Arthur Bernardeli
March 24, 2025
Exam 1 (Midterm) - AGRO 932

Genomic Patterns in Distinct Soybean Breeding Pools: Selection and FST Dynamics

Soybean (*Glycine max* (L.) Merr.) is a globally important crop, cultivated for its high-value seeds rich in both oil and protein. It serves as a critical component of food, feed, and industrial markets, with significant production hubs across the Americas and Asia (US Department of Agriculture Foreign Agricultural Service, 2017). In the United States, soybean is a cornerstone of agricultural output and economic stability, occupying millions of acres annually. Its dual utility, approximately 40% protein and 20% oil in seed composition makes it especially valuable for both human consumption and livestock feed. Given its essential role in global nutrition and supply chains, improving soybean traits remains a primary objective for plant breeding programs worldwide.

Among the most economically and nutritionally significant traits in soybeans are seed yield and seed protein content. However, these traits are negatively correlated, presenting a substantial challenge for breeders. Genetic gains in seed yield have often been accompanied by declines in protein concentration, and efforts to increase protein content frequently come at the expense of yield potential. This antagonistic relationship is rooted in complex genetic and physiological trade-offs, which are not yet fully understood. As a result, breeding programs have often pursued these traits in parallel but separate pipelines, selecting elite lines for either high yield or high protein. While this strategy enables targeted improvements, it limits opportunities for simultaneous enhancement of both traits and underscores the need to better understand the genetic mechanisms underlying this trade-off.

Despite significant advances in genomic scan studies in soybean seed composition traits, such as in Zhang et al. (2016, 2018), the genetic architecture and selection dynamics that maintain or potentially resolve the yield-protein trade-off remain poorly defined. There is limited insight into how sustained directional selection shapes allele frequencies at trait-associated loci, and how such divergence may manifest across the genome in breeding populations that share a common elite ancestry. Population genomics approaches, such as the fixation index (FST), offer a powerful means of quantifying genetic differentiation and identifying regions of the genome that have responded to trait-specific selection pressures. Moreover, the integration of recombination and inter-group crossing offers the potential to break unfavorable linkage blocks and generate superior recombinants that transcend historical trait limitations.

This study aims to investigate these genomic and breeding dynamics using a simulation-based approach. We simulated elite soybean populations subjected to divergent selection for either high yield or high protein content. These simulations are designed to mimic sequential selection and mating cycles commonly performed by soybean breeding programs, enabling a controlled and realistic examination of selection outcomes over time. The simulated populations initially share a common elite background, from which two selection paths emerge. Each path undergoes repeated intra-population selection, resulting in phenotypic advancement and genomic divergence. Genotypic and phenotypic data are collected over multiple selection cycles, enabling the analysis of genome-wide FST, chromosome-specific patterns of divergence, and changes in within-group diversity and trait distributions.

The first aim of this study is to evaluate the degree of genetic differentiation between the high-yield and high-protein populations following divergent selection. Using genome-wide and chromosome-specific FST estimates, we assess how selection shapes allele frequency distributions in each breeding pool. We hypothesize that loci associated with yield or protein content will exhibit moderate to high differentiation values as selection cycles advance.

The second aim is to determine how repeated cycles of intra-population selection impact both genetic structure and phenotypic means within each group. We hypothesize that selection will reduce within-group diversity by favoring specific haplotypes, thereby increasing genetic homogeneity and driving further divergence at key loci across the genome. This will be reflected in rising phenotypic means and increasing FST values relative to the founder generation.

The third aim explores the consequences of crossing between the divergent breeding pools, followed by selection in the recombinant progeny. This scenario is designed to assess whether recombination can effectively break the negative correlation between yield and protein, allowing for the development of superior lines that combine favorable alleles from both parental groups.

Material and Methods

Population Simulation and Structure

To investigate genomic divergence and selection responses in soybean breeding pools, we simulated one diverse founder population, each consisting of 20 high-yield lines and 20 high-protein lines. These population will then represent breeding pools with distinct selection histories, one selected for high seed yield and the other for high seed protein content. Genotypic data were simulated using 6,000 evenly spaced single-nucleotide polymorphism (SNP) markers, distributed across the 20 soybean chromosomes (~300 markers per chromosome), to mimic the marker density commonly used in breeding programs and genomic studies. Genotype data were imputed leveraging a next-generation sequencing dataset based on the *Glycine max* reference genome Williams 82 (*www.soybase.org*).

The high yield lines (group Y) was designed to reflect an elite germplasm background with a historical focus on improving seed yield. This group had a high base population mean for yield and a moderate mean for protein content, consistent with the documented negative correlation between the two traits. The high protein lines (group P) was simulated from a similar elite background but selected for increased seed protein concentration. Due to the antagonistic relationship between the traits, this group exhibited a reduced yield mean. Both populations shared common genetic ancestry, emulating real-world breeding scenarios where selection diverges from a shared elite base.

Simulation Platform and Trait Architecture

Simulations were conducted using the AlphaSimR (Gaynor et al., 2021) package in R (R Core Team, 2021), a widely used platform for modeling complex trait inheritance, recombination, and selection in plant breeding, according to specific simulation parameters. The simulated traits included seed yield and seed protein content, with a genetic correlation of -0.45 imposed between them to reflect the biological constraint observed in empirical breeding data. Heritability values were set at 0.45 for yield and 0.70 for protein.

Trait values were assigned at the founder level, generating variation within the founder population for groups Y and P. This variation allowed for the initial assessment of divergence and provided a foundation for within-population selection in subsequent cycles.

Experimental Procedure

Step 1: Founder Population Simulation

The initial step involved simulating one founder population of 40 individuals, each with assigned genotypes and trait values according to the parameters described earlier. These populations served as the baseline for evaluating genomic differentiation and trait performance.

Step 2: Intra-Group Recurrent Selection Cycles

Each simulated population underwent five generations of trait-specific selection. In each cycle, the top 10% of individuals were selected based on phenotypic values for their respective target trait, yield for the Y group and protein content for the P group.

Selected individuals were intermated within their group to produce progeny for the next generation. Recombination and segregation were simulated during mating, and new phenotypic values were assigned based on inherited genotypes and environmental effects. This process was repeated for five cycles, allowing the accumulation of selection responses and genetic changes across generations.

Step 3: FST, PCA, and phenotypic analysis

Following the founder simulation, genome-wide and chromosome-specific FST values were calculated between the Y and P groups using SNP data. This allowed for the quantification of genetic differentiation attributable to divergent selection histories. In parallel, principal component analysis (PCA) was performed to visualize the overall genetic structure and clustering of individuals by selection group. FST and PCA analyses were repeated in each cycle to evaluate changes in population structure and genetic clustering. Comparisons were made between each generation and its respective founder population, as well as between the high-yield and high-protein populations in cycles 0 and 5. For soybeans, a similar procedure was performed by Yang et al. (2022), Silva et al. (2025), and Andrijanić et al. (2023) to compute the fixation index in distinct soybean pools. Phenotypic means of yield and protein content were recorded for each group at every selection cycle to assess phenotypic gains and trade-offs over time.

Codes

Codes are available in the Supplementary Material section.

Results and Discussion

This study investigated the genomic consequences and phenotypic benefits of selection in two distinct soybean groups over five generations. Each group was initially designed with identical genetic backgrounds but selected for contrasting traits, one for high protein content (group P) and the other for high yield (group Y). Using fixation index (FST) and principal component analysis (PCA), the genetic structure and differentiation of the populations were tracked over selection cycles. Mean phenotypic responses were monitored to evaluate the efficiency of selection. This approach enabled us to understand the magnitude and direction of genomic change resulting from selection and how it correlates with phenotypic performance.

Initial Genetic Structure and Differentiation (Cycle 0)

Before selection was applied, the two founder groups displayed minimal genetic differentiation. Chromosome-wise FST values (Figure 1) were low, indicating a shared genetic base between the two groups. Although these groups were designed for distinct selection objectives (protein versus yield), the genetic structure had not yet diverged. PCA results (Figure 2) confirmed this, with individuals from both groups overlapping in multivariate space, showing no discernible clusters. This genetic similarity served as a baseline, allowing the effects of subsequent selection to be clearly attributed to breeding pressure rather than founder variation.

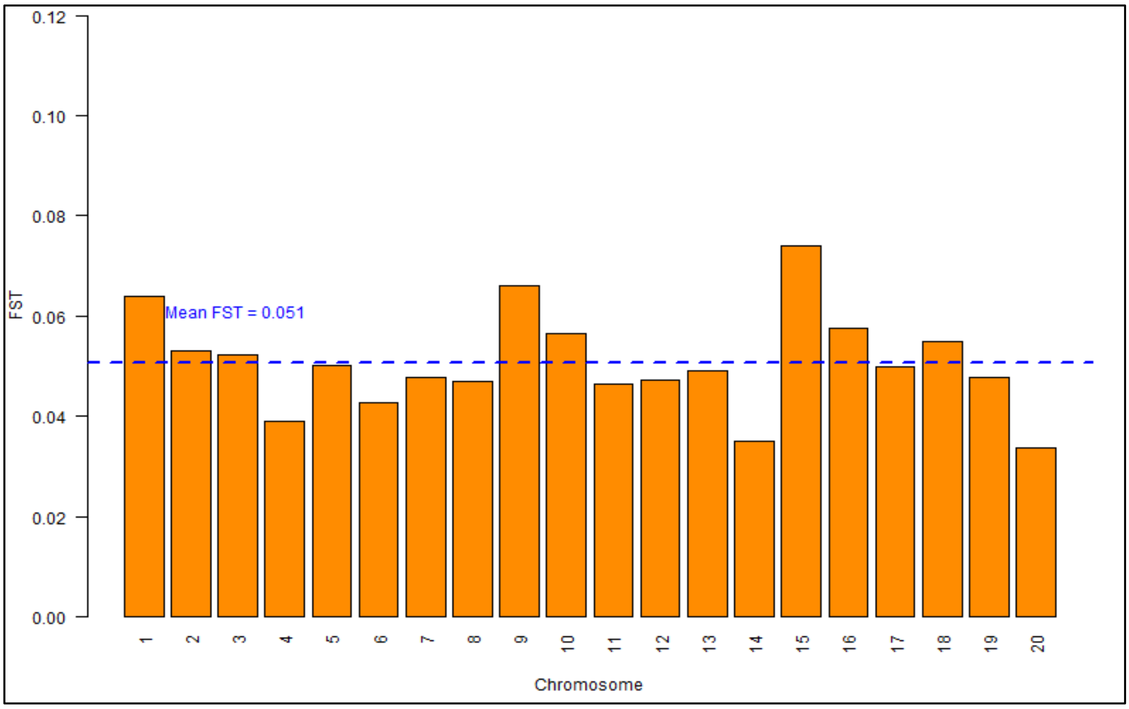

Figure 1. Chromosome-wise FST between individuals of the founder population (cycle 0).

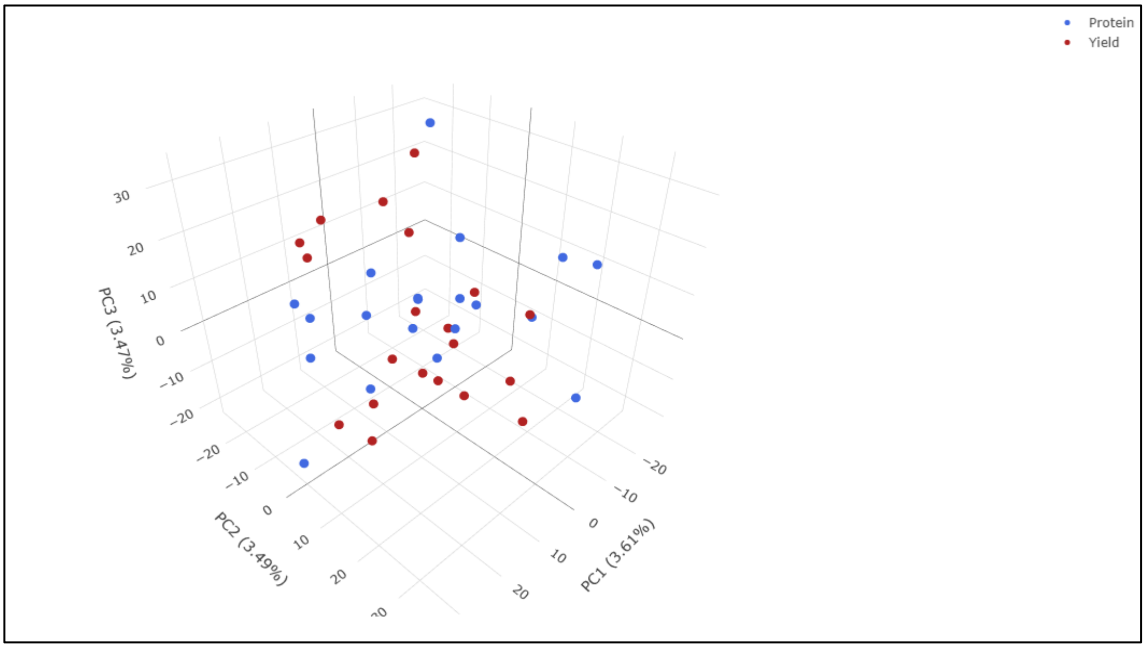

Figure 2. Principal component analyses plot (PCA) between individuals of the founder population (cycle 0).

FST Across Selection Cycles

Following selection, an increase in FST values was observed in both populations relative to their founder state. In group P, FST rose from 0.243 in Cycle 1 to 0.589 in Cycle 5 (Figure 3), while in group Y, it increased from 0.238 to 0.626 over the same period (Figure 4). These trends reflect selection pressure, favoring alleles associated with the respective target traits. The steady rise in FST over cycles is indicative of both reduced within-population genetic diversity and increased divergence from the ancestral gene pool. This suggests that only a subset of alleles, likely those conferring favorable phenotypic effects, were retained through recombination and selection.

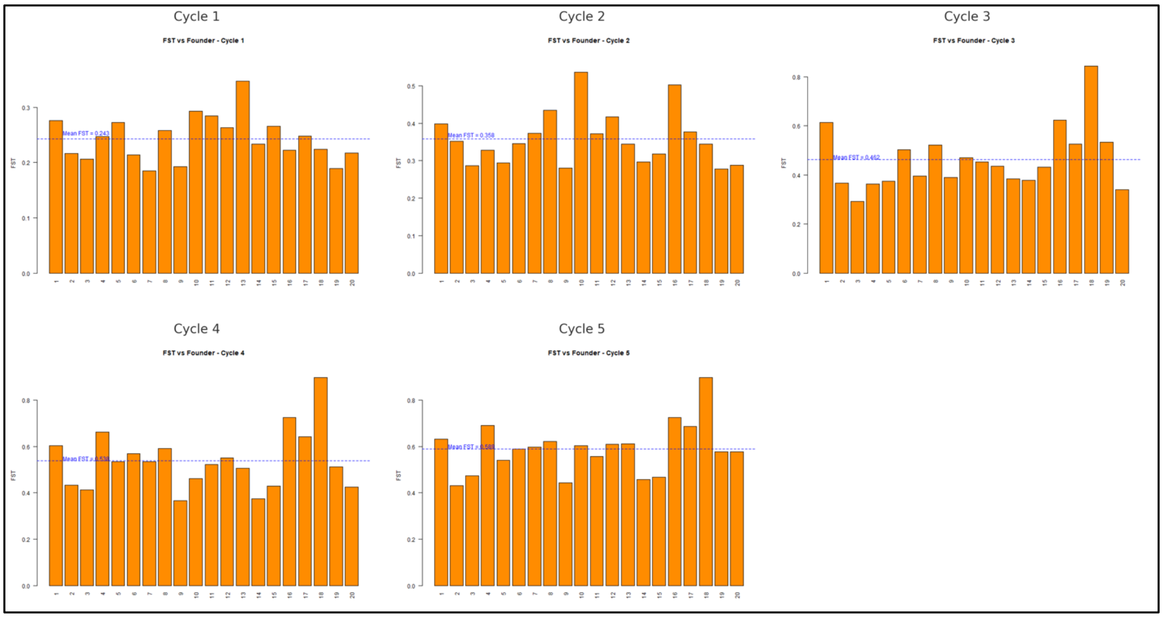

Figure 3. FST between protein population and founder population for each selection cycle.

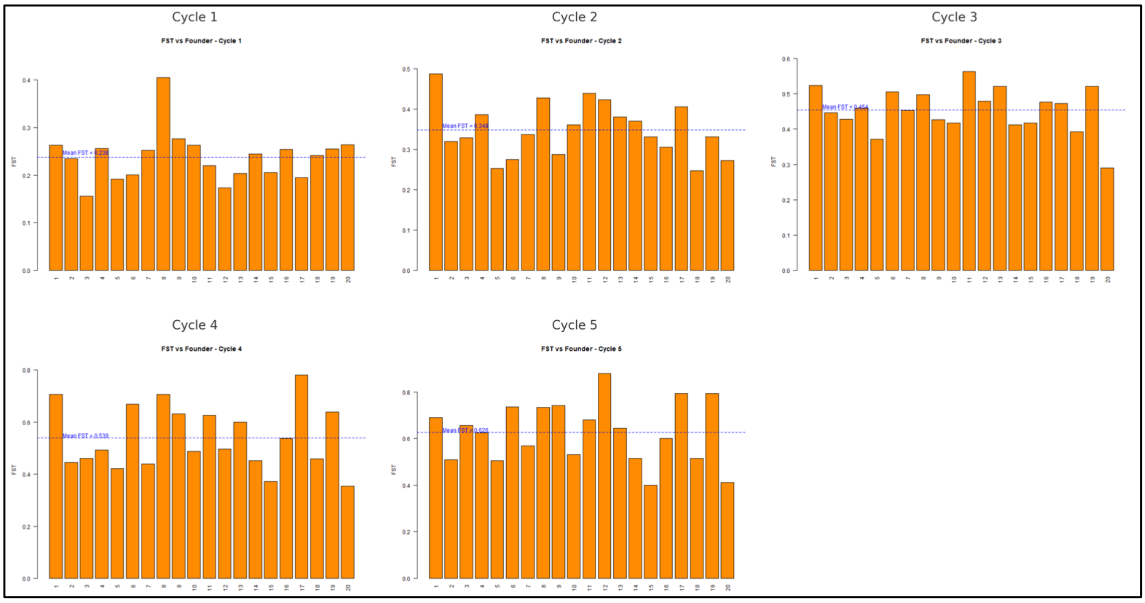

Figure 4. FST between yield population and founder population for each selection cycle.

To further investigate the impact of trait-based selection, we directly compared groups P and Y. Table 1 shows the interpopulation FST for each cycle, which increased from 0.277 in Cycle 1 to 0.622 by Cycle 5. This growing divergence emphasizes how selection for contrasting phenotypes reshapes the genome in divergent directions. As selection advanced, each population fixed or retained different allelic combinations that improved their respective trait values, which cumulatively enhanced genetic separation. This is particularly relevant in applied breeding programs where the trade-off between protein and yield must be managed strategically.

Table 1. FST between groups P and Y for each selection cycle.

Cycle	**FST**
Cycle 1	0.277
Cycle 2	0.389
Cycle 3	0.494
Cycle 4	0.566
Cycle 5	0.622

PCA Across Selection Cycles

The genomic divergence described by FST was also clearly visualized using PCA. In the early cycles, the overlap between groups was still visible, but from Cycle 1 onward, distinct clusters emerged for both protein- and yield-selected lines (Figures 5 and 6). In Cycle 5, PCA plots demonstrated complete separation, confirming that directional selection had generated distinct genetic trajectories. This pattern was reinforced when comparing the two groups at each cycle (Figure 7). These plots are consistent with our hypothesis of loss of shared alleles and accumulation of beneficial mutations or combinations specific to the selection regime.

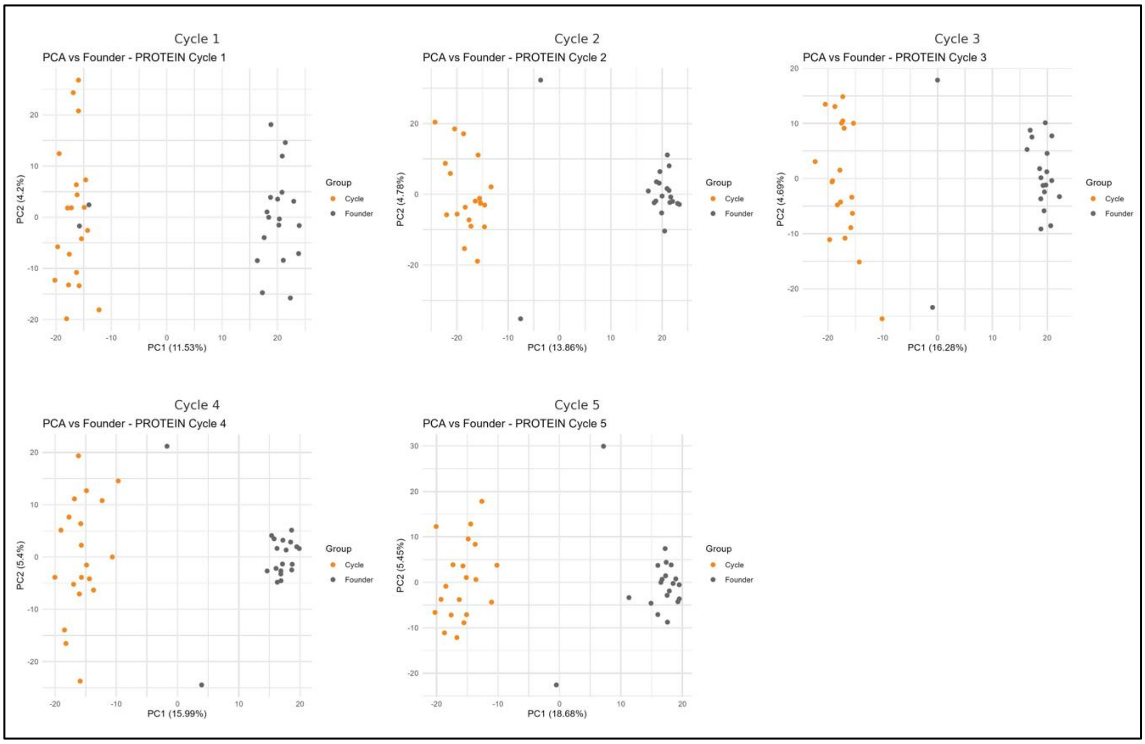

Figure 5. PCA between group P and founder population for each selection cycle.

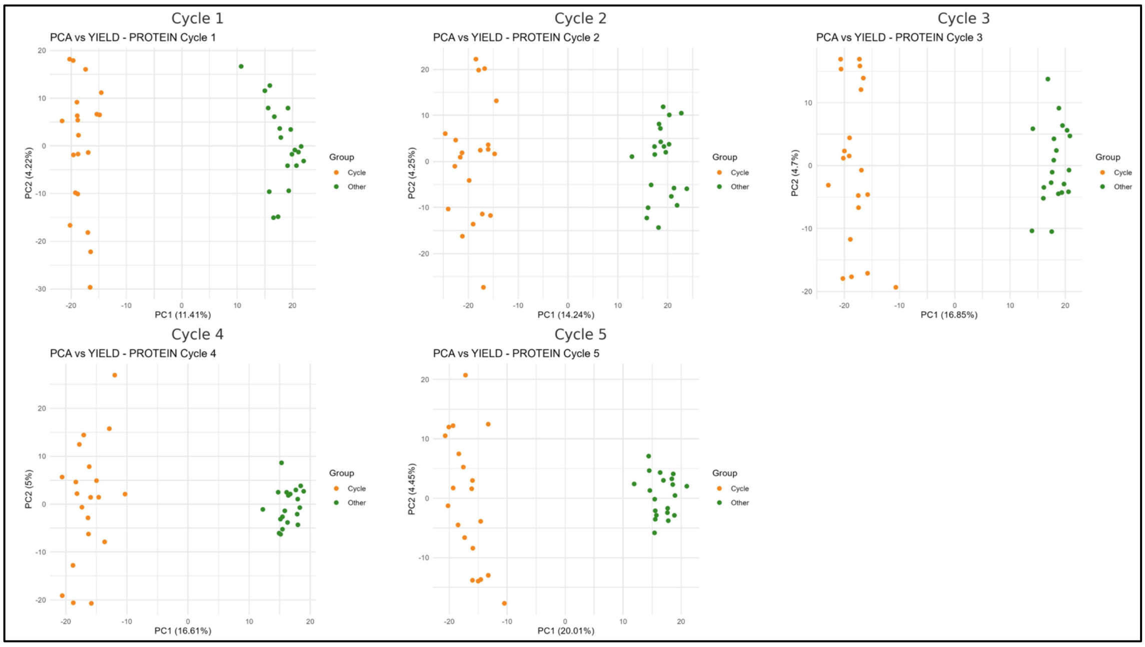

Figure 6. PCA between Y group and founder population for each selection cycle.

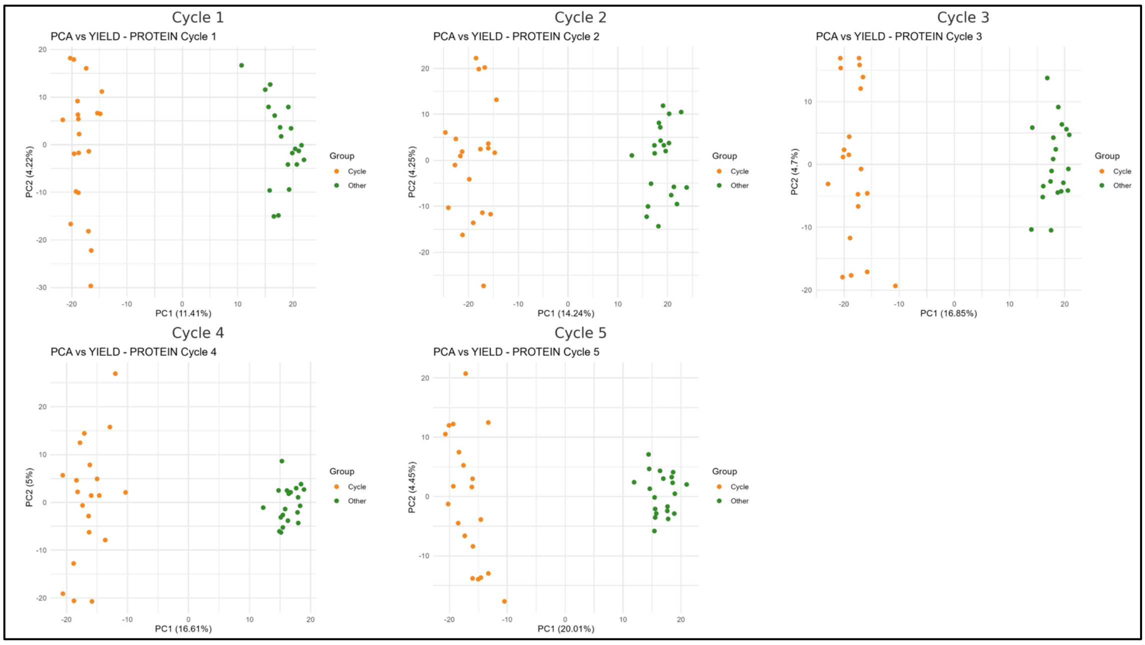

Figure 7. PCA between P and Y groups for each selection cycle.

FST and PCA in Cycle 5

Chromosome-specific FST analysis between the P and Y groups in Cycle 5 (Figure 8) revealed variable patterns of divergence. Some chromosomes showed stronger differentiation, potentially harboring loci with major effects for protein or yield traits. This heterogeneity is expected given the polygenic nature of both traits. It also suggests that despite whole-genome selection, specific genomic regions were disproportionately impacted. PCA of Cycle 5 (Figure 9) supports this observation, with tight clustering within populations and broad separation between them.

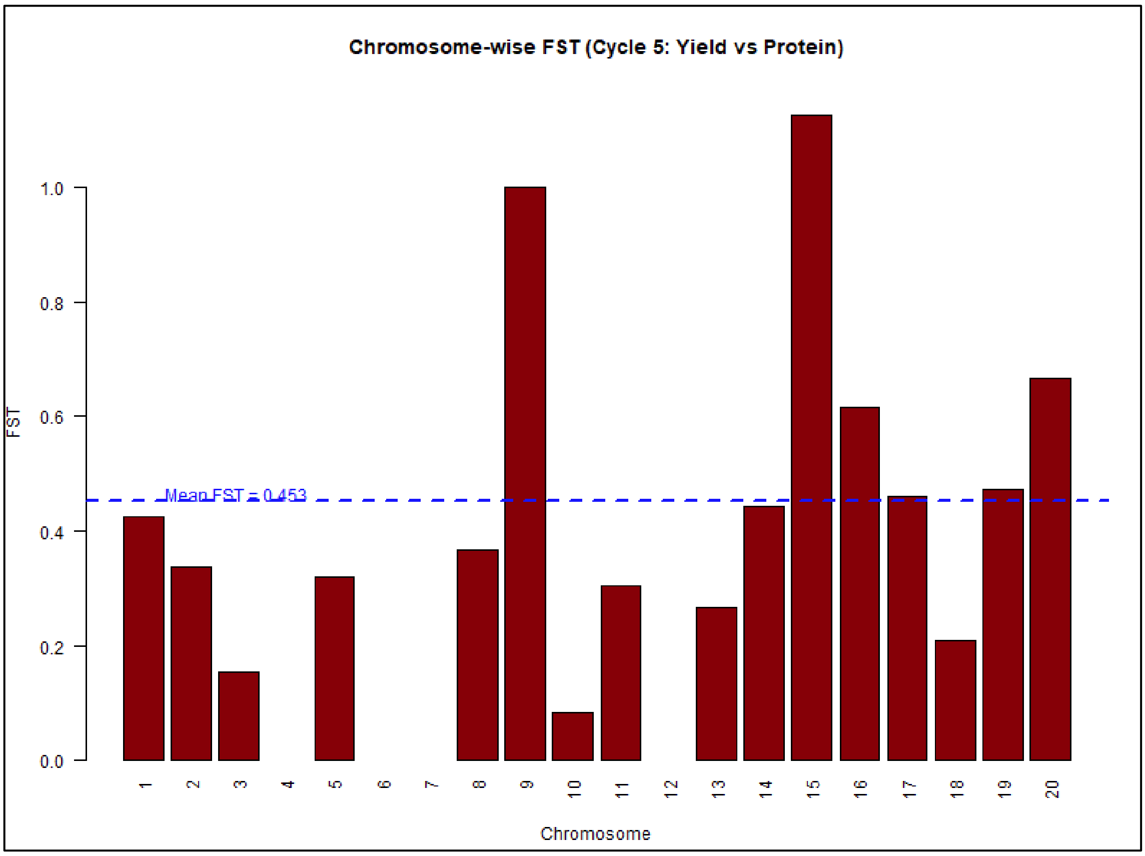

Figure 8. Chromosome-wise FST between yield and protein population in cycle.

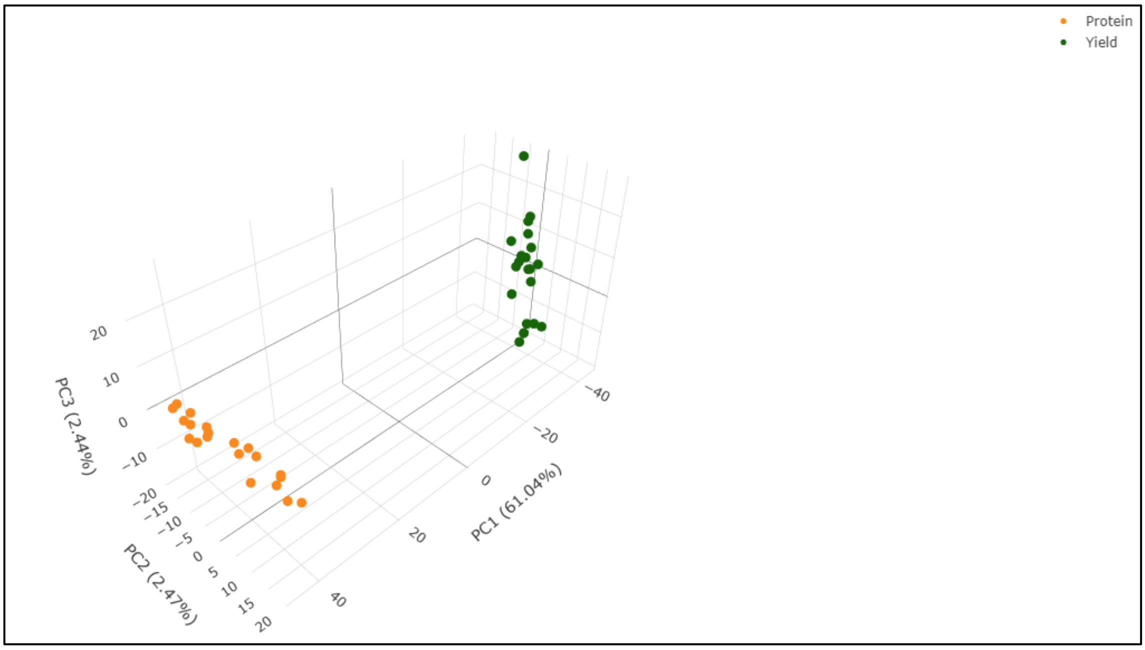

Figure 9. PCA between yield and protein population in cycle.

FST Values and Changes in Phenotypes Across Cycles

The consistent increase in FST coincided with favorable shifts in mean phenotypic values. For group P, average protein content rose across selection cycles (Figure 10). Likewise, group P exhibited rising yield values (Figure 11), highlighting selection success. It also reflects the narrowing of genetic diversity due to cycles of selection.

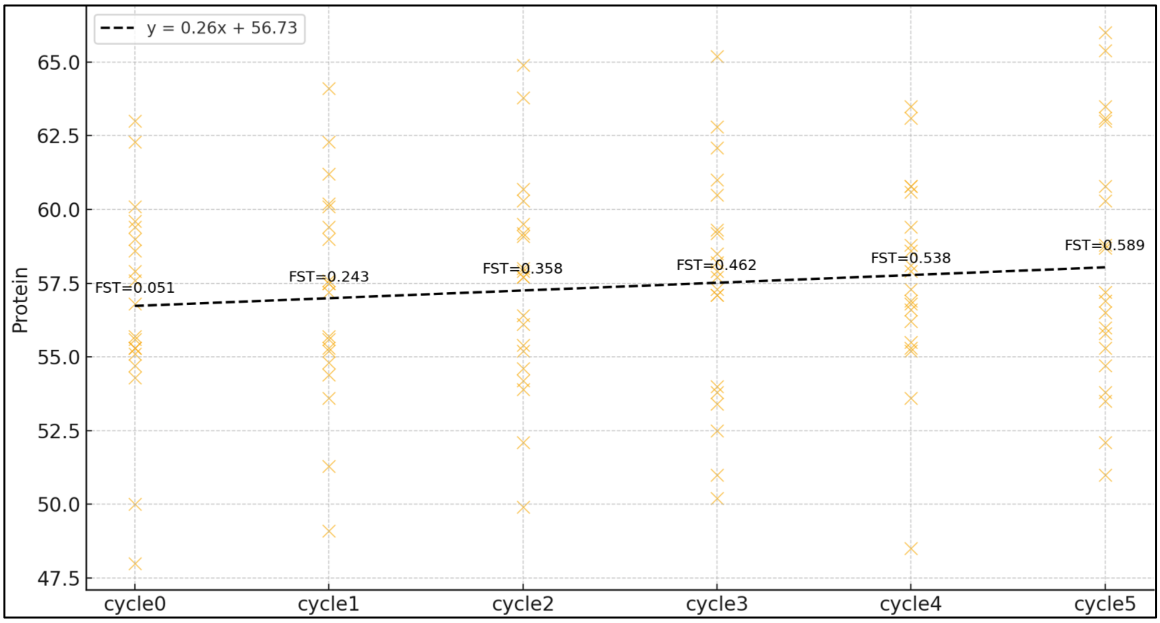

Figure 10. Changes in FST and protein content mean for each selection cycle for group P. Protein content (y-axis) is represented on a percentage basis.

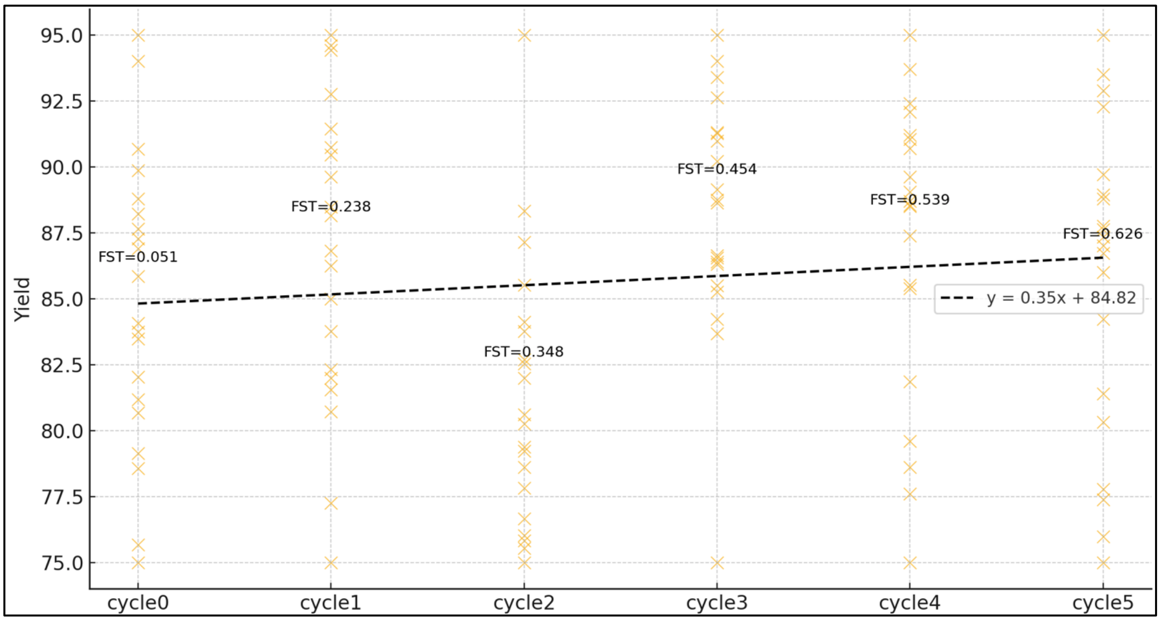

Figure 11. Changes in FST and yield content mean for each selection cycle for group Y. Yield (y-axis) is represented on a bushels per acre basis.

This study demonstrates the power and consequences of selection in breeding populations by combining population genomics with selection scenarios. It contributes to our understanding of how elite breeding pools diverge and interact, ultimately informing strategies to improve both yield and protein content in soybeans. Even starting from genetically similar founders, consistent selection for distinct traits produced rapid genomic divergence and substantial phenotypic gains. These outcomes reinforce the utility of genomic tools such as FST and PCA in monitoring genetic change. For breeders, the results underscore the trade-offs inherent in selection: while targeted gains are possible, they often come at the cost of genetic diversity. In practice, this necessitates striking a balance between progress and the maintenance of long-term variability. Moreover, the divergence observed here provides a rationale for developing pools based on protein and yield traits.

Future directions

The next phase of this study will assess the impact of FST in a new population generated from progenies resulting from crosses between the high-yield and high-protein lines in cycle 5. We expect that the mean FST will decrease relative to the founder population, as the cycle 6 population is expected to redistribute genetic variability from both parental pools across the genome. For the same reason and given the negative correlation between yield and protein content, it is also expected that the mean phenotypic values for both traits will decline compared to those observed in Cycle 5.

References

Andrijanić, Z., Nazzicari, N., Šarčević, H., Sudarić, A., Annicchiarico, P., & Pejić, I. (2023). Genetic diversity and population structure of European soybean germplasm revealed by single nucleotide polymorphism. *Plants, 12*(9), 1837.

Gaynor RC, Gorjanc G, Hickey JM (2021). “AlphaSimR: an R package for breeding program simulations.” G3 Gene|Genomes|Genetics, 11(jkaa07). https://doi.org/10.1093/g3journal/jkaa017.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Silva, A. C. D., Gregorio da Silva, D. C., Ferreira, E. G. C., Abdelnoor, R. V., Borém, A., Arias, C. A., ... & Marcelino-Guimarães, F. C. (2025). Genetic diversity, population structure in a historical panel of Brazilian soybean cultivars. *PloS one*, *20*(1), e0313151.

US Department of Agriculture Foreign Agricultural Service. (2017). Oilseeds: World Production Markets and Trade Reports (Washington, DC: FAS).

Yang, C., Yan, J., Jiang, S., Li, X., Min, H., Wang, X., & Hao, D. (2022). Resequencing 250 soybean accessions: new insights into genes associated with agronomic traits and genetic networks. *Genomics, Proteomics & Bioinformatics*, *20*(1), 29-41.

Zhang, J., Song, Q., Cregan, P. B., & Jiang, G. L. (2016). Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycine max). Theoretical and Applied Genetics, 129, 117-130.

Zhang, J., Wang, X., Lu, Y., Bhusal, S. J., Song, Q., Cregan, P. B., ... & Jiang, G. L. (2018). Genome-wide scan for seed composition provides insights into soybean quality improvement and the impacts of domestication and breeding. Molecular Plant, 11(3), 460-472.