when those accessions are grown in Indonesia. Consequently, we focused on using only the genotype data to screen the accessions and generate the core collection, since phenotypes are basically the expression of genotypes [9].

The genotype-based core collection development focused on selecting accessions with maximum genetic dissimilarity from each other, while maintaining the allelic richness of the original population. Once the core accessions were selected, their genetic properties and the distribution of phenotypic values in the core collection were evaluated to ensure that the genetic diversity is also reflected in the phenotypic diversity. The accessions identified in this study can be requested from USDA using their GRIN website [10] to enrich our soybean collection and encourage their utilization for future breeding and genetic studies.

2. Materials and methods

2.1. Genotypic data acquisition and core collection identification
The USDA soybean germplasm SNP data was downloaded from Soybase [7] in vcf format and based on Wm82.a1 reference genome. The population SNP data was then opened in the SNPRelate package in R [11], and the number of SNP was trimmed using ‘snpgdsLDpruning’ command on default settings. The pruned SNP data was then used to identify the core collection in Core Hunter 3 package [12] using default settings to generate a population size of 0.2x of the original population size, and 0.019x of the original population size.

2.2. Evaluation of genetic characteristics of the core population
Genetic characterization was performed by comparing the core population with the same number of random accessions and some Indonesian accessions sequenced at ICABIOGRAD [13,14]. Random accessions were selected by assigning a number to each accession using the ‘RANDBETWEEN’ function in Microsoft Excel, followed by selection of 382 or 4000 accessions with the smallest assigned random number. Unless otherwise stated, genotypic characterization was performed using the pruned SNP data sets outlined in section 2.1.

Phylogenetic analysis was carried out in TASSEL 5 [15] using the ‘Create tree’ command and ‘neighbor-joining’ option selected. The resulting tree was exported as a newick file, and visualized using iTol [16]. Principal Component Analysis (PCA) was also performed in TASSEL using default options, and the resulting EigenValues were charted using ggplot2 library in R [17].
The software ADMIXTURE [18] was used to assess genetic admixture in the core population, using the default options and k-value from 1 to 12 to simulate the existence of 1 to 12 hypothetical subpopulations. Genetic admixture was plotted in iToL using multibar plot option alongside the phylogenetic tree. Allele frequencies in each population was calculated using Vcftools [19], with the ‘--keep’ command added to select the core or random population and ‘--freq’ command to output the allele frequency values in the population. The frequency distribution was then summarized using boxplot in ggplot2 library in R.

2.3. Phenotypic distribution analysis in selected populations

The phenotypic data of the USDA soybean germplasm collection was downloaded from NPGS GRIN-Global Training Site 1.10.6 [8], with special interests on branch number, maturity group, seed weight, and overall yield. The distribution of each phenotype in the original, random, and core population was charted using boxplot in ggplot2 library in R.

3. Results and discussions

3.1. Identification of candidate core accessions

The Core Hunter 3 package was designed to select representative accessions by maximizing the minimum distance between accessions [12], hence all selected accessions will be as different as possible from other selected accessions. Since this study completely relied on genotypic (SNP) data, the package selected the accessions based on genetic distances between them. The SNP data were first pruned to eliminate markers that are located in the same linkage disequilibrium (LD) block and correlated with each other ($r^2 > 0.2$) in order to eliminate ascertainment bias [20]. The pruning step also reduced the marker number considerably and lightened the computational burden during the core accession identification.

The population size of the core collection was set at around 384 to aid future DNA assays and analysis that are based on 96 well PCR plates. This population size was also deemed large enough for several genome-wide association studies (GWAS) [21–24], and still manageable for planting and performing phenotypic evaluations in the field. By setting the target population size to 1.9% of the population, the Core Hunter package selected 382 soybean accessions, which originated from at least 33 countries (table 1). Additional analysis to generate core collection that is 20% of the size of the original population was also performed to observe and compare the characteristics of different population sizes selected by the software package.

3.2. Phylogeny and genetic admixture in selected core accessions

Phylogenetic analysis was performed to visualize and check the genetic distances between each selected accession. Clusters of accessions, which indicate groupings between genetically similar accessions, are still visible in the core collection with 4000 accessions and a population of randomly selected 382 accessions, but harder to detect in the core population with 382 individuals (figure 1). This could mean that the Core Hunter had successfully selected individuals that are approximately equidistant to each other and caused phylogenetic analysis that relies on genetic dissimilarity to place the individuals on tree branches that are more or less similar in length and distances.

Genetic admixture analysis was also performed on both the selected core population and the random population. The result suggested a high likelihood that there are 10 different ancestral subpopulations in the collection, and figure 1 shows the proportion of genetic admixtures from those color-coded subpopulations in each individual. It is apparent that one of the side effects of maximum genetic distance-based selection is that it favored admixed individuals as there are very few individuals whose genetic ancestry is purely composed by one subpopulation (figure 1C). This contrasts with the admixture pattern observed in the randomly selected population, where many individuals had nearly pure genotypes from one subpopulation (figure 1B).
| Accession (Country) | Accession (Country) | Accession (Country) | Accession (Country) |
|--------------------|--------------------|--------------------|--------------------|
| FC19976-2 (Japan)  | PI437848B (China)  | P1567333B (China)  | P1603617 (China)   |
| FC19979-6 (Japan)  | PI437902C (China)  | P1567334 (China)   | P1603662A (China)  |
| FC31676 (Unknown)  | PI438019B (China)  | P1567354 (China)   | P1603698F (China)  |
| FC3659 (China)     | PI438065 (China)   | P1567366A (China)  | P1603758D (China)  |
| PI103088 (China)   | PI438112B (China)  | P1567367 (China)   | P1603764B (China)  |
| PI103415 (China)   | PI438257B (China)  | P1567379A (China)  | P16039115C (N_Korea) |
| PI123439 (Myanmar) | PI438282A (Japan)  | P1567382A (China)  | P16039115E (N_Korea) |
| PI155682 (El_Salvador) | PI438292 (Japan) | P1567384 (China) | P1605758D (Vietnam) |
| PI157401 (S_Korea) | PI438314 (Algeria) | P1567392 (China)   | P16058878B (Vietnam) |
| PI157432 (S_Korea) | PI438330A (France) | P1567404E (China)  | P1612609 (Australia) |
| PI157476 (S_Korea) | PI438332 (France)  | P1567406B (China)  | P1612610 (N_Korea) |
| PI157487B (S_Korea) | PI438335 (Algeria) | P1567410A (China)  | P1612734 (China)   |
| PI159322 (S_Africa) | PI438509A (USA)   | P1567410B (China)  | P1612744 (China)   |
| PI170896 (S_Africa) | PI438509B (USA)   | P1567432A (China)  | P1612754 (China)   |
| PI179826 (China)   | PI445810A (Germany) | P1567473A (China) | P1612759A (China)  |
| PI189923 (France)  | PI445810C (Germany) | P1567507B (China) | P1615473 (Vietnam) |
| PI198067 (Sweden)  | PI458264 (S_Korea) | P1567510B (China)  | P162202-2 (China) |
| PI200490 (Japan)   | PI459025C (China)  | P1567515 (China)   | P162204 (China)    |
| PI200520 (Japan)   | PI462312 (India)   | P1567540A (China)  | P1628945 (Brazil)  |
| PI209332 (Japan)   | PI467322B (China)  | P1567543C (China)  | P1631437 (USA)     |
| PI209334 (Japan)   | PI467347 (China)   | P1567545 (China)   | P1632650 (Vietnam) |
| PI209839B (Nepal)  | PI468911 (China)   | P1567552 (China)   | P1632943B (Vietnam) |
| PI219698 (Pakistan) | PI475822C (China) | P1567593B (China)  | P163945 (China)    |
| PI227320 (Japan)   | PI47605A (China)   | P1567622 (China)   | P1639528A (Russia) |
| PI253656B (China)  | PI490765 (China)   | P1567635 (China)   | P1639530 (Unknown) |
| PI257429 (Germany) | PI497964B (India)  | P1567776 (China)   | P1639550E (Moldova) |
| PI274454 (Japan)   | PI497966 (India)   | P1572265D (Georgia) | P1639559A (Ukraine) |
| PI279088 (Unknown) | PI504495 (Taiwan)  | P1572297 (USA)     | P1639561 (Philippines) |
| PI281891B (Indonesia) | PI506507 (Japan) | P1574478A (China)  | P1639569 (Colombia) |
| PI283328 (Taiwan)  | PI506628 (Japan)   | P1578058 (USA)     | P1639602 (Unknown) |
| PI307898 (India)   | PI506639 (Japan)   | P1578399 (China)   | P1639621 (Unknown) |
| PI323574 (India)   | PI506737 (Japan)   | P1578474 (China)   | P1653874B (Vietnam) |
| PI330635 (S_Africa) | PI506762B (Japan)  | P1584527 (USA)     | P1656647 (USA)     |
| PI342434 (Japan)   | PI506825 (Japan)   | P1587550A (China)  | P166879 (China)    |
| PI346306 (India)   | PI506908 (Japan)   | P1587555A (China)  | P171564 (China)    |
| PI346308 (India)   | PI507017 (Japan)   | P1587606C (China)  | P179691-4 (China) |
| PI346309 (India)   | PI507293A (Japan)  | P1587612B (China)  | P179732-3 (China)  |
| PI360851 (Japan)   | PI507341 (Japan)   | P1587662A (China)  | P179825 (China)    |
| PI361062A (Romania) | PI507417 (Japan)  | P1587668C (China)  | P180466-1 (Japan)  |
| PI361080 (Russia)  | PI507472 (Japan)   | P1587828 (China)   | P180473 (Japan)    |
| PI378658 (Ukraine) | PI507476 (Japan)   | P1587893 (China)   | P180837 (Japan)    |
| PI378682B (Japan)  | PI507480 (Japan)   | P1587911A (China)  | P180845-2 (Japan)  |
| PI378693A (Japan)  | PI507486 (Japan)   | P1587996C (China)  | P181029 (Japan)    |
| PI378696B (Unknown) | PI507501 (Japan)  | P1588006A (China)  | P181037-5 (Japan) |
| PI391577 (China)   | PI507521 (Japan)   | P1588007A (China)  | P181041 (Japan)    |
| PI398368 (S_Korea) | PI507567 (Japan)   | P1588039 (China)   | P181042-2 (Japan) |
| PI398481 (S_Korea) | PI507681B (Uzbekistan) | P1592938 (China) | P181785 (Japan)    |
| PI398587 (S_Korea) | PI507686C (Moldova) | P1592939 (China) | P182278 (S_Korea) |
| PI398875 (S_Korea) | PI507718B (N_Korea) | P1592947 (China)  | P182302 (S_Korea)  |
| PI407191 (Unknown) | PI509108 (S_Korea) | P1592951 (China) | P182312 (S_Korea) |
| PI407658B (China)  | PI509113 (China)   | P1593999B (S_Korea) | P183881 (N_Korea) |
| PI407771 (S_Korea) | PI512322A (Georgia)| P1594449 (China)  | P183925 (Japan)    |
| PI407937-1 (S_Korea) | PI538377 (China)  | P1594471E (China) | P184594 (S_Korea) |
Since soybean is largely self-pollinating [25], genetically admixed individuals are more likely caused by artificial crossing than random mating. Thus, the Core Hunter software might inadvertently favor improved cultivars developed by plant breeding than landraces, which usually arose from a continual selection of natural variants and not from deliberate hybridization of different soybean accessions [26]. This may be more beneficial for breeding purposes, as improved cultivars are typically bred to obtain superior traits. Thus, a population of improved cultivars should theoretically have more agriculturally desirable alleles in them. On the other hand, landraces were likely selected based on their superior adaptation to a certain environment, which could also be beneficial for some breeding programs or genetic studies, so their absence in the core population can be concerning.

The admixture pattern may also be useful during the parental selection stage in plant breeding or genetic studies. For example, to generate a genetically diverse progeny population from some
Indonesian cultivars like Grobogan, which has mostly red subpopulation ancestry, it should be crossed with accessions carrying different colored subpopulation ancestry (figure 1.C). Further diversity checks can be performed by examining the full allelic and phenotypic differences among the candidate parental lines. For breeding purposes, admixed accessions that are specifically crossed and bred for cultivation also tend to have less linkage drag since non-desirable traits are usually selected against during the breeding process [27].

![Figure 1. Phylogenetic tree and admixture pattern of various populations: A. Core collection set at 20% of the total population and some Indonesian cultivars (●B3293, ●Tambora, ●Grobogan, ●Anjasmoro); B. Randomly selected 382 accessions; C. Core collection with 382 accessions plus some Indonesian cultivars. Note that the same color in each admixture chart does not refer to the same subpopulation.](image)

3.3. Principal components, allele frequencies, and phenotype distribution
Principal component analysis (PCA) was also carried out to see the distribution pattern of selected accessions in relation to the rest of the accessions in the population. In the eigenvalue plot of the two largest components of genetic variations in the population, it can be seen that the Core Hunter selection algorithm removed accessions with extremely high or low eigenvalues (figure 2), since accessions located in the edges of the scatterplot were not selected. This occurred even when the selection intensity was fairly relaxed at only 20% of the population and became even more pronounced when only 382 accessions were selected from 20,095 total accessions. A similar pattern was observed in a wheat core collection development using different software [28]. Further investigation is required to examine the effects of those extreme eigenvalues on accession and whether eliminating such accessions will be detrimental for future genetic studies or breeding programs.

In terms of allele frequency preservation, random selection and Core Hunter selection did not significantly alter the distribution of allele frequencies in the population (figure 3), although it appeared that some alleles with extremely low frequencies were more common in the core population than the random population. A closer examination of the raw allele frequency table in the core population confirmed that all alleles were retained in the population and no rare alleles disappeared even after 98.1% of the population was removed to form the core collection. Thus the Core Hunter package had successfully preserved all genotyped alleles in the core population.

The distribution of phenotypic values appeared to undergo some shifts in the core collection. We examined the distribution of four traits that had been evaluated by the USDA, namely branch number, seed weight, yield, and maturity group (figure 4). There were no obvious changes in the distribution of yield in the original, core, and random population. However, the median value and the overall range of seed weight were lower in the core population compared to the original and randomly chosen population. Large seed weight is typically found in elite cultivars [29], so one possible explanation of the reduction in median seed weight is the elimination of many elite cultivars in the core collection due to their close genetic relationship to each other. Nevertheless, eliminating accessions from the original population with extremely high seed weight is worrying as this may mean that the alleles causing that trait are absent in the core collection. Another possibility is that the alleles still exist in the core collection but no longer exist together in a single accession, which means that multiple crosses will be needed to reassemble the allelic combinations that produce extremely high seed weight.
In contrast, while the median maturity group is the same in the core, original, and random population, the interquartile range of the core population shifted toward the higher numbered maturity group.
Soybean maturity group in North America is numbered from 000 to IX, with the lowest number most suited to high latitude regions like Canada while maturity group IX is most suited to Southern Florida [30]. The shift toward higher numbered maturity groups means that the core collection has an increased proportion of accessions adapted to lower latitude regions (especially group V and VI). It is possible that accessions from maturity group 000 to IV, which make up 64% of the original collection, had lower genetic diversity between them so that most of them were eliminated by Core Hunter in the core collection. For the branching trait, the core, random, and original population are dominated by non-branching accessions, although outliers with up to 4 branches exist in both core and random populations. Accessions with branches are actually quite rare since they only exist in <14% of the population. The fact that those branched accessions are still included in the core collection means that those rare accessions have sufficient genetic diversity between them, resulting in their inclusion in the core collection.

4. Conclusions
A soybean core population has been successfully identified and it can be used by genebanks or soybean breeding programs as a starting point in sampling the genetic diversity of worldwide soybean collection. Our data shows that while the Core Hunter 3 package successfully selected accessions that are approximately genetically equidistant to each other and avoid over-representation of dominant sub-populations in the original collection, it might also inadvertently eliminate accessions with extreme eigenvalues for the main principal components, accessions with pure ancestral subpopulation genetic background, and accessions with very high seed weight. Since the number of accessions is still relatively small, it is feasible to add those eliminated accessions to the core collection if it is deemed necessary. The genetic data can also be used as additional screening tools to select parental lines that produce progenies with high genetic diversity.

Acknowledgements
The authors thank the Soybase Database and USDA GRIN soybean germplasm collection for their generosity in opening their data and collection. For this manuscript, both authors contributed equally as the main contributors.

References
[1] Hasan N, Suryani E and Hendrawan R 2015 Procedia Computer Science 72 605–12
[2] FAOSTAT Retrieved from http://www.fao.org/faostat/en/#data/QC
[3] Bertan I, Carvalho F I F De and Oliveira A C De 2005 J Crop Sci Biotech 10 211–22
[4] Haupt M and Schmid K 2020 Plant. Cell Environ 43 1421–36
[5] van Hintum T J L, Brown A H D and Spillane C 2000 Core Collections of Plant Genetic Resources (Rome: IPGRI)
[6] Kaga A, Shimizu T, Watanabe S, Tsubokura Y, Katayose Y, Harada K, Vaughan D A and Tomooka N 2012 Breed. Sci. 61 566–92
[7] SoySNP50K iSelect BeadChip - SoyBase Retrieved from https://soybase.org/snps/
[8] Descriptors - GRIN-Global Web v 1.10.6.1 Retrieved from https://training.ars-grin.gov/gringlobal/descriptors.aspx
[9] Bochner B 2003 Nat. Rev. Genet. 4 309–14
[10] Accessions - GRIN-Global Web v 1.10.6.2 Retrieved from https://npgsweb.ars-grin.gov/gringlobal/search.aspx
[11] Zheng X, Levine D, Shen J, Gogarten S M, Laurie C and Weir B S 2012 Bioinformatics 28 3326–8
[12] De Beukelaer H, Davenport G F and Fack V 2018 BMC Bioinformatics 19 203
[13] Satyawan D, Rijzaani H and Tasma I M 2014 Plant Genet. Resour. Characterisation Util. 12 S1
[14] Pusat Genom Pertanian Indonesia Retrieved from http://genom.litbang.pertanian.go.id
[15] Bradbury P J, Zhang Z, Kroon D E, Casstevens T M, Ramdoss Y and Buckler E S 2007 *Bioinforma. Appl.* 23 2633–5
[16] Letunic I and Bork P 2019 *Nucleic Acids Res.* 47 W256–9
[17] Wickham H 2016 *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York)
[18] Alexander D H and Lange K 2011 *BMC Bioinformatics* 12 246
[19] Danecek P, Auton A, Abecasis G, Albers C A, Banks E, Depristo M A, Handsaker R E, Lunter G, Marth G T, Sherry S T, Mcvean G, Durbin R and Project G 2011 *Bioinforma. Appl. NOTE* 27 2156–8
[20] Malomane D K, Reimer C, Weigend S, Weigend A, Sharifi A R and Simianer H 2018 *BMC Genomics* 19 22
[21] Hwang E Y, Song Q, Jia G, Specht J E, Hyten D L, Costa J and Cregan P B 2014 *BMC Genomics* 15 1
[22] Zatybekov A, Abugalieva S, Didorenko S, Gerasimova Y, Sidorik I, Anuarbek S and Turuspekov Y 2017 *BMC Plant Biol.* 17 179
[23] Dhanapal A P, Ray J D, Singh S K, Hoyos-Villegas V, Smith J R, Purcell L C, Andy King C, Cregan P B, Song Q and Fritschi F B 2014 *Theor. Appl. Genet.* 128 73–91
[24] Zeng A, Chen P, Korth K, Hancock F, Pereira A, Brye K, Wu C and Shi A 2017 *Mol. Breed.* 37 1–14
[25] Yan H, Zhang J Y, Zhang C B, Peng B, Zhang W L, Wang P N, Ding X Y, Liu B H, Feng X Z and Zhao L M 2019 *J. Integr. Agric.* 18(9) 1971–9
[26] Casañas F, Simó J, Casals J and Prohens J 2017 *Front. Plant Sci.* 08 145
[27] Kumar A, Sandhu N, Venkateshwarlu C, Priyadarshi R, Yadav S, Majumder R R and Singh V K 2020 *Sci. Rep.* 10 13073
[28] Jeong S, Kim J Y, Jeong S C, Kang S T, Moon J K and Kim N 2017 *PLoS One* 12 e0181420
[29] Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, Fang C, Shen Y, Liu T, Li C, Li Q, Wu M, Wang M, Wu Y, Dong Y, Wan W, Wang X, Ding Z, Gao Y, Xiang H, Zhu B, Lee S H, Wang W and Tian Z 2015 *Nat. Biotechnol.* 33 408–14
[30] MU Gene Zoos and Crop Gardens - Soybean Gene Zoo Retrieved from http://genezoo.missouri.edu/soybean.cfm