Arthur Bernardeli
March 24, 2025
Exam 2 (Final) - AGRO 932

# Genomic Patterns in Distinct Soybean Breeding Pools: $F_{st}$ Dynamics and impacts of selection strategies on gains

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. Regarding soybean history, it was domesticated from wild soybean (*Glycine soja*) in East Asia around 6,000-9,000 years ago. Genomic and archaeological evidence points to the Huang-Huai Valley in Central China as the center of domestication. A semi-wild form, *Glycine gracilis*, appears to be a transitional species in this process. Recent studies confirm a gradual domestication pathway shaped by gene flow from wild populations (Sedivy et al., 2017).

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 ($F_{st}$), 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 intra-group crossing offers the potential to break unfavorable linkage blocks and generate superior recombinants that transcend historical trait limitations. These newly generated populations can be further selected using conventional phenotyping of quantitative traits (e.g., yield and protein), or using modern techniques such as genomic selection. Genomic selection leverages the kinship at a nucleotide level among individuals that are directly related or not, and it has shown its advantages to boos gains (Bandillo et al., 2023). However, comparing its long-term effectiveness is also fundamental for the success of breeding programs (Jannink, 2010).

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 $F_{st}$, chromosome-specific patterns of divergence, and changes in within-group diversity, trait distributions, and phenotypic and genomic selection dynamics over gains.

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 $F_{st}$ 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 $F_{st}$ values relative to the founder generation.

The third aim explores the consequences of selecting the two distinct breeding pools using phenotype and genome-based selection. This will mimic a long-term plan of selection cycles in soybean breeding programs, helping breeders decide which selection

strategy best answers their needs and when new genetic variability needs to be reintroduced in the breeding pipeline.

## Material and Methods

### *Population Simulation and Structure*

To investigate genomic divergence and selection responses in soybean breeding pools, we simulated one diverse founder population per year for five years, each consisting of 20 high-yield lines and 20 high-protein lines. These populations 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) were 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) were 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 software (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 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. The heritability and correlation values used in this simulation were derived from empirical estimates from a real-world dataset provided by the University of Nebraska-Lincoln Soybean Breeding Program. This dataset comprises approximately 3,000 breeding lines evaluated across 18 locations over five years (2020-2024). Although the data set is unbalanced, reflecting a typical breeding pipeline's structure, broad-sense heritability was estimated for yield and protein content. The resulting average heritability estimates were used to parameterize the simulation, ensuring that the simulated trait architecture closely reflects observed variability and genetic control in elite soybean germplasm.

Trait values were assigned at the founder level, generating variation for groups Y and P in the founder population. This variation allowed for the initial assessment of divergence and provided a foundation for within-population selection in subsequent cycles. This study assumes a fixed number of QTLs and a set of genetic correlations. Future studies could extend this framework by simulating different numbers of underlying QTLs and varying levels of linkage disequilibrium between yield- and protein-associated loci.

*Experimental Procedure*

Step 1: Founder Population Simulation

The initial step involved simulating one founder population of 40 individuals per year in five years, each with assigned genotypes and trait values according to the simulation parameters. 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 before being assigned to crosses. One population was transferred to selection in year *n*, with their individuals redirected to crosses in year *n+1*, to ensure the continuity of the breeding program. 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. This step is usually taken when establishing breeding programs, where diverse populations with minimal information about qualitative and quantitative traits, such as yield and protein content in soybeans, are available. At the end of the fifth cycle, the breeder is expected to have two core pools: high-yield lines and high-protein lines.

Step 3: $F_{st}$, PCA, and phenotypic analysis

Following the founder simulation, genome-wide and chromosome-specific $F_{st}$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. $F_{st}$ 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. $F_{st}$ was calculated according to the formula:

$$F_{st} = \frac{\sigma^2}{\bar{p}\bar{q}} \text{ (Equation 1),}$$

Where $\sigma^2$ is the observed sample variance in the frequency of allele $A_1$ among populations, and $\bar{p}$ and $\bar{q}$ is the average frequence of alleles $A_1$ and $A_2$ among populations.

Step 4: Selection scenarios

After the fifth cycle of recurrent selection, the five top-ranked individuals in groups Y and P were intermated within groups in a complete diallel to generate 10 new populations, each contributing 100 lines to form a set of 1000 individuals to be tested and selected according to phenotypic and genomic-selection strategies. The top five selected individuals would then be intermated, and their progenies would be reevaluated according to phenotypic and genomic selection until a plateau in genetic gains was reached. This will be fundamental to identify the potential and limitations of phenotypic and genomic selection for major traits in soybean breeding and the key aspects of exploiting genetic variation in quantitative traits. For this section, the crossing (intermating) parameters were the same as described in the *Simulation Platform and Trait Architecture* section,

and the simulations were performed using the AlphaSimR package (Gaynor et al., 2021) in R software (R Core Team, 2021).

A ridge-regression best linear unbiased prediction (RR-BLUP; Endelman, 2011) model was deployed to select individuals for the genomic selection pipeline. Parallel to that, 500 individuals were randomly pulled from the original group of 1000 individuals to assess the prediction accuracy of the proposed model, where 250 individuals were assigned to the training set, and the remaining 250 were assigned to the testing set. This is a routine procedure in plant breeding programs that deploy genomic selection yearly to know the expected accuracy when selecting their materials in a real-world genomic selection scenario. This context of updating training sets and model fitting approaches to maximize the potential of the genomic selection and the test-half predict-half pipeline is well described in Atanda et al. (2021a, 2021b). For this study, the RR-BLUP model fitting was as follows:

$$y = \mu + Zu + e \text{ (Equation 2)},$$

where $y$ is the vector of phenotypic values of protein or yield, $\mu$ is the overall mean, $Z$ is the design matrix (incidence matrix) of markers, $u \sim N(0, \sigma_u^2)$ is the vector of random marker effects, and $e \sim N(0, \sigma^2)$ is the vector of residuals. The genomic estimated breeding value (GEBV) was estimated as:

$$GEBV_i = \sum_{j=1}^{m} Z_{ij}\hat{u}_j \text{ (Equation 3)},$$

where $m$ is the total number of $j$ markers for each $i$ phenotypic observation. The marker effects are estimated as:

$$\hat{u} = (Z^T Z + {}^{\sigma^2}/_{\sigma_u^2} I)^{-1} Z^T y \text{ (Equation 4)},$$

where $I$ represents an identity matrix. The ridge regression analyses were executed using the rrBlup package (Endelman et al., 2011) in R software (R Core Team, 2021).

*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 ($F_{st}$) 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 $F_{st}$ 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.

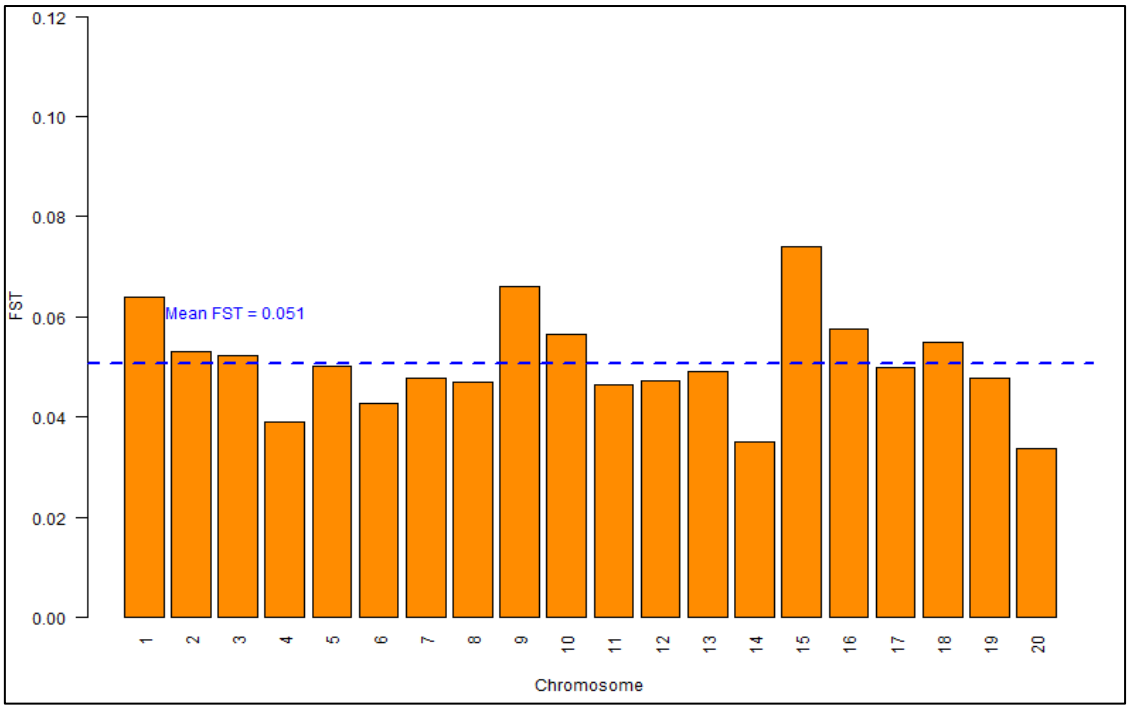



**Figure 1.** Chromosome-wise $F_{st}$ between individuals of the founder population (cycle 0).

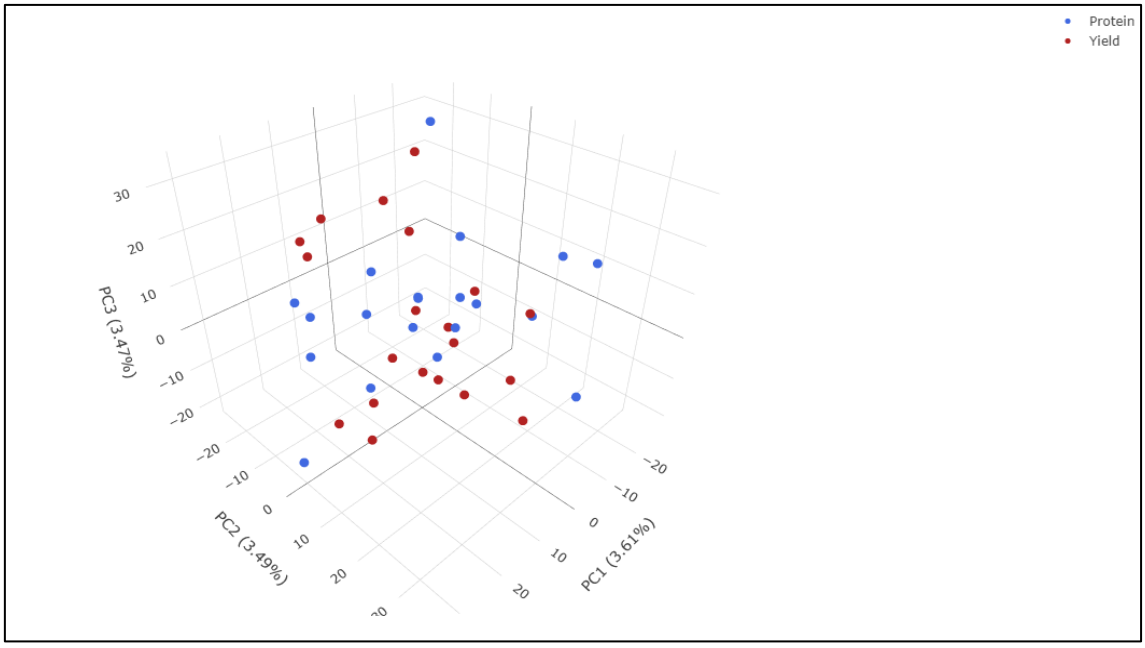



**Figure 2.** Principal component analyses plot (PCA) between individuals of the founder population (cycle 0).

## $F_{st}$ *Across Selection Cycles*

Following selection, an increase in $F_{st}$ values was observed in both populations relative to their founder state. In group P, $F_{st}$ 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 $F_{st}$ 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.

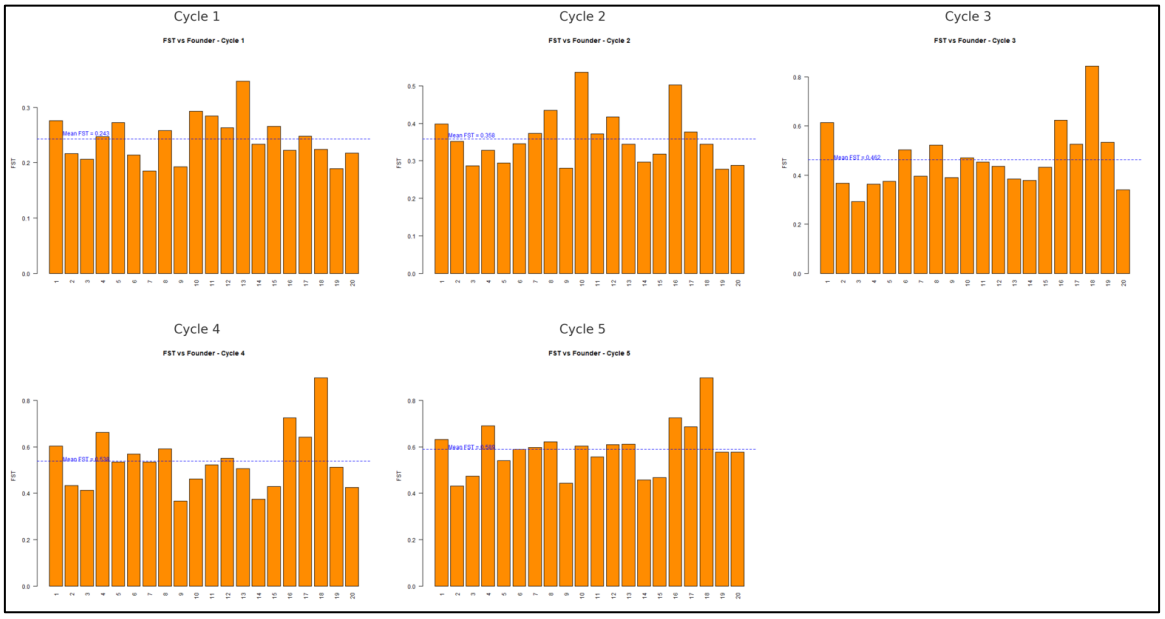



**Figure 3.** $F_{st}$ between protein population and founder population for each selection cycle.

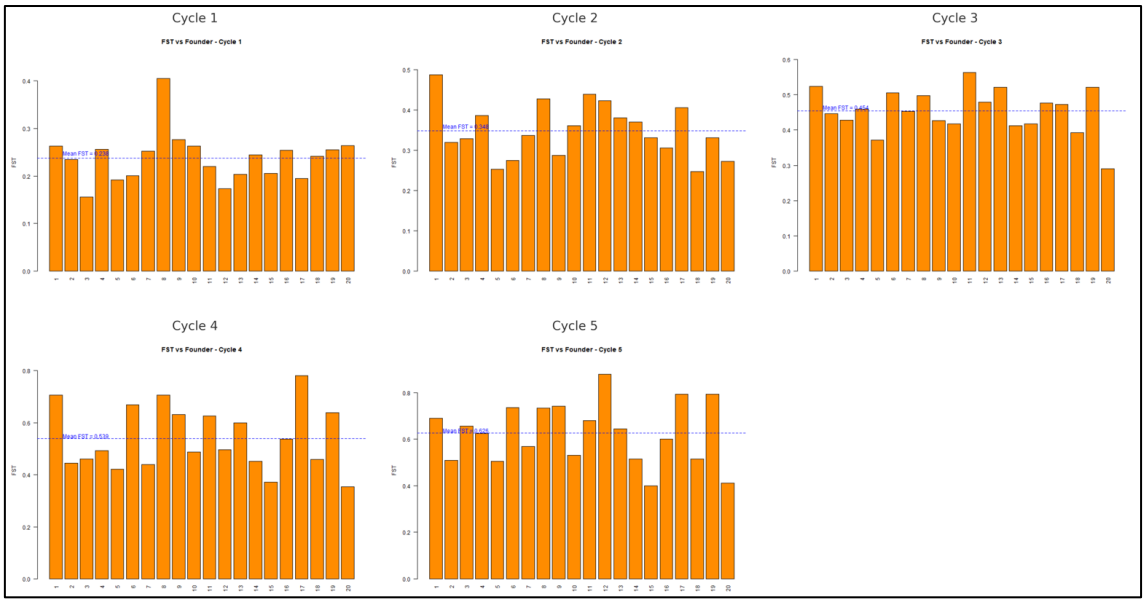



**Figure 4.** $F_{st}$ between yield population and founder population for each selection cycle.

To further investigate the impact of trait-based selection, we directly compared groups P and Y. Table 1 shows the interpopulation $F_{st}$ 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.** $F_{st}$ 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 $F_{st}$ 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.

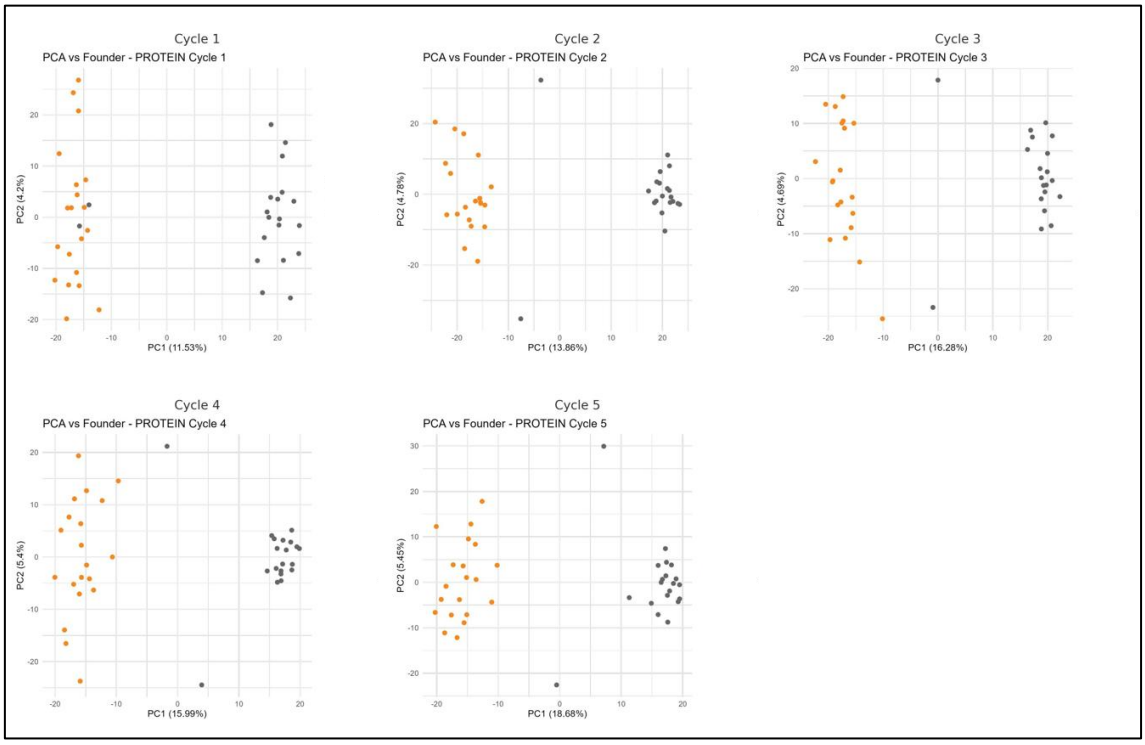



**Figure 5.** PCA between the group P and the founder population for each selection cycle. Yellow dots represent individuals from the improved populations for the current cycle, and gray dots represent individuals from the founder population.

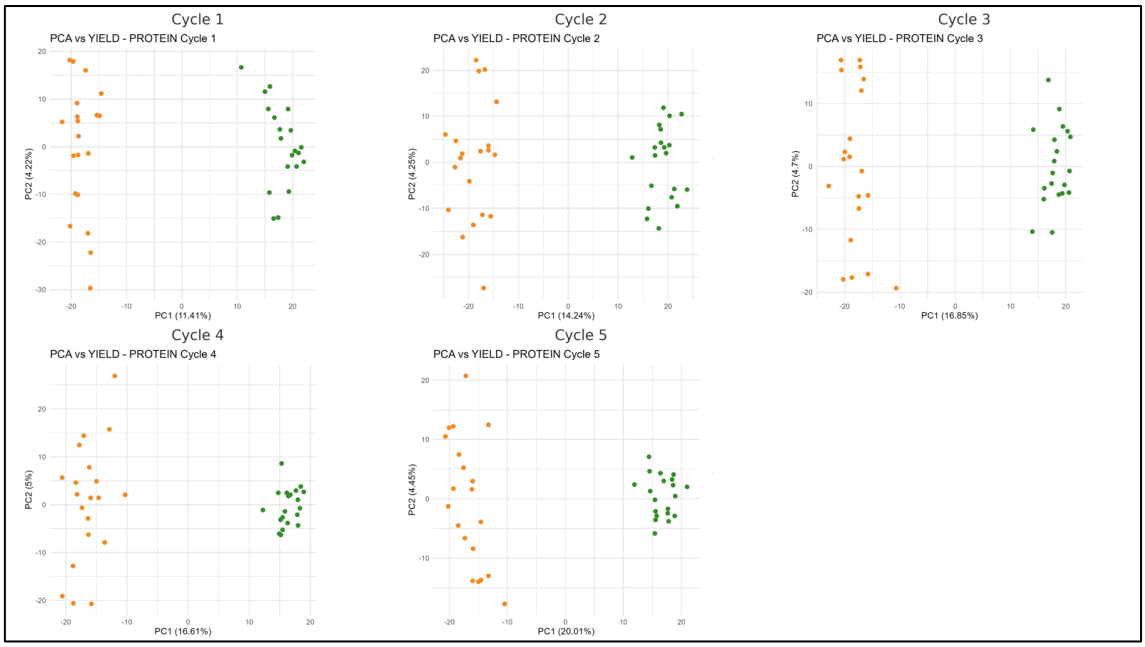



**Figure 6.** PCA between the Y group and the founder population for each selection cycle. Yellow dots represent individuals from the improved populations for the current cycle, and green dots represent individuals from the founder population.

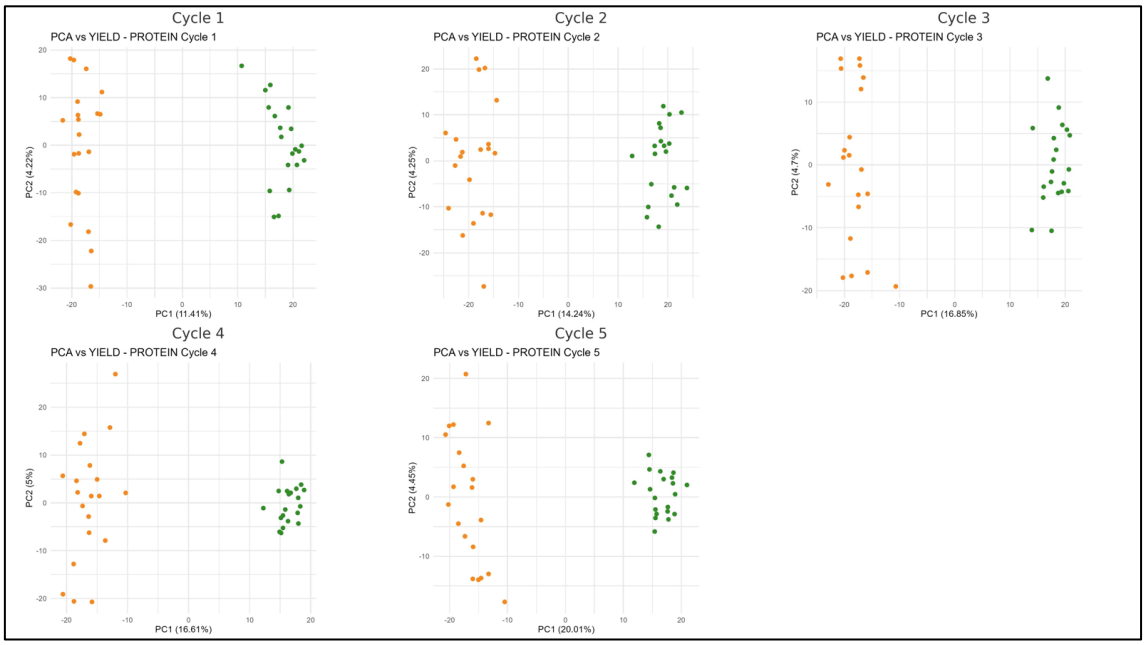



**Figure 7.** PCA between P and Y groups for each selection cycle. Yellow dots represent individuals from the P group, and green dots represent individuals from the Y group.

*FST and PCA in Cycle 5*

Chromosome-specific $F_{st}$ 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.

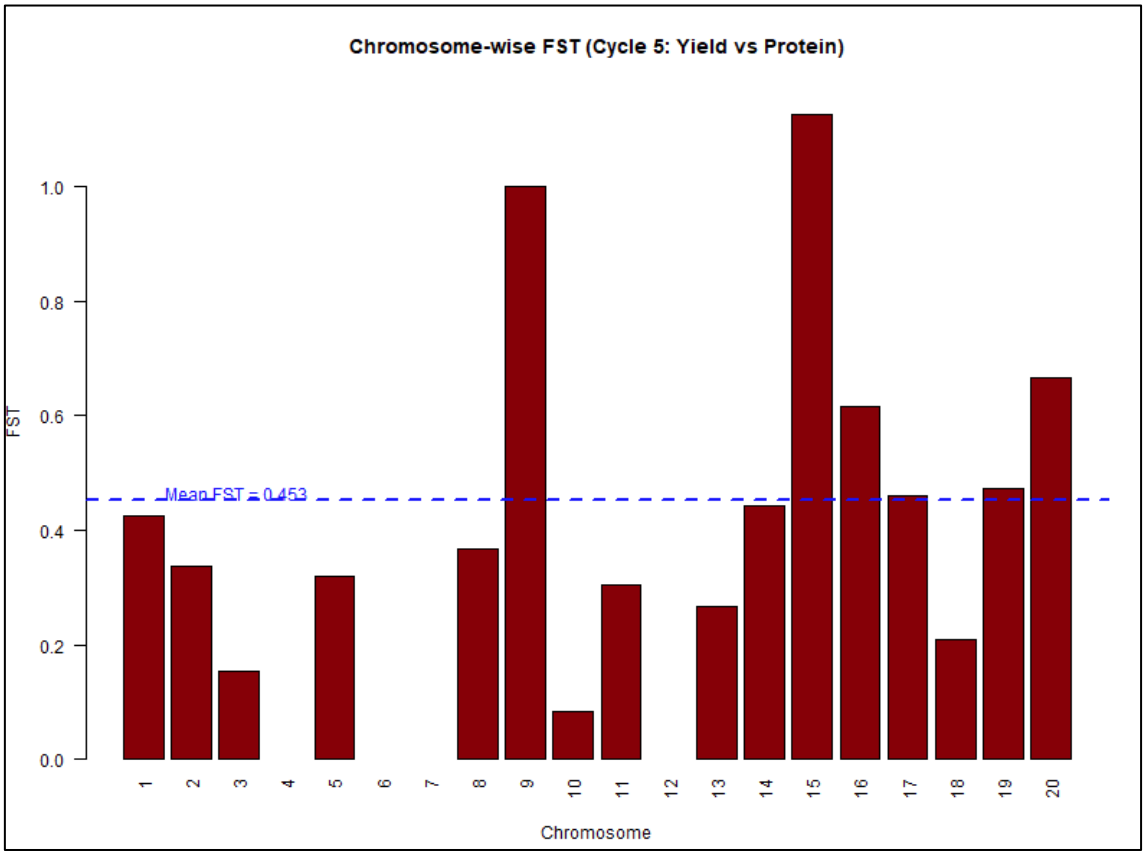



**Figure 8.** Chromosome-wise $F_{st}$ between yield and protein population in cycle 5.

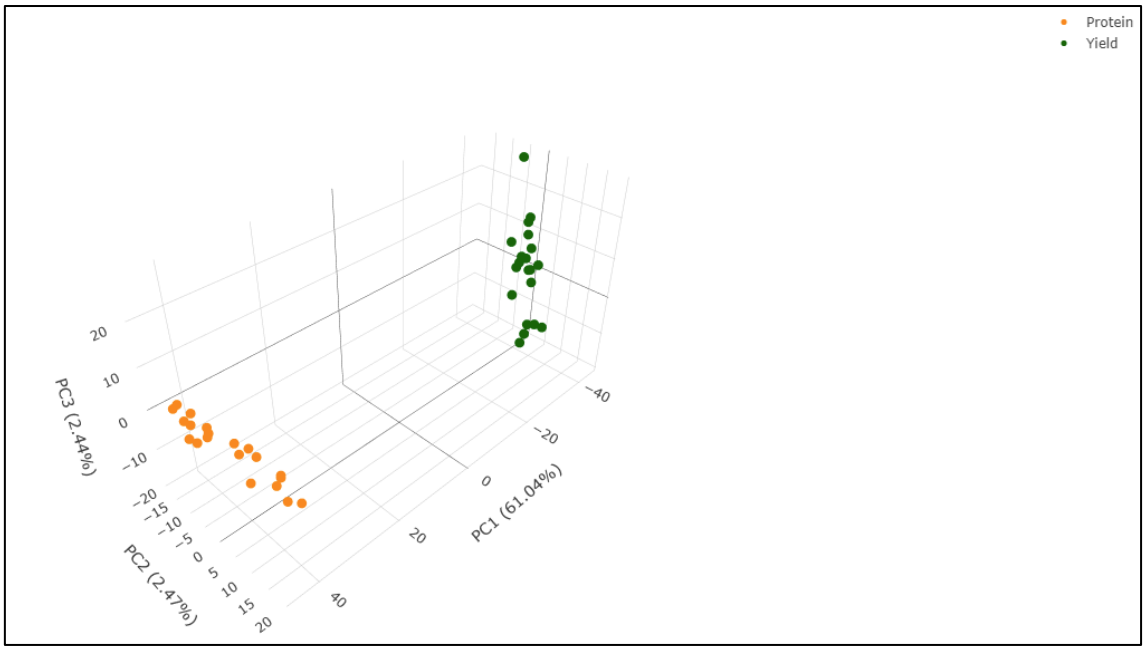



**Figure 9.** PCA between yield and protein population in cycle 5.

245 $F_{st}$ *Values and Changes in Phenotypes Across Cycles*

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

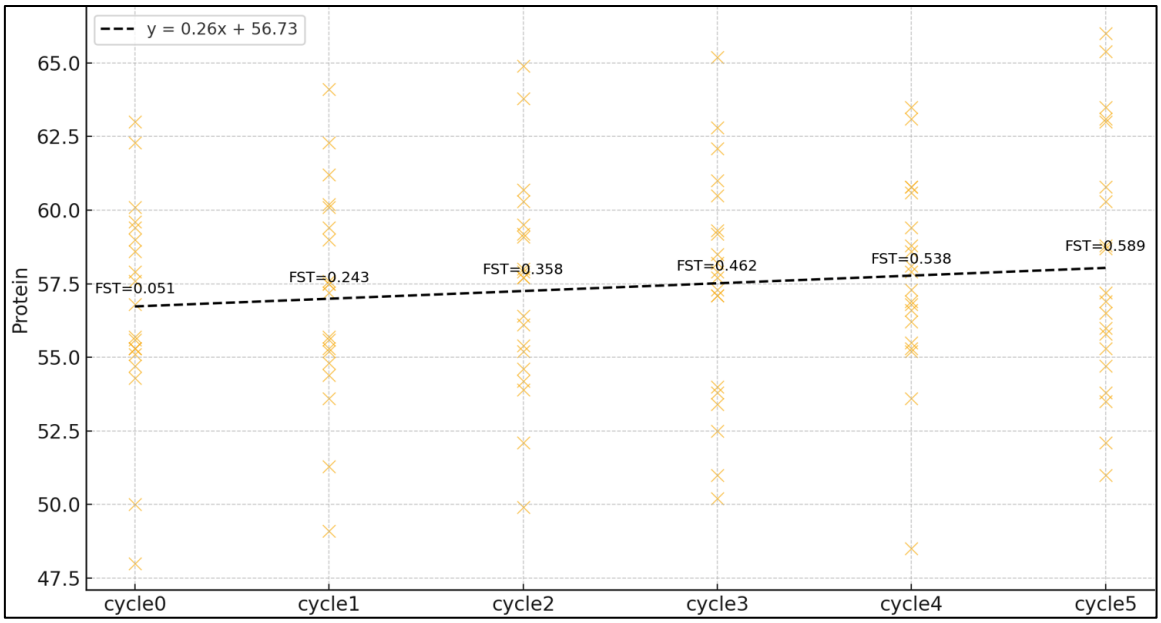



**Figure 10.** Changes in $F_{st}$ and protein content mean for each selection cycle for group P. Protein content (y-axis) is represented on a percentage basis.

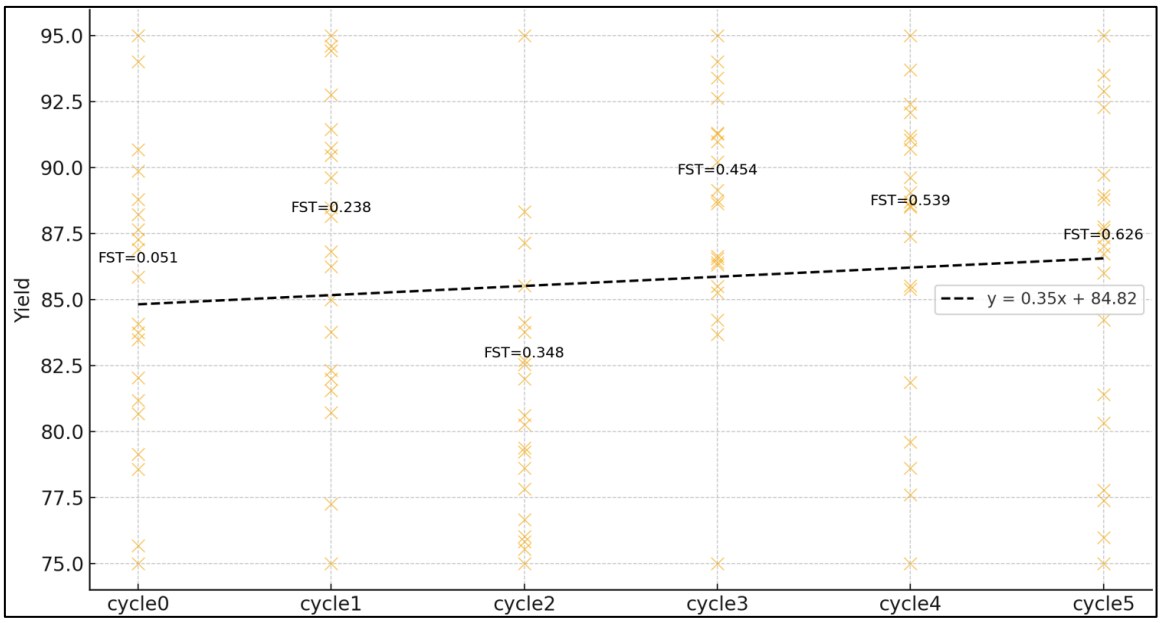



**Figure 11**. Changes in $F_{st}$ and yield content mean for each selection cycle for group Y. Yield (y-axis) is represented on a bushels per acre basis.

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 $F_{st}$ 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.

*Phenotypic and genomic selection strategies*

Recurrent selection followed by phenotypic or genomic selection is important to boost gains in breeding programs. Our study simulated five recurrent selection cycles and ten consecutive phenotypic or genomic selection cycles on protein and yield populations.

For the phenotypic selection, yield showed an asymptotic increase in gains through cycles. These increases were more evident between cycles 1-6, but a plateau was reached between cycles 7-10 (Figure 12). This suggests that the diminishing gains result from finite genetic variance because of selection (Falconer & Mackay, 1996). The initial rapid response to selection was due to the high frequency of favorable alleles. However, these alleles were fixed across initial cycles, which led to a reduction in the variability and consequently lower gains.

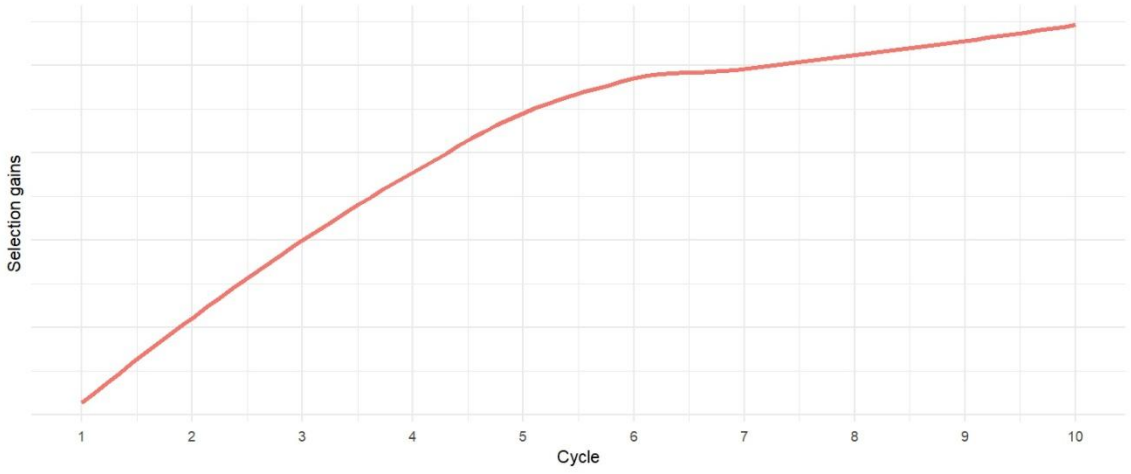



**Figure 12**. Selection gains for yield in group Y across cycles of phenotypic selection.

The results for phenotypic selection of protein for group P also show cumulative positive gains across cycles of selection but little cumulative gain after cycle 6 (Figure 13). This is similar to what was displayed for phenotypic selection on yield, where the variance was exhausted over time when no new genetic variability was introduced to the breeding pipeline.

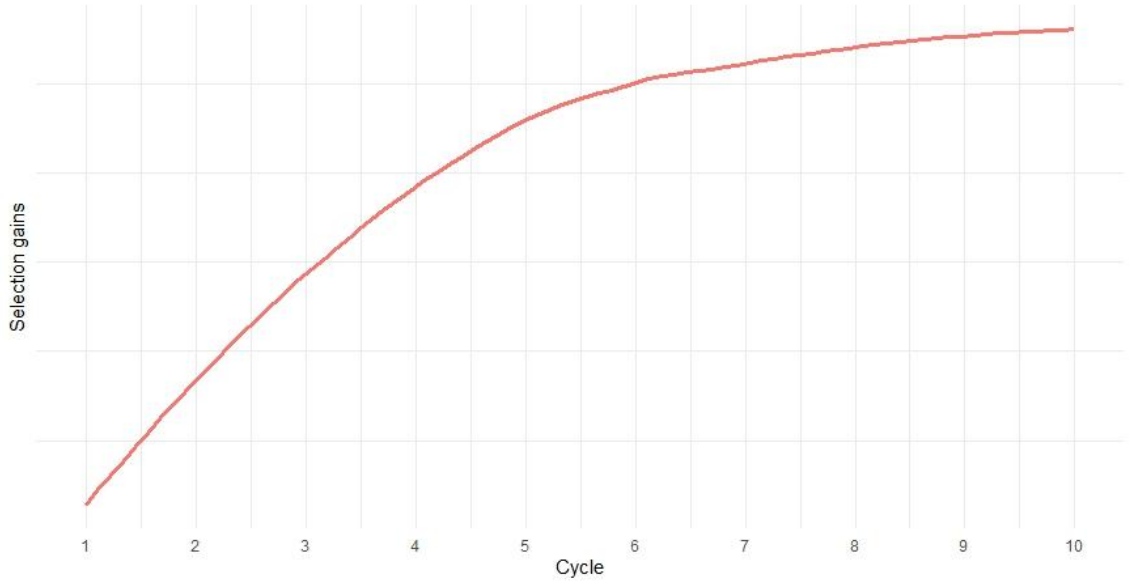



**Figure 13**. Selection gains for protein in group P across cycles of phenotypic selection.

Opposed to what was shown for phenotypic selection for yield and protein, the genomic selection showed initial decreases in gain in initial cycles for both traits (protein in group P and yield in group Y). In both cases, the initial decrease in gains was followed by a sharp increase; then a plateau was reached. For protein, the initial drop was smoother, and the increase in gains initiated in cycle five and continued until cycle seven, where it peaked. However, it decreased rapidly in the following cycles (Figure 14). For yield, despite the more evident initial decrease in gains, genomic selection remained positive for a more extended period, from cycle three until cycle 7, followed by a sharp decrease in subsequent cycles (Figure 15).

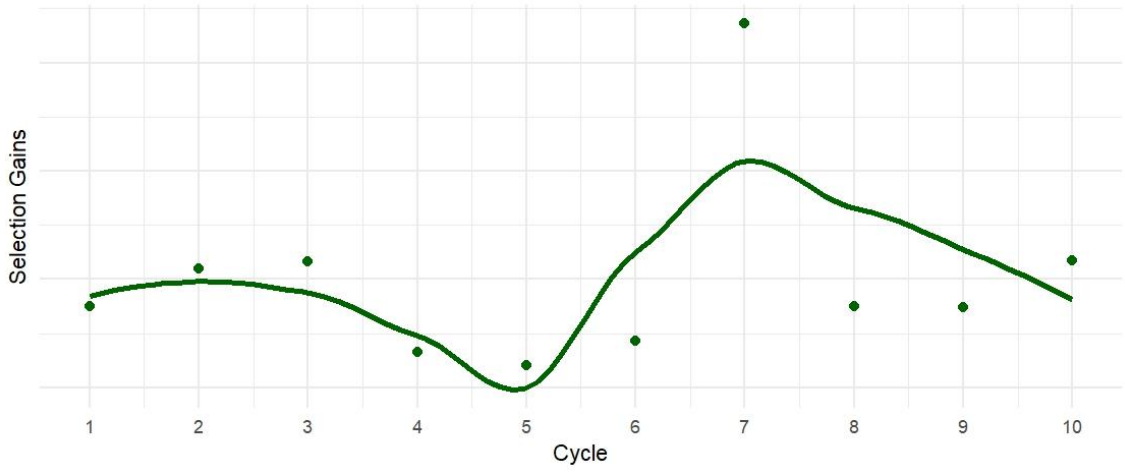
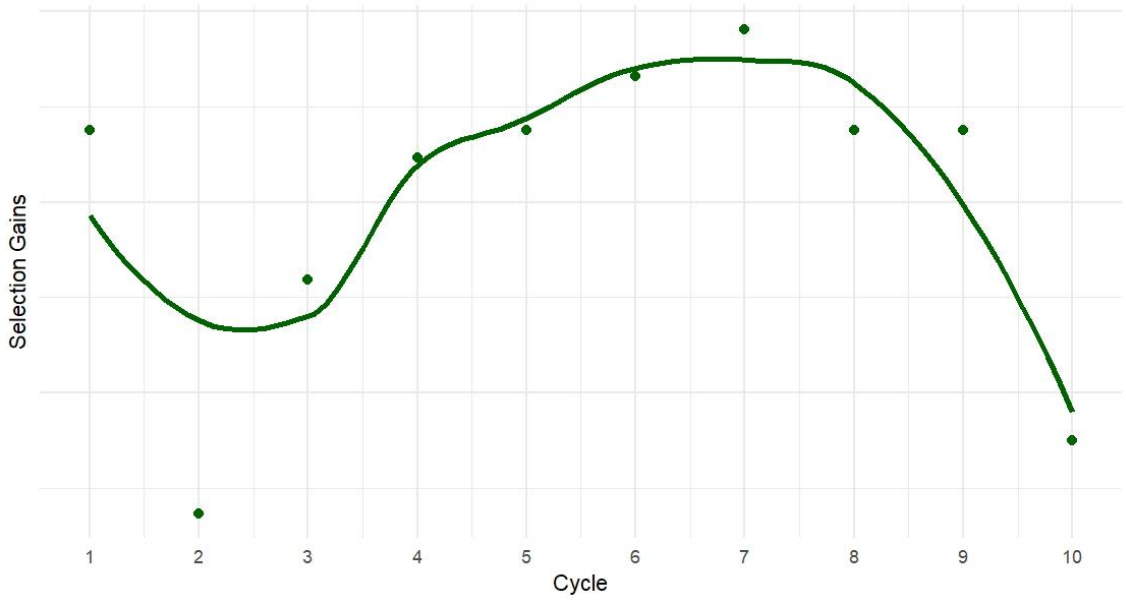



**Figure 14**. Selection gains for protein in group P across cycles of genomic selection.

The gain patterns are aligned to the prediction accuracies from the RR-BLUP. For yield, accuracy ranged from 33.05 to 43.98%, indicating moderate predictive ability. For protein, accuracies were higher and ranged from 39% to 56.02% (Table 2). Lower accuracies coincide with drops in selection gains, steady accuracies coincide with little to no increment in selection gains, and large accuracy values coincide with increments in selection gains.




303

304 **Figure 15**. Selection gains for yield in group Y across cycles of genomic selection.

305 Lower or negative gains in initial cycles might be attributed to poor optimization
306 of the training population, which reflects inadequate selection in the target population
307 (Akdemir et al, 2015). The increase in gains might be attributed to the RR-BLUP model
308 selecting highly related individuals over time due to the marker shrinkage, which favors
309 lines whose marker profiles resemble the training population or previously selected
310 individuals (Goddard, 2009). It can be said that this was an unintentional training set
311 optimization; however, in a long-term perspective, it can lead to narrow variability of the
312 breeding pool (Jannink, 2010). Finally, the plateau or drop in both genomic and
313 phenotypic selection can be attributed to the lack of introducing new variability to the
314 breeding pool, where no more available genetic variation could be exploited.

315 **Table 2**. Prediction accuracies of the RR-BLUP model for genomic selection across
316 cycles for yield (group Y) and protein (group P).

317

| Cycle | Yield | Protein |
|---|---|---|
| 1 | 43.05 | 49.01 |
| 2 | 40.01 | 53.97 |
| 3 | 33.99 | 48.97 |
| 4 | 38.99 | 39 |
| 5 | 40.92 | 38.01 |
| 6 | 43.98 | 55.99 |
| 7 | 41.01 | 56.02 |
| 8 | 40.57 | 50.99 |
| 9 | 39.98 | 49.99 |
| 10 | 33.05 | 47.97 |

## Conclusions

This study illustrates germplasm adaptation to elite breeding pools (cycles one through five). This was followed by genomic and phenotypic selection over the elite breeding pools (second round of ten cycles). Results of phenotypic and genomic selection were contrasting despite the same objectives initially outlined for both: to maximize gains. However, the genomic selection approach was more volatile, and it showed that it is necessary to rethink the genomic selection pipeline according to multiple criteria such as training and target population optimization, model-fitting optimization, and high-quality field testing for accurate records of phenotypic values.

## References

Akdemir, D., Sanchez, J. I., & Jannink, J. L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, *47*, 1-10.

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.

Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., ... & Robbins, K. R. (2021a). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theoretical and Applied Genetics*, *134*, 279-294.

Atanda, S. A., Olsen, M., Crossa, J., Burgueño, J., Rincent, R., Dzidzienyo, D., ... & Robbins, K. R. (2021b). Scalable sparse testing genomic selection strategy for early yield testing stage. *Frontiers in Plant Science*, *12*, 658978.

Bandillo, N. B., Jarquin, D., Posadas, L. G., Lorenz, A. J., & Graef, G. L. (2023). Genomic selection performs as effectively as phenotypic selection for increasing seed yield in soybean. *The Plant Genome*, *16*(1), e20285.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The plant genome*, *4*(3).

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.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, *136*(2), 245-257.

Jannink, J. L. (2010). Dynamics of long-term genomic selection. *Genetics Selection Evolution*, *42*, 1-11.

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/.

Sedivy, E. J., Wu, F., & Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. New Phytologist, 214(2), 539-553.

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.