ABSTRACT

of the dissertation work of Doszhanova Botakoz
titled: **"Genome-wide association study of agronomic traits in soybean (*Glycine max* (L.) Merr.) for practical breeding"** submitted for the academic degree "Doctor of Philosophy" (Ph.D.), education program 8D05105 – "Biotechnology"

The relevance of research. Soybean (*Glycine max* (L.) Merr.) is one of the most important legume crops worldwide, serving as a major source of protein and oil for food, feed, and industrial applications. In Kazakhstan, due to the economic and practical value of this crop, new cultivation areas are being developed each year. The identification of genetic factors associated with important agronomic traits in opens up new opportunities for enhanced breeding efficiency and the development of high-productivity varieties. Modern biotechnological approaches using molecular genetic markers can accelerate the development of this area of breeding. Global experience in applying the results of such studies significantly speeds up the development of new crop varieties with desirable traits, while saving financial and time resources. In this regard, the study of genetic resources of soybean from Kazakhstan and foreign varieties and lines using advanced genetic methods is of great importance both for fundamental research and for practical applications in molecular soybean breeding.

Research objects. The object of the study was a collection consisting of 252 soybean (*Glycine max* (L.) Merr.) varieties and lines of diverse origin, including 31 samples of Kazakhstan breeding programs.

Research subjects. The subject of the study was a genome-wide association study (GWAS) of a soybean collection grown under the conditions of Almaty region, aimed to identify marker-trait associations (MTAs) and KASP (*Kompetitive Allele-Specific PCR*) DNA markers associated with traits determining soybean productivity and quality.

The goal of research. The goal of the research is identifying marker-trait associations for traits determining productivity and quality in soybean (*Glycine max* (L.) Merr.) based on the study of a global collection of diverse origin using the method of association mapping.

Tasks of research:

1. To study the phenotypic variability of the soybean collection based on the analysis of a set of traits determining plant productivity under the growing conditions of Almaty region of Kazakhstan;
2. To select promising soybean varieties and lines based on the studied agronomically valuable traits to enhance soybean breeding;
3. To study the genetic diversity and characterize the genetic structure of the soybean collection, including accessions from Kazakhstan, Eastern and Western Europe, East Asia, and North America, using SNP (*single nucleotide polymorphism*) markers based on whole genome resequencing data (WGRS);
4. To identify MTAs determining soybean productivity and seed quality using the GWAS approach.

5. To analyze haplotypes of significant associations and develop new DNA markers for KASP associated with agronomically valuable traits in soybean.

Research methods. The present study employed field experiments, as well as biochemical, molecular-genetic, and statistical methods of investigation. Field experiments were conducted at the Kazakh Research Institute of Agriculture and Plant Growing (KRIAPG). Grain quality was assessed in accordance with GOST 32749-2014 in the Laboratory of Biotechnology, Biochemistry, and Grain Quality at KRIAPG using the NIRS DS2500 grain analyzer (FOSS, Hillerød, Denmark). Molecular-genetic research, in particular, the procedure of DNA extraction from soybean seedlings, quality assessment, and visualization, was performed in the Molecular Genetics Laboratory at the Institute of Plant Biology and Biotechnology (IPBB). Resequencing was carried out using the Illumina HiSeq X Ten system at the Department of the School of Natural Sciences, Guangzhou University, Guangzhou, China. Biometrical statistical methods included analysis of variance (ANOVA), Pearson correlation analysis, Student's *t*-test, and visualization of results using RStudio packages. Statistical methods also included regression analysis models, such as the multi-locus mixed model (MLMM) implemented using the GAPIT package in the R platform, neighbor joining (NJ) clustering, and the Markov Chain Monte Carlo (MCMC) method using the STRUCTURE software. During the course of study, close collaboration was established with partners from China and Japan. In 2019, the PhD student completed a research internship on the topic of the study at the Hokkaido Research Center under the supervision of a foreign academic advisor.

The scientific novelty of the research. Using SNP markers, a comparison of the genetic profiles of Kazakhstan soybean varieties and lines with foreign accessions was conducted, and the population structure of the global soybean collection was analyzed. The scientific novelty of the study is the identification of 46 stable marker-trait associations (MTAs) related to agronomically important traits in soybean, including 25 presumably novel MTAs, based on the application of modern molecular genetic methods. Among the stable MTAs, six pleiotropic associations were identified, each associated with multiple traits. The highest number of MTAs was detected for yield components (20), followed by traits related to protein and oil content (18), seed size (5), and pod shattering resistance (3).

The obtained results highlight the importance of genetic and breeding research under specific growing conditions, where the interaction between genotype (G) and environment (E) must be taken into account. The identified associations have the potential for further investigation and detailed identification of candidate genes controlling agronomically important traits in soybean. These findings provide a theoretical foundation for future breeding strategies aimed at developing improved soybean varieties.

The study's practical significance. Based on the identified significant and stable associations, effective haplotypes associated with important agronomic traits were revealed. These associations serve as a foundation for the development of new informative KASP markers—valuable tools for soybean breeding aimed at improving seed quality and overall crop productivity. A patent titled "Method for the identification of valuable soybean breeding lines (*Glycine max* (L.) Merr.) using a

set of KASP markers" (No. 9289) was developed and registered, enabling the efficient selection of valuable lines for further breeding. The obtained results can be implemented in breeding programs to improve key desirable traits of soybean in Kazakhstan.

The theoretical significance. One of the main hypotheses of this study was to explore the possibility of identifying novel and previously reported MTAs for agronomically important traits in soybean under the growing conditions of a global collection in southeastern Kazakhstan. The identified MTAs have potential for further investigation and the possible identification of candidate genes. The obtained results can be effectively utilized in soybean genetic and breeding research, taking into account the specific interactions between genotype and environment, and may serve as a theoretical foundation for future breeding strategies aimed at developing improved soybean varieties.

The main provisions of the dissertation submitted for defense:

1. The studied collection, consisting of 252 soybean varieties and lines from Kazakhstan, Eastern and Western Europe, East Asia, and North America, demonstrates a high level of phenotypic diversity under the environmental conditions of the Almaty region. Based on the combination of ten agronomically valuable traits, the collection serves as an effective genetic resource for conducting genome-wide association studies (GWAS) aimed at identifying novel genetic factors influencing the variability of important agronomic traits in soybean.

2. As a result of the phenotypic evaluation of the soybean collection for 10 agronomically valuable traits, 9 domestic and 42 foreign genotypes were identified with high productivity and seed quality indicators, surpassing those of the standard variety in the Almaty region. These genotypes represent a valuable resource for further use in breeding programs aimed at developing improved soybean varieties.

3. Based on whole-genome resequencing data, the genetic structure of the world soybean collection was divided into three main subpopulations according to their geographical origin and characterized by a high level of genetic diversity. The identified genetic structure and degree of population diversity meet the key requirements for conducting genome-wide association studies (GWAS).

4. A total of 106 significant trait associations related to soybean productivity and seed quality were identified through GWAS, including 46 stable associations. These findings provide important insights for uncovering genetic factors controlling complex quantitative traits in this crop species. Among the 46 stable associations, GWAS revealed 25 presumably novel marker–trait associations (MTAs) for the studied agronomically valuable traits. Additionally, 21 MTAs were previously identified in other studies, confirming the reliability of provided results.

5. Based on the significant associations, effective haplotypes were identified for traits determining soybean productivity and seed quality. These haplotypes can be applied at early stages of breeding programs aimed at developing high-yielding, high-quality domestic soybean varieties. Additionally, the most significant associations related to multiple yield-determining traits were converted into KASP markers — *ipbb_gm_141*, *ipbb_sd_16,* and *ipbb_sd_17* — which demonstrated high efficiency for use in breeding programs.

Relation to the plan of main scientific works. The dissertation research was carried out within the framework of grant-funded projects of the Ministry of Science and Higher Education of the Republic of Kazakhstan: No. AR05131592 “Genome-wide association study of soybean resistance to fungal diseases in Kazakhstan,” AP13068118 “Identification of genes associated with agronomically important traits in soybean (*Glycine max* (L.) Merr.) using whole-genome resequencing of a global collection” (2022–2024), and the scientific and technical program BR24992903 “Practical implementation of modern molecular-genetic, physiological-biochemical, biotechnological methods, and digital phenotyping in the breeding of economically important agricultural crops” (2024–2026).

The personal contribution of a Ph.D. student to the preparation of each publication. It includes data collection on the research topic, the implementation of theoretical and experimental studies, as well as the preparation of articles for publication.

Approbation of research work. The results of the dissertation research were presented and published at international scientific and practical conferences, including: the International Scientific Conference of Students and Young Scientists "Farabi Alemi" (Almaty, Kazakhstan, 2019, 2020, 2025), the International Scientific Conference of Young Scientists dedicated to the 40th anniversary of the Institute of Molecular Biology and Biochemistry (Almaty, Kazakhstan, 2023), the International Conference on Plant Biology and Biotechnology (Almaty, Kazakhstan, 2024), and the 4th Young Scientists School “Plant Genetics, Genomics, Bioinformatics, and Biotechnology” PlantGen School 2024 (Novosibirsk, Russia, 2024). The main results of the dissertation research were presented annually at the scientific seminars of the department of biotechnology, biology and biotechnology faculty, Al-Farabi Kazakh National University, and at scientific seminars of the laboratory of molecular genetics IPBB. The results of the dissertation research were partially included in the research reports for projects AP05131592, AP13068118 and BR24992903.

Publications. The main content of the dissertation has been published in 12 printed works, including 5 articles in international peer-reviewed journals in SCOPUS and/or Web of Science, 1 patent (utility model), 5 abstracts in the international conferences, and 1 poster presented at an international conference abroad.

Conclusions. In the present study, a comprehensive analysis was conducted to identify marker–trait associations related to important agronomic traits in soybean (*Glycine max* (L.) Merr.), including yield components, seed quality characteristics, and morphological features. The main findings of the study are summarized as follows:

1) The study revealed that the analyzed collection, comprising 252 soybean cultivars and lines from Kazakhstan, Eastern and Western Europe, East Asia, and North America, shows a high level of phenotypic diversity under the environmental conditions of the Almaty region. Based on the combined analysis of ten agronomic

traits, the collection represents an effective genetic resource for conducting genome-wide association studies (GWAS) aimed at identifying novel genetic factors influencing trait variability in soybean. ANOVA results indicated that among all studied traits, the genotype (G) had the most significant effect on seed quality traits ($P < 2\times10^{-16}$), whereas the environment (E) had a stronger impact on yield components ($P < 2\times10^{-16}$). The genotype-by-environment interaction (G × E) was significant for most traits, underscoring the importance of testing cultivars across different agroecological zones ($P < 0.01$). Pearson correlation analysis revealed a significant negative correlation between protein content and oil content, as well as with yield components, which plays a key role in breeding strategies ($P < 0.05$). In contrast, seed sizes as width and thickness showed a positive correlation with oil content and yield components ($P < 0.05$).

2) As a result of the phenotypic evaluation of the soybean collection based on 10 agronomic traits, 9 local (Danaya, Pamyat' YuGK , lines 186/1, 126/1, 173/1, 261/1, 391, 209/1, 350/1) and 42 foreign genotypes were identified that demonstrated superior productivity and seed quality compared to the standard cultivar used in the Almaty region. These genotypes represent a valuable resource for further use in breeding programs aimed at developing improved soybean cultivars.

3) Genotyping of the 252 soybean accessions using whole-genome resequencing data resulted in the identification of 4,923,660 SNPs, of which 44,385 high-quality SNP markers were selected for further analysis of genetic diversity and high-resolution genome-wide association studies (GWAS). The population structure analysis revealed that the collection primarily consists of three major subpopulations, differentiated by geographic origin. Notably, the 31 accessions from Kazakhstan were distributed across all three subpopulations, indicating the extensive use of germplasm from various regions of the world in soybean breeding programs in Kazakhstan.

4) As a result of the GWAS, 106 significant marker–trait associations (MTAs) were identified, including 46 stable associations that exceeded the statistical significance threshold. Among these 46 stable MTAs, 25 were considered potentially novel, as they have not been previously reported in the literature. The remaining 21 MTAs had been identified in earlier studies, supporting the reliability of the findings in this research. Among the 25 potentially novel MTAs, 15 were associated with yield components, 5 with seed quality traits, 3 with seed size, and 2 with pod shattering resistance.

5) Based on the significant associations, effective haplotypes were identified for traits related to soybean productivity and quality ($P < 1\times10^{-4}$). These haplotypes can be applied at early stages of breeding programs aimed at developing high-productivity and high-quality local soybean cultivars. In addition, the most significant associations related to multiple traits, particularly yield components, were converted into KASP markers - *ipbb_gm_141, ipbb_sd_16* и *ipbb_sd_17*, which demonstrated high efficiency for application in breeding programs ($P < 0.05$).