

Identification of subpopulation in Russian soybean germplasm collection using advanced genetic methods

Habib Y.^{1*}, Zamalutdinov A.¹, Boldyrev S.¹, Schegolkov A.², Ben C.¹, Gentzbittel L.¹

¹ Skolkovo Institute of Science and Technology, Moscow, Russia

² SOKO, Krasnodar, Russia

* Yawar.Habib@skoltech.ru

Key words: soybean germplasm; admixture; Discriminant Analysis of Principal Components (DAPC)

Motivation and Aim: Globally recognized as a crucial legume, soybean (*Glycine max* (L.) Merr.) is esteemed for its high protein and oil content, serving as a key component in diverse food products. The foundation of successful breeding programs requires a comprehensive understanding of the genetic variation and population dynamics existing within contemporary germplasm repositories. Assessing genetic diversity and identifying the number of subpopulations is crucial for enhancing breeding strategies and preserving genetic resources. This study aims to employ sophisticated techniques such as Admixture, Discriminant Analysis of Principal Components (DAPC), integrated with K-means clustering, to identify various subgroups within Russian soybean collections and elucidate their genetic diversity and population structure.

Methods and Algorithms: SNP calling for GBS data was performed using “ref_map.pl” from the Stacks software package with default parameters, while for WGS data, GATK 4.2.6.1 Haplotype Caller was used [1]. Quality control was conducted with PLINK v1.07 [2], using thresholds for missing genotype and individual rates at 0.6 and 0.01 respectively, MAF at 0.05. To remove SNPs in linkage disequilibrium, haplotype blocks were constructed using the D’ method [3] with PLINK v1.07, defining strong LD by a 95 % confidence bound on D’ > 0.98 and < 0.9 for recombination evidence, with a 200kb window size. Admixture patterns were analyzed using the software ADMIXTURE v1.3 [4] with 50 bootstrap replicates, and cross-validation was done to determine the optimal K value. DAPC, implemented in R’s adegenet 2.0 package, was used to identify genetic clusters via K-means clustering and Bayesian Information Criterion (BIC), transforming data with PCA and performing Discriminant Analysis on retained PCs using cross-validation.

Results: In this study, a set of 59,830 SNPs, genotyped from 344 soybean accessions (50 by WGS and 294 by GBS), was utilized. Following quality control procedures, 25,388 SNPs were retained across 323 soybean genotypes, achieving a genotyping rate of 93.58 % after excluding SNPs and individuals that did not meet the criteria for missingness and minor allele frequency threshold. The number of haplotype blocks computed were 3,225 and two SNPs from each block were selected based on higher MAF and the maximum distance among each other. The lowest value of cross validation error was associated with K = 15 showing in Fig. 1(a). Fig. 1(b) displays the proportion of ancestral heritage for these 15 groups as determined by admixture analysis. Many individuals display homogenous ancestry, with bars of a single color indicating over 90 % genetic material from one ancestral group. Clusters of individuals from 250 to 268 and 288 to 307 exemplify this, having similar parental lines. In total, 93 individuals showed homogenous ancestry. Conversely, some individuals have highly mixed

ancestry. For instance, individual 208's highest single ancestry proportion is only 17.5 %, while individual 203's is 18.7 %. Others, like individuals 221, 205, 135, and 202, have single ancestry proportions ranging from 18.9 % to 21.5 %, reflecting diverse genetic contributions.

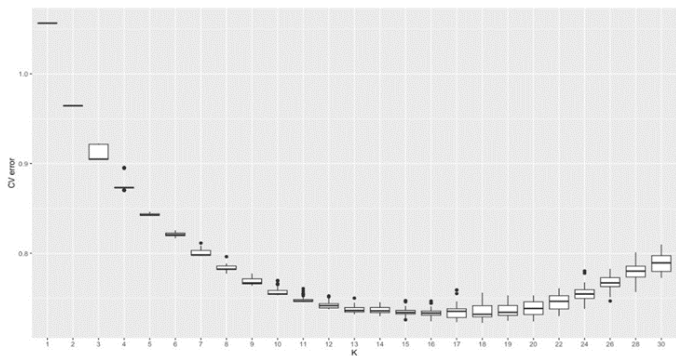


Fig. 1. (a) Cross-validation error for each value of K, indicating the optimal number of ancestral populations

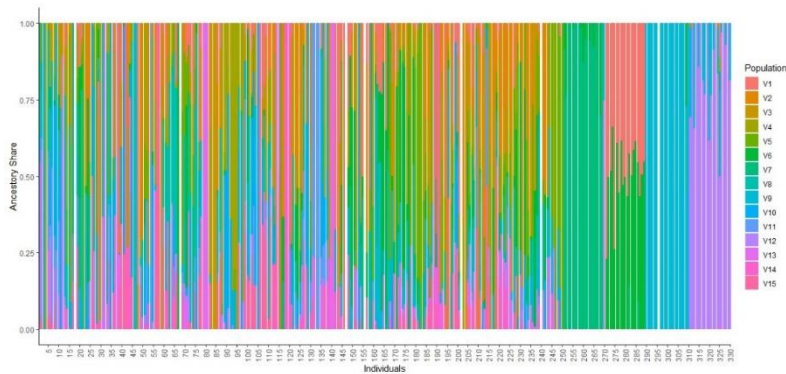


Fig. 1. (b) Ancestry proportions for each individual in the study, as inferred from the Admixture analysis with K=15 (15 ancestral populations)

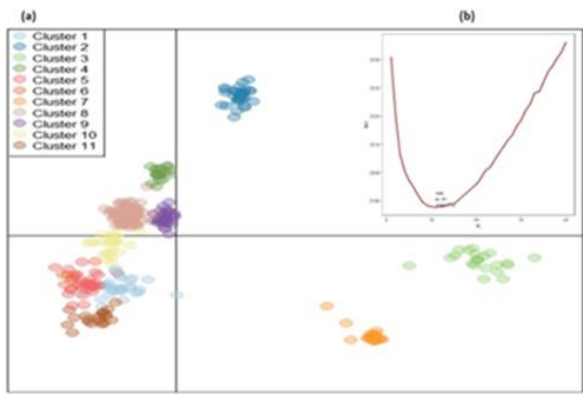


Fig. 2. (a) Shows Discriminant Analysis of Principal components (DAPC) scatter plots with PC1 and PC2 for identification of 11 unique soybean groups with embedded fig. 2 (b) which shows K means clustering visualization based on Bayesian Information Criteria (BIC)

The valley value plot in Fig 2(b) for K means clustering shows an elbow at K = 11, which will be considered the best number of clusters for this soybean germplasm collection. The scatter plot in Fig. 1(a) from the DAPC shows these 11 distinct groups, with 30 principal components and 6 discriminant functions used to describe how these clusters are related. Clusters 2 and 7 are densely packed with genetically similar lines from specific parent crosses, as seen in the admixture analysis. This dense clustering confirms the quality of the genetic data and the accuracy of the clustering. In contrast, the less dense clusters show more genetic diversity, which is good for the genetic diversity of the germplasm. These diverse clusters are important because they might possess different traits valuable for different breeding targets. The number of accessions in each subpopulation varied significantly. Subpopulation VIII had the highest number with 70 accessions, while Subpopulation III and XI had the lowest numbers, with 5 and 20 accessions, respectively. The distribution aids in elucidating the genetic architecture of the soybean germplasm, serving as a guide for breeding and conservation initiatives. *Conclusion:* This study mapped out the genetic makeup of Russian soybean germplasm, revealing 11 subpopulations using DAPC and 15 ancestral populations through admixture analysis. Despite the differences in the number of clusters, there was notable consistency in individual clustering, particularly in densely populated clusters like Clusters 2 and 7, which showed high genetic similarity. The genetic diversity observed among subpopulations is crucial for the preservation of germplasm and for guiding breeding programs focused on developing new cultivars. These findings underscore the importance of integrating different analytical approaches to obtain comprehensive insights into genetic diversity and population structure, with important implications for future studies on association mapping and crop improvement.

References

1. O'Connor B.D., Van der Auwera G. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media, 2020
2. Purcell S., Neale B., Todd-Brown K. et al. PLINK: A Toolset for Whole Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet.* 2007;81(3):559-575
3. Gabriel S.B., Schaffner S.F., Nguyen H. et al. The structure of haplotype blocks in the human genome. *Science.* 2002;296(5576):2225-2229
4. Alexander D.H., Novembre J., Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009;19(9):1655-1664