1 **Genomic Patterns in Distinct Soybean Breeding Pools:** ~~Selection and~~ ~~F_{st}~~**FST**

2 **Dynamics and impacts of selection strategies on gains**

3   Soybean (*Glycine max* (L.) Merr.) is a globally important crop, cultivated for its

4 high-value seeds rich in both oil and protein. It serves as a critical component of food,

5 feed, and industrial markets, with significant production hubs across the Americas and

6 Asia (US Department of Agriculture Foreign Agricultural Service, 2017). In the United

7 States, soybean is a cornerstone of agricultural output and economic stability, occupying

8 millions of acres annually. Its dual utility, approximately 40% protein and 20% oil in seed

9 composition makes it especially valuable for both human consumption and livestock feed.

10 Given its essential role in global nutrition and supply chains, improving soybean traits

11 remains a primary objective for plant breeding programs worldwide. Regarding soybean

12 history, it was domesticated from wild soybean (*Glycine soja*) in East Asia around 6,000-

13 9,000 years ago. Genomic and archaeological evidence points to the Huang-Huai Valley

14 in Central China as the center of domestication. A semi-wild form, *Glycine gracilis*,

15 appears to be a transitional species in this process. Recent studies confirm a gradual

16 domestication pathway shaped by gene flow from wild populations (Sedivy et al., 2017).

17   Among the most economically and nutritionally significant traits in soybeans are

18 seed yield and seed protein content. However, these traits are negatively correlated,

19 presenting a substantial challenge for breeders. Genetic gains in seed yield have often

20 been accompanied by declines in protein concentration, and efforts to increase protein

21 content frequently come at the expense of yield potential. This antagonistic relationship

22 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}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~~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

2

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}~~FST~~, chromosome-specific patterns of divergence, and changes in within-group diversity, ~~and~~ 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}~~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 F_{st}~~FST~~ 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

72   strike: strategy best answers their needs and when new genetic variability needs to be

73   underline: reintroduced in the breeding pipeline. strike: crossing between the divergent breeding pools,

74   followed by selection in the recombinant progeny. This scenario is designed to assess

75   whether recombination can effectively break the negative correlation between yield and

76   protein, allowing for the development of superior lines that combine favorable alleles

77   from both parental groups.

78   **Material and Methods**

79   *Population Simulation and Structure*

80   To investigate genomic divergence and selection responses in soybean breeding

81   pools, we simulated one diverse founder population per year for five years, each

82   consisting of 20 high-yield lines and 20 high-protein lines. These populations will then

83   represent breeding pools with distinct selection histories, one selected for high seed yield

84   and the other for high seed protein content. Genotypic data were simulated using 6,000

85   evenly spaced single-nucleotide polymorphism (SNP) markers, distributed across the 20

86   soybean chromosomes (~300 markers per chromosome), to mimic the marker density

87   commonly used in breeding programs and genomic studies. Genotype data were imputed

88   leveraging a next-generation sequencing dataset based on the *Glycine max* reference

89   genome Williams 82 (*www.soybase.org*).

90   The high-yield lines (group Y) were high yield lines (group Y) was designed to

91   reflect an elite germplasm background with a historical focus on improving seed yield.

92   This group had a high base population mean for yield and a moderate mean for protein

93   content, consistent with the documented negative correlation between the two traits. The

94   high protein lines (group P) was high-protein lines (group P) were simulated from a similar

95   elite background but selected for increased seed protein concentration. Due to the

4

96 antagonistic relationship between the traits, this group exhibited a reduced yield mean.

97 Both populations shared common genetic ancestry, emulating real-world breeding

98 scenarios where selection diverges from a shared elite base.

99 *Simulation Platform and Trait Architecture*

100 Simulations were conducted using the AlphaSimR (Gaynor et al., 2021) package

101 in R software (R Core Team, 2021), a widely used platform for modeling complex trait

102 inheritance, recombination, and selection in plant breeding, according to specific

103 simulation parameters. The simulated traits included seed yield and ~~seed~~ protein content,

104 with a genetic correlation of -0.45 imposed between them to reflect the biological

105 constraint observed in empirical breeding data. Heritability values were set at 0.45 for

106 yield and 0.70 for protein. The heritability and correlation values used in this simulation

107 were derived from empirical estimates from a real-world dataset provided by the

108 University of Nebraska-Lincoln Soybean Breeding Program. This dataset comprises

109 approximately 3,000 breeding lines evaluated across 18 locations over five years (2020-

110 2024). Although the data set is unbalanced, reflecting a typical breeding pipeline's

111 structure, broad-sense heritability was estimated for yield and protein content. The

112 resulting average heritability estimates were used to parameterize the simulation,

113 ensuring that the simulated trait architecture closely reflects observed variability and

114 genetic control in elite soybean germplasm.

115 Trait values were assigned at the founder level, generating variation ~~within the~~

116 ~~founder population for groups Y and P~~for groups Y and P in the founder population. This

117 variation allowed for the initial assessment of divergence and provided a foundation for

118 within-population selection in subsequent cycles. This study assumes a fixed number of

119 QTLs and a set of genetic correlations. Future studies could extend this framework by

120 simulating different numbers of underlying QTLs and varying levels of linkage

121 disequilibrium between yield- and protein-associated loci.

122 *Experimental Procedure*

123 Step 1: Founder Population Simulation

124 The initial step involved simulating one founder population of 40 individuals per

125 year in five years, each with assigned genotypes and trait values according to the

126 simulation parameters described earlier. These populations served as the baseline for

127 evaluating genomic differentiation and trait performance.

128 Step 2: Intra-Group Recurrent Selection Cycles

129 Each simulated population underwent five generations of trait-specific selection

130 before being assigned to crosses. One population was transferred to selection in year n,

131 with their individuals redirected to crosses in year $n+1$, to ensure the continuity of the

132 breeding program. In each cycle, the top 10% of individuals were selected based on

133 phenotypic values for their respective target trait, yield for the Y group, and protein

134 content for the P group. Selected individuals were intermated within their group to

135 produce progeny for the next generation. Recombination and segregation were simulated

136 during mating, and new phenotypic values were assigned based on inherited genotypes

137 and environmental effects. This process was repeated for five cycles, allowing the

138 accumulation of selection responses and genetic changes across generations. This step is

139 usually taken when establishing breeding programs, where diverse populations with

140 minimal information about qualitative and quantitative traits, such as yield and protein

141 content in soybeans, are available. At the end of the fifth cycle, the breeder is expected to

142 have two core pools: high-yield lines and high-protein lines.

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

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

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

Where σ^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, μ 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:

189
$$GEBV_i = \sum_{j=1}^{m} Z_{ij}\hat{u}_j \text{ (Equation 3)},$$

190   where m is the total number of j markers for each i phenotypic observation. The marker

191   effects are estimated as:

192
$$\hat{u} = (Z^T Z + \sigma^2/\sigma_u^2 I)^{-1} Z^T y \text{ (Equation 4)},$$

193   where I represents an identity matrix. The ridge regression analyses were executed

194   using the rrBlup package (Endelman et al., 2011) in R software (R Core Team, 2021).

195   *Codes*

196   Codes are available in the Supplementary Material section.

197   **Results and Discussion**

198   This study investigated the genomic consequences and phenotypic benefits of

199   selection in two distinct soybean groups over five generations. Each group was initially

200   designed with identical genetic backgrounds but selected for contrasting traits, one for

201   high protein content (group P) and the other for high yield (group Y). Using fixation index

202   (F_{st}) and principal component analysis (PCA), the genetic structure and

203   differentiation of the populations were tracked over selection cycles. Mean phenotypic

204   responses were monitored to evaluate the efficiency of selection. This approach enabled

205   us to understand the magnitude and direction of genomic change resulting from selection

206   and how it correlates with phenotypic performance.

207   *Initial Genetic Structure and Differentiation (Cycle 0)*

208   Before selection was applied, the two founder groups displayed minimal genetic

209   differentiation. Chromosome-wise F_{st} values (Figure 1) were low, indicating a shared

210 genetic base between the two groups. Although these groups were designed for distinct

211 selection objectives (protein versus yield), the genetic structure had not yet diverged. PCA

212 results (Figure 2) confirmed this, with individuals from both groups overlapping in

213 multivariate space, showing no discernible clusters. This genetic similarity served as a

214 baseline, allowing the effects of subsequent selection to be clearly attributed to breeding
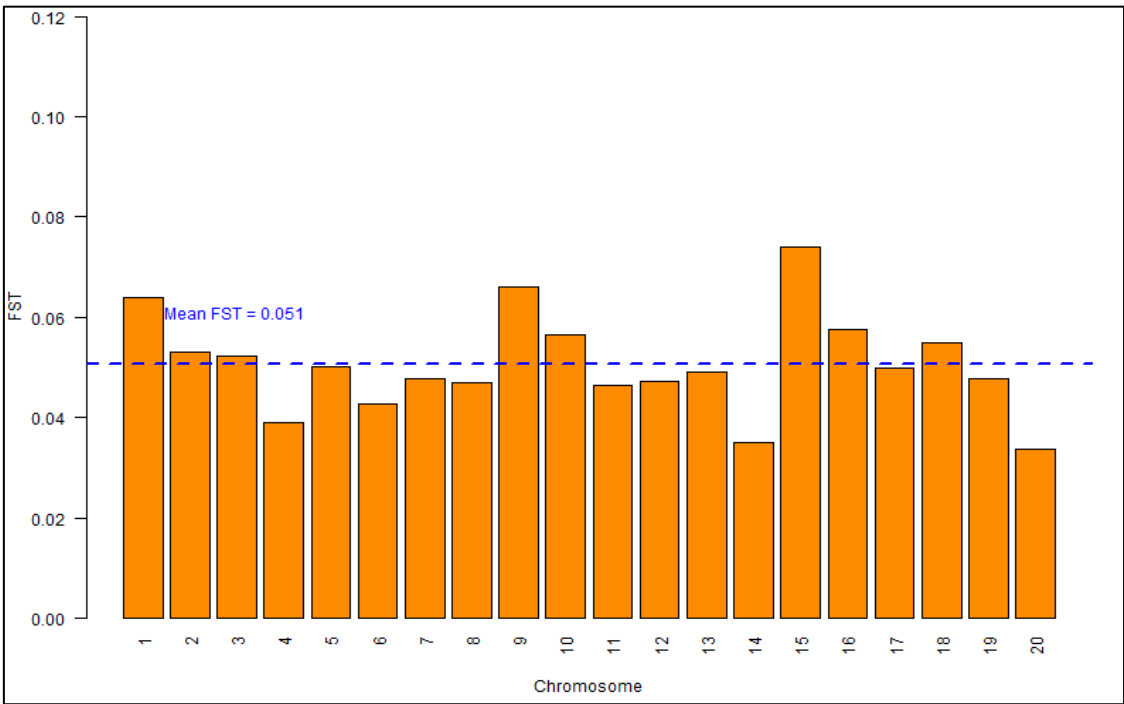
215 pressure rather than founder variation.

Figure 1. Chromosome-wise F_{st}FST between individuals of the founder population (cycle 0).
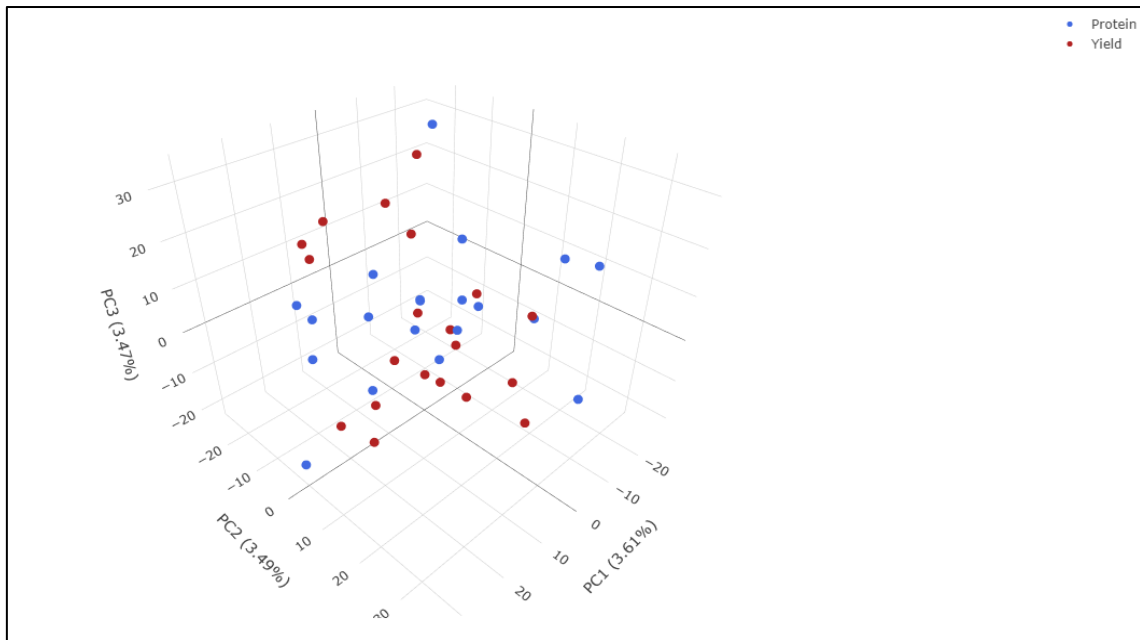
Figure 2. Principal component analyses plot (PCA) between individuals of the founder population (cycle 0).

F_{st}~~FST~~ *Across Selection Cycles*

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

11
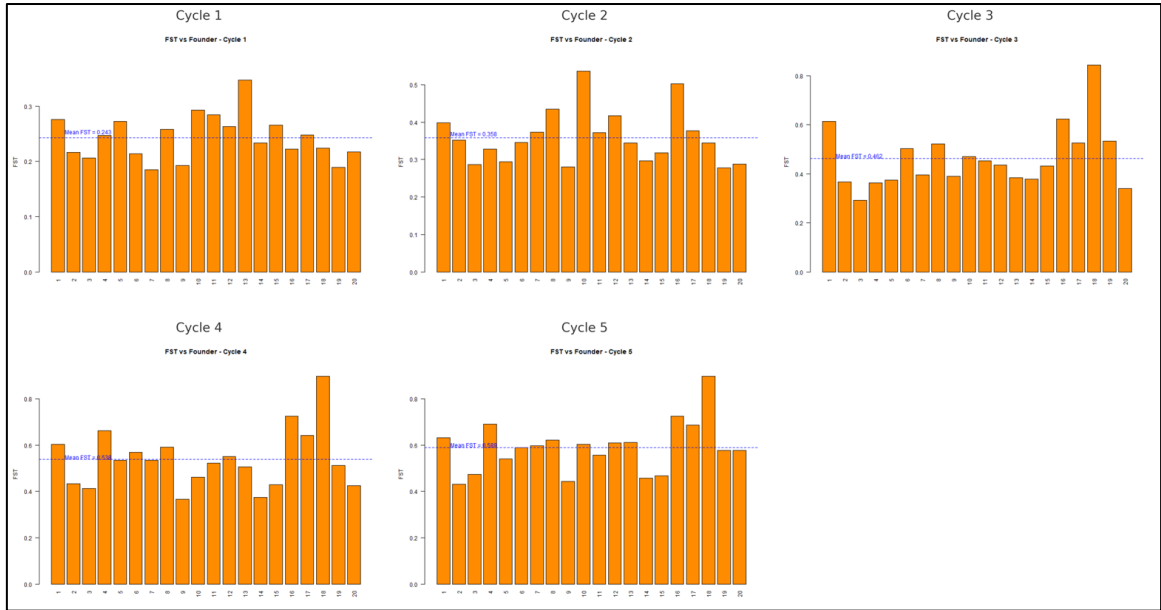
Figure 3. F_{st}F̶S̶T̶ between protein population and founder population for each selection cycle.
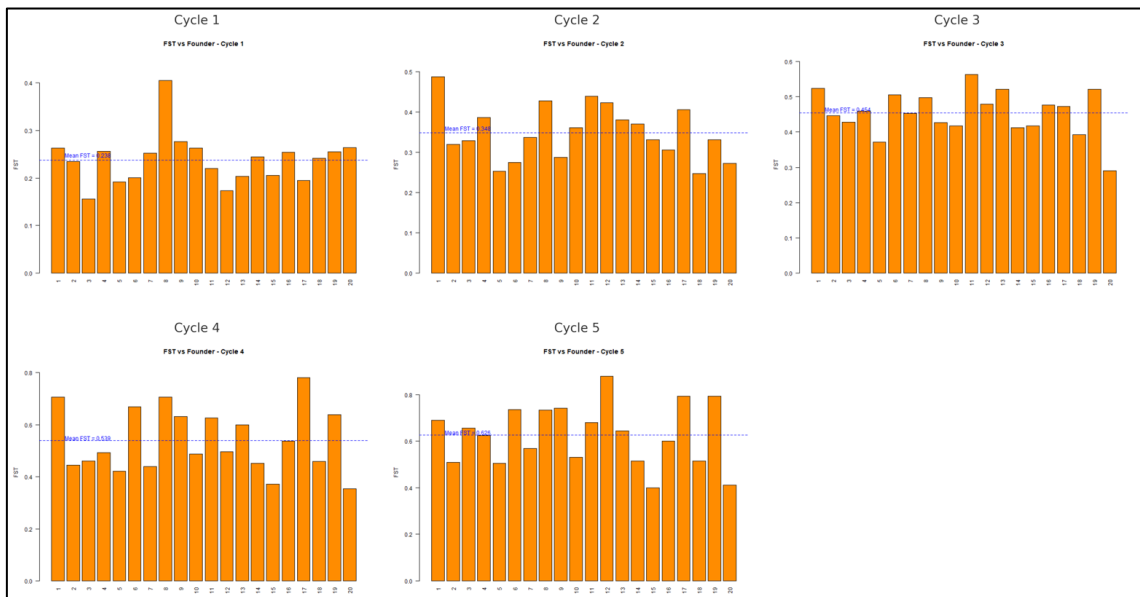
Figure 4. F_{st}F̶S̶T̶ between yield population and founder population for each selection cycle.

226   To further investigate the impact of trait-based selection, we directly compared

227   groups P and Y. Table 1 shows the interpopulation F_{st}F̶S̶T̶ for each cycle, which increased

228   from 0.277 in Cycle 1 to 0.622 by Cycle 5. This growing divergence emphasizes how

229   selection for contrasting phenotypes reshapes the genome in divergent directions. As

230 selection advanced, each population fixed or retained different allelic combinations that

231 improved their respective trait values, which cumulatively enhanced genetic separation.

232 This is particularly relevant in applied breeding programs where the trade-off between

233 protein and yield must be managed strategically.

Table 1. F_{st}~~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

234 *PCA Across Selection Cycles*

235   The genomic divergence described by F_{st}~~FST~~ was also clearly visualized using

236 PCA. In the early cycles, the overlap between groups was still visible, but from Cycle 1

237 onward, distinct clusters emerged for both protein- and yield-selected lines (Figures 5 and

238 6). In Cycle 5, PCA plots demonstrated complete separation, confirming that directional

239 selection had generated distinct genetic trajectories. This pattern was reinforced when

240 comparing the two groups at each cycle (Figure 7). These plots are consistent with our

241 hypothesis of loss of shared alleles and accumulation of beneficial mutations or

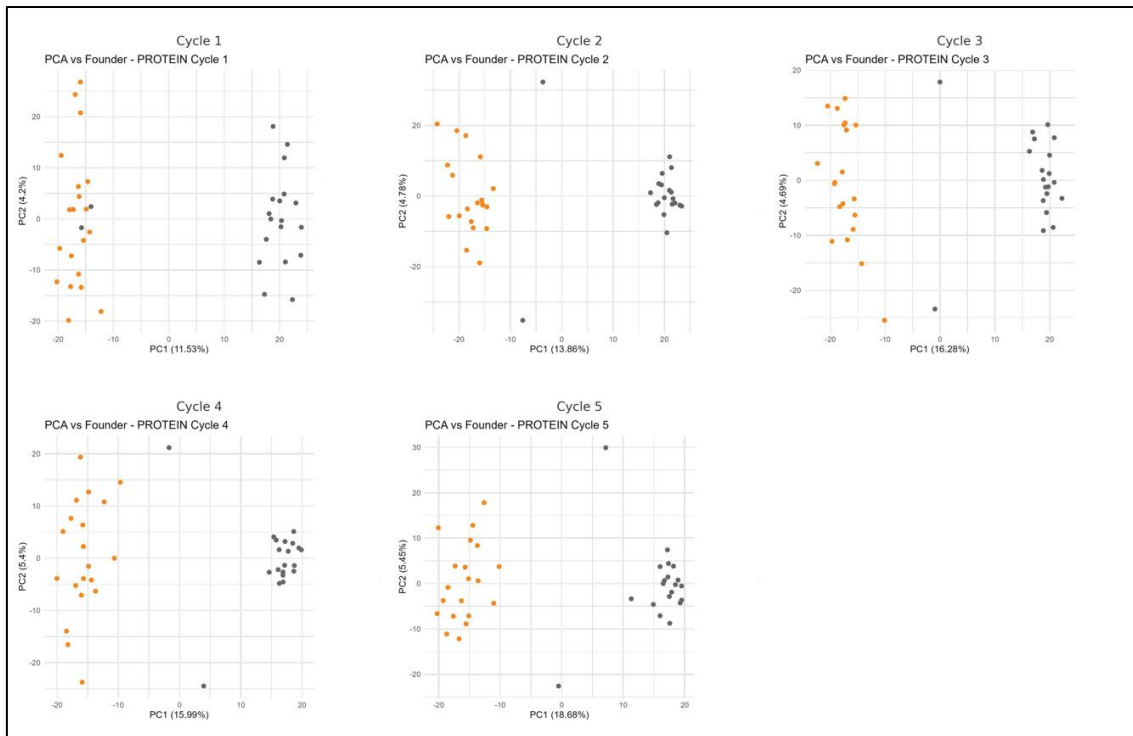242 combinations specific to the selection regime.

13

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.
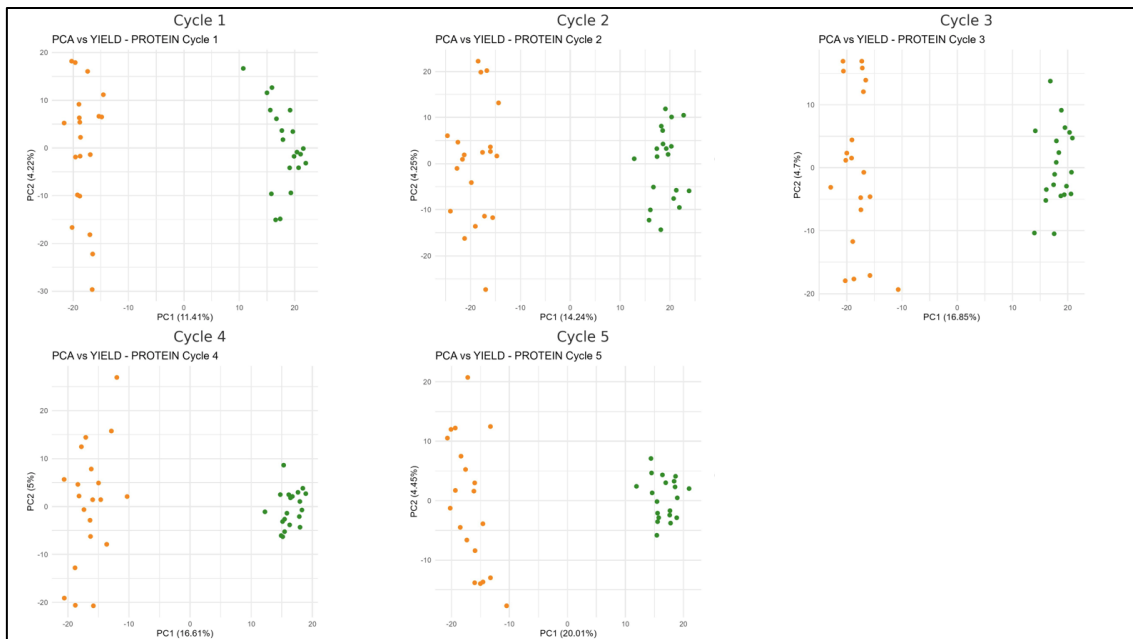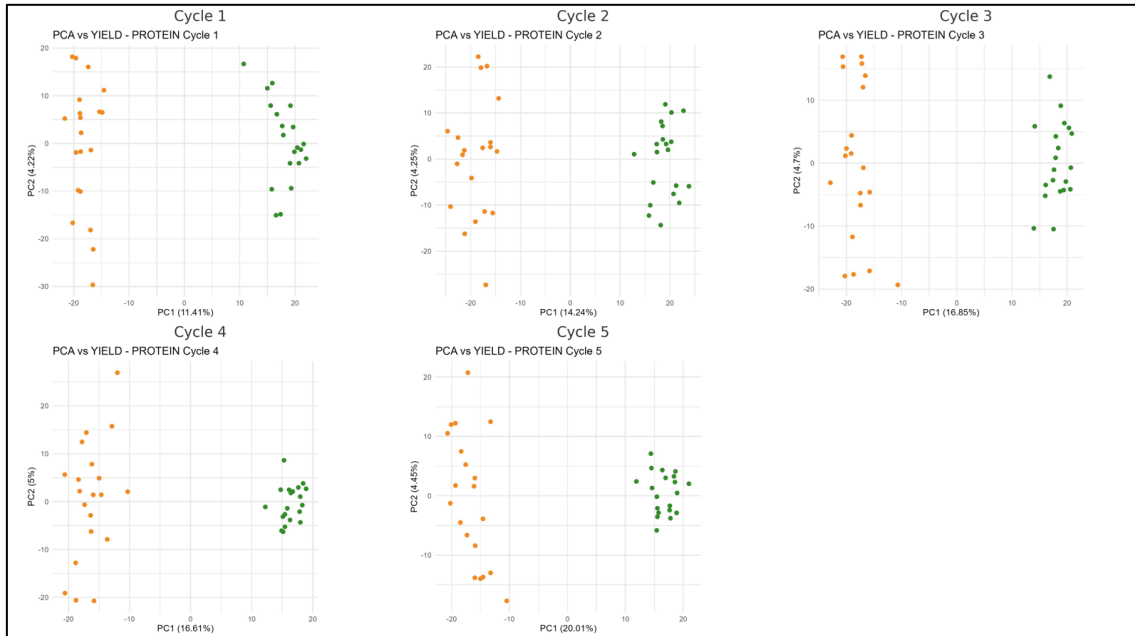

Figure 6. PCA between Y group and the Y group and the founder population for each selection cycle. Yellow dots represent individuals from the improved populations for the current cycle, and green dots represent individuals from the founder population.

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

244   Chromosome-specific F_{st}FST analysis between the P and Y groups in Cycle 5

245   (Figure 8) revealed variable patterns of divergence. Some chromosomes showed stronger

246   differentiation, potentially harboring loci with major effects for protein or yield traits.

247   This heterogeneity is expected given the polygenic nature of both traits. It also suggests

248   that despite whole-genome selection, specific genomic regions were disproportionately

249   impacted. PCA of Cycle 5 (Figure 9) supports this observation, with tight clustering

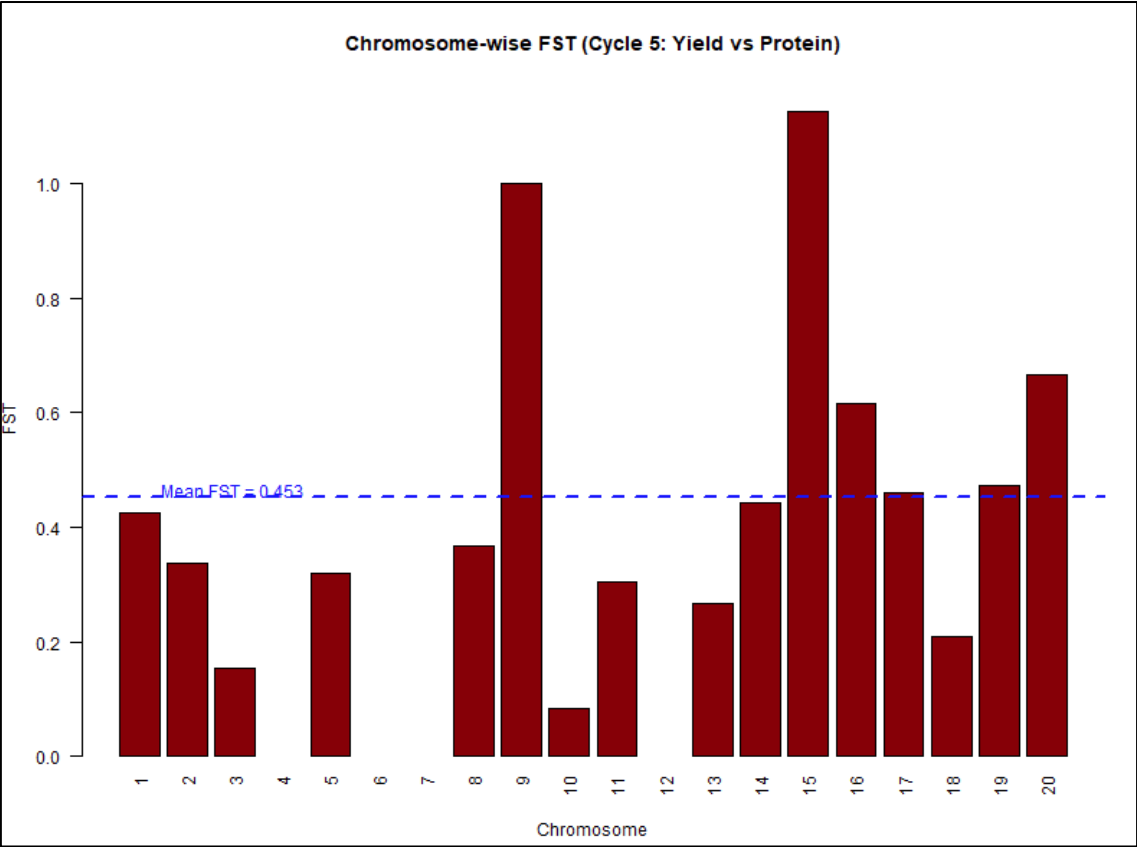250   within populations and broad separation between them.

15

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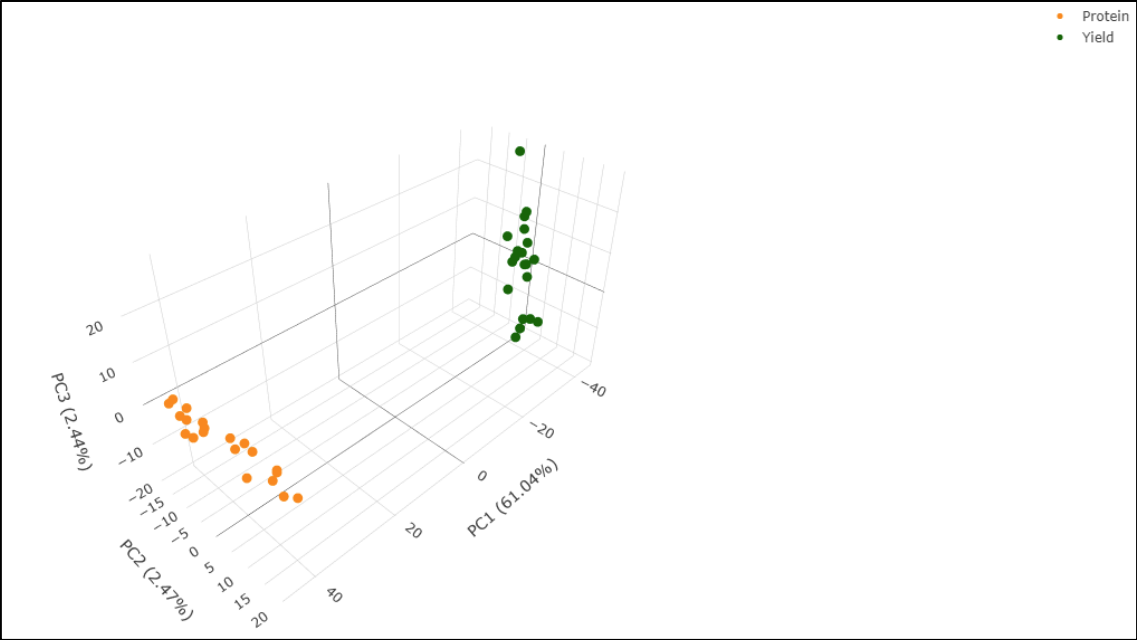

Figure 9. PCA between yield and protein population in cycle 5.

251   *F_{st}~~FST~~ Values and Changes in Phenotypes Across Cycles*

252   The consistent increase in F_{st}FST coincided with favorable shifts in mean

253   phenotypic values. For group P, average protein content rose across selection cycles

254   (Figure 10). Likewise, group P exhibited rising yield values (Figure 11), highlighting

255   selection success. It also reflects the narrowing of genetic diversity due to cycles of

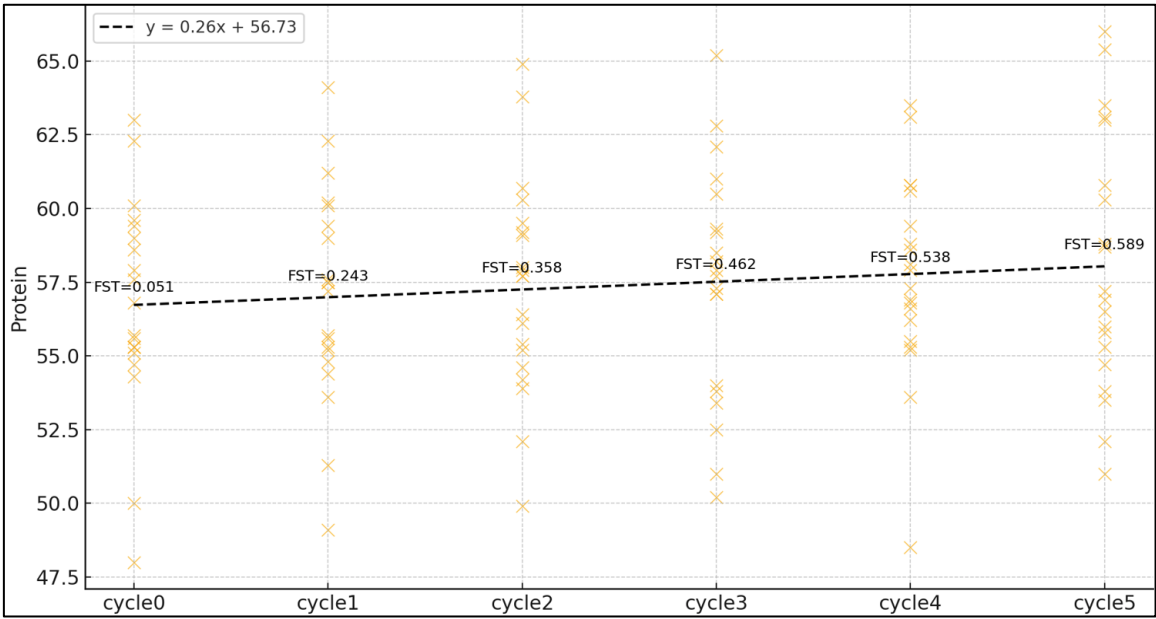256   selection.

Figure 10. Changes in F_{st}FST and protein content mean for each selection cycle for group P. Protein content (y-axis) is represented on a percentage basis.
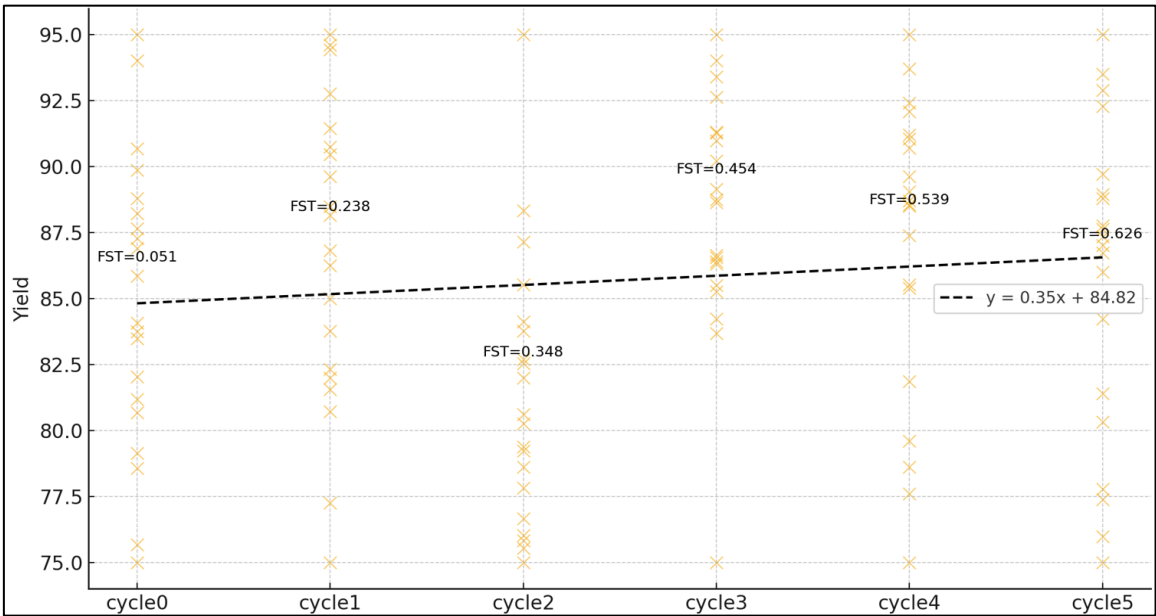
Figure 11. Changes in F_{st}FST and yield content mean for each selection cycle for group Y. Yield (y-axis) is represented on a bushels per acre basis.

257   This study demonstrates the power and consequences of selection in breeding

258 populations by combining population genomics with selection scenarios. It contributes to

259 our understanding of how elite breeding pools diverge and interact, ultimately informing

260 strategies to improve both yield and protein content in soybeans. Even starting from

261 genetically similar founders, consistent selection for distinct traits produced rapid

262 genomic divergence and substantial phenotypic gains. These outcomes reinforce the

263 utility of genomic tools such as F_{st}FST and PCA in monitoring genetic change. For

264 breeders, the results underscore the trade-offs inherent in selection: while targeted gains

265 are possible, they often come at the cost of genetic diversity. In practice, this necessitates

266 striking a balance between progress and the maintenance of long-term variability.

267 Moreover, the divergence observed here provides a rationale for developing pools based

268 on protein and yield traits.

269 *Phenotypic and genomic selection strategies*

270   Recurrent selection followed by phenotypic or genomic selection is important to

271 boost gains in breeding programs. Our study simulated five recurrent selection cycles and

272 ten consecutive phenotypic or genomic selection cycles on protein and yield populations.

273   For the phenotypic selection, yield showed an asymptotic increase in gains

274 through cycles. These increases were more evident between cycles 1-6, but a plateau was

275 reached between cycles 7-10 (Figure 12). This suggests that the diminishing gains result

276 from finite genetic variance because of selection (Falconer & Mackay, 1996). The initial

277 rapid response to selection was due to the high frequency of favorable alleles. However,

278 these alleles were fixed across initial cycles, which led to a reduction in the variability
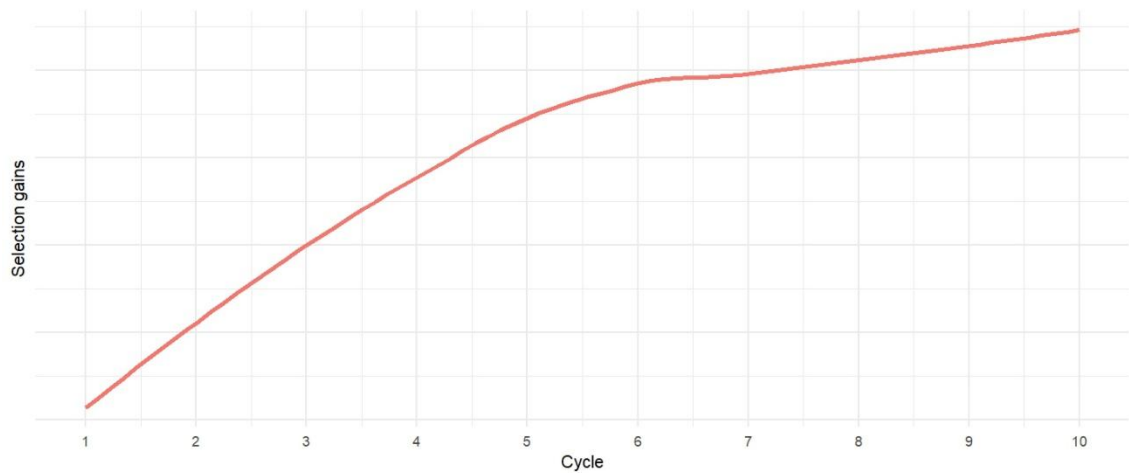
279 and consequently lower gains.

18

280

281

282

283

284

285 The results for phenotypic selection of protein for group P also show cumulative positive

286 gains across cycles of selection but little cumulative gain after cycle 6 (Figure 13). This

287 is similar to what was displayed for phenotypic selection on yield, where the variance

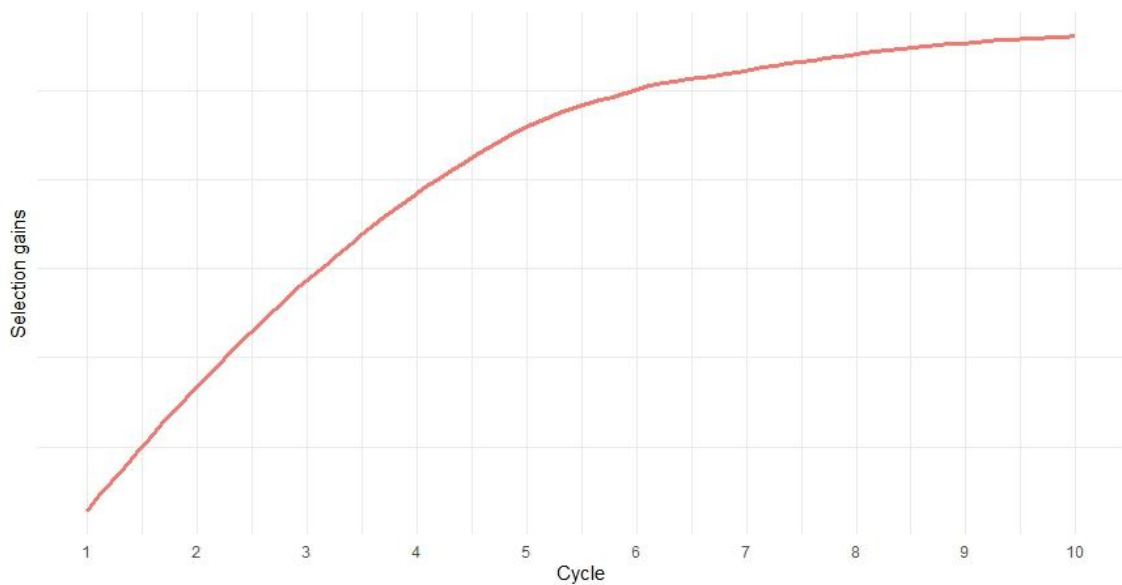288 was exhausted over time when no new genetic variability was introduced to the breeding

289 pipeline.

290

292   Opposed to what was shown for phenotypic selection for yield and protein, the

293   genomic selection showed initial decreases in gain in initial cycles for both traits (protein

294   in group P and yield in group Y). In both cases, the initial decrease in gains was followed

295   by a sharp increase; then a plateau was reached. For protein, the initial drop was smoother,

296   and the increase in gains initiated in cycle five and continued until cycle seven, where it

297   peaked. However, it decreased rapidly in the following cycles (Figure 14). For yield,

298   despite the more evident initial decrease in gains, genomic selection remained positive

299   for a more extended period, from cycle three until cycle 7, followed by a sharp decrease
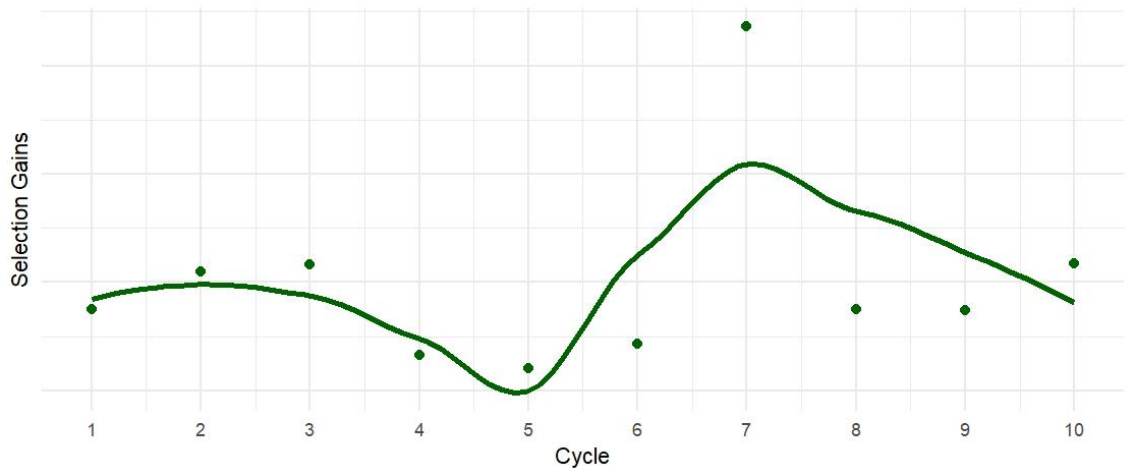
300   in subsequent cycles (Figure 15).

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.
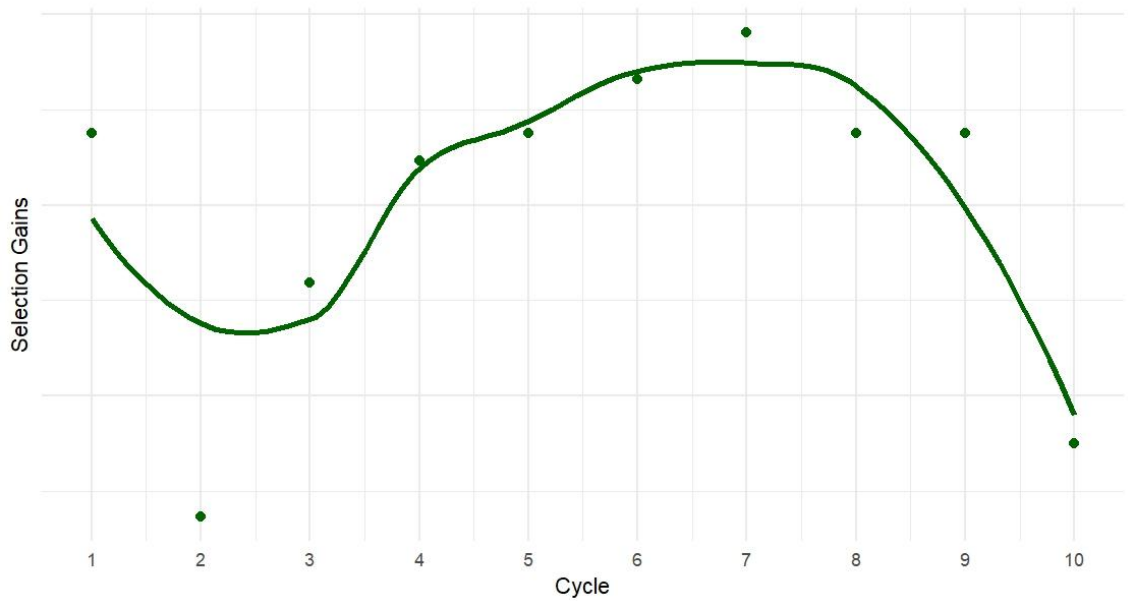
21

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

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

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

Future directionsConclusions

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.

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

342 correlation between yield and protein content, it is also expected that the mean phenotypic

343 values for both traits will decline compared to those observed in Cycle 5.

344 **References**

345 Akdemir, D., Sanchez, J. I., & Jannink, J. L. (2015). Optimization of genomic selection

346   training populations with a genetic algorithm. *Genetics Selection Evolution*, *47*,

347   1-10.

348 Andrijanić, Z., Nazzicari, N., Šarčević, H., Sudarić, A., Annicchiarico, P., & Pejić, I.

349   (2023). Genetic diversity and population structure of European soybean

350   germplasm revealed by single nucleotide polymorphism. *Plants*, *12*(9), 1837.

351 Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., ... &

352   Robbins, K. R. (2021a). Maximizing efficiency of genomic selection in

353   CIMMYT's tropical maize breeding program. *Theoretical and Applied

354   Genetics, 134*, 279-294.

355 Atanda, S. A., Olsen, M., Crossa, J., Burgueño, J., Rincent, R., Dzidzienyo, D., ... &

356   Robbins, K. R. (2021b). Scalable sparse testing genomic selection strategy for

357   early yield testing stage. *Frontiers in Plant Science*, *12*, 658978.

358 Bandillo, N. B., Jarquin, D., Posadas, L. G., Lorenz, A. J., & Graef, G. L. (2023).

359   Genomic selection performs as effectively as phenotypic selection for increasing

360   seed yield in soybean. *The Plant Genome*, *16*(1), e20285.

361 Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R

362   package rrBLUP. *The plant genome*, *4*(3).

363 Gaynor RC, Gorjanc G, Hickey JM (2021). "AlphaSimR: an R package for breeding

364 program simulations." G3 Gene|Genomes|Genetics, 11(jkaa07).

365 https://doi.org/10.1093/g3journal/jkaa017.

366 Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long

367 term response. *Genetica*, *136*(2), 245-257.

368 Jannink, J. L. (2010). Dynamics of long-term genomic selection. *Genetics Selection*

369 *Evolution*, *42*, 1-11.

370 R Core Team (2021). R: A language and environment for statistical computing. R

371 Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-

372 project.org/.

373 Sedivy, E. J., Wu, F., & Hanzawa, Y. (2017). Soybean domestication: the origin, genetic

374 architecture and molecular bases. *New Phytologist*, *214*(2), 539-553.

375 Silva, A. C. D., Gregorio da Silva, D. C., Ferreira, E. G. C., Abdelnoor, R. V., Borém, A.,

376 Arias, C. A., ... & Marcelino-Guimarães, F. C. (2025). Genetic diversity,

377 population structure in a historical panel of Brazilian soybean cultivars. *PloS*

378 *one*, *20*(1), e0313151.

379 US Department of Agriculture Foreign Agricultural Service. (2017). Oilseeds: World

380 Production Markets and Trade Reports (Washington, DC: FAS).

381 Yang, C., Yan, J., Jiang, S., Li, X., Min, H., Wang, X., & Hao, D. (2022). Resequencing

382 250 soybean accessions: new insights into genes associated with agronomic traits

383 and genetic networks. *Genomics, Proteomics & Bioinformatics*, *20*(1), 29-41.

384   Zhang, J., Song, Q., Cregan, P. B., & Jiang, G. L. (2016). Genome-wide association study,

385   genomic prediction and marker-assisted selection for seed weight in soybean

386   (Glycine max). Theoretical and Applied Genetics, 129, 117-130.

387   Zhang, J., Wang, X., Lu, Y., Bhusal, S. J., Song, Q., Cregan, P. B., ... & Jiang, G. L.

388   (2018). Genome-wide scan for seed composition provides insights into soybean

389   quality improvement and the impacts of domestication and breeding. Molecular

390   Plant, 11(3), 460-472.