Statement to reviewers:

Dear reviewers -

Your comments were addressed, and the changes were performed accordingly. You can follow the corrected content in the version where the track changes tool was enabled. There is also a second version of the manuscript where all the content (updates for the final exam and corrections) is not highlighted to facilitate reading.

Find below the answers to your comments.

Looking forward to hearing from you about the manuscript.

Best regards,

Arthur Bernardeli.

Answers to reviewers' questions:

Reviewer 1

1. I agree with the reviewer that the introduction is thoughtful and well-placed. The three objectives are clearly defined, and the literature review is well done.

   Dear reviewer, thank you for the positive comment on the introduction and the literature review.

2. There are still some typos. For example, in line 66, 'these population' should be 'these populations.' As you revise your paper, please be sure to proofread carefully and correct all typos and grammatical errors

   Dear reviewer, the typos that were found were corrected throughout the whole paper.

3. In lines 68–69, evenly spaced SNPs can be a good starting point. However, to make the simulation more realistic, why not use real-world SNP distribution data for your second round of simulations?

   Dear reviewer, thank you for your constructive suggestion. We agree that incorporating real-world SNP distributions can improve the simulation. However, we found that many of these resources (such as SoySNP50K) lack associated phenotypic and pedigree data. These two features were included in our hypothesis and questions for developing elite populations for yield and protein. In a future approach, we plan to incorporate a real SNP dataset in our analyses, although it might impact our hypotheses and assumptions.

4. In lines 88–89, consider expanding the discussion on the heritability of the two traits. For example, explain under what mating schemes and field experimental designs yield heritability was estimated in different studies. Are these values consistent across studies, or do they vary? Clarify why you chose one estimate over the others.

   Dear reviewer, these values for trait heritability were chosen from a real breeding dataset from the University of Nebraska-Lincoln Soybean Breeding Program. The dataset contains approximately 3000 lines tested in 18 locations across five years (2020-2024). Although the dataset is unbalanced, representing the breeding pipeline, heritability was estimated for yield and protein, and averaged to serve as input for our simulation study. Now, this is detailed in the materials and methods section.

5. In Materials and Methods section, considering to reorganize them into two several subsections “Population simulation procedure” and “Population genetics statistical estimation”, “Statistical Methods and Codes”, etc.

Dear reviewer, the Materials and Methods section was reorganized. Now it shows the following structure:

- **Material and Methods**
  - *Population Simulation and Structure*
  - *Simulation Platform and Trait Architecture*
  - *Experimental Procedure*
    - Step 1: Founder Population Simulation
    - Step 2: Intra-Group Recurrent Selection Cycles
    - Step 3: FST, PCA, and phenotypic analysis
    - Step 4: Selection scenarios
  - *Codes*

6. What does pairwise FST mean in your context? If it is not a standard Fst statistic, you should define it clearly in the Material and Methods and provide an interpretation. Why not use heterozygosity or theta values we learned in the class?

   Dear reviewer, we used FST in this study to quantify genetic differentiation between groups under divergent selection (e.g., yield vs. protein) and between each selected group and its founder population. We chose FST because it directly compares the divergence between specific population pairs at each selection cycle. Our initial focus was on quantifying intra- and between-population divergence caused by directional selection. Therefore, pairwise FST was the most relevant metric for our objectives. FST formula is included in the Materials and Methods section now.

7. The PCA figures are not very informative. I assume you're trying to demonstrate that, over several generations of divergent selection, the populations become

increasingly differentiated. You could illustrate this more clearly with a single PCA plot using color coding to represent generations 0 through 5.

Dear reviewer, we initially attempted to merge all the PC plots into one; however, even using transparent filling for each genotype (scores), it was challenging to visualize each cycle separately due to overlapping. Therefore, we decided to maintain it separately.

8. Finally, it's not clear to me what you expect to learn from this simulation. Perhaps, in your revision, you want to redefine the genetic architecture underlying the two traits (protein and yield) by simulating different numbers of QTLs and varying levels of linkage between them. If so, by tracking changes over several generations of selection, you might gain insight into the potential causes of the antagonistic relationship between the two traits.

Dear reviewer, this simulation aimed to evaluate how divergent selection for yield and protein, starting from a shared founder background, can lead to genomic divergence (FST), changes in genetic structure (PCA), and phenotypic gains over time under controlled selection scenarios (genomic and phenotypic selection). Our current simulation assumes a fixed polygenic architecture for both traits with a known negative genetic correlation, reflecting observations in empirical soybean breeding data. Importantly, the simulation pipeline was designed to closely mimic the structure and operations of a real-world soybean breeding program, incorporating recurrent selection, selection intensity, trait heritability, the timing of intermating and advancement cycles, and genomic and phenotypic selections. This is essential and can be interpreted in an applied breeding context. We added in the Materials and Methods section, under the subitem Simulation Platform and

Trait Architecture, our reason for choosing an even number of QTLs (SNP markers).

Reviewer 2

1. This manuscript from Arthur Bernardeli addresses the key issues in the soybean breeding program. Starting with the importance of soybeans and the economic and nutritional value of the crop, it explained the key challenge between yield and seed protein content in soybeans, which are negatively correlated with each other. Improving one trait often leads to the detriment of another, and the aim is to optimize both traits. It explains the poorly understood selection dynamics that maintain or resolve the yield protein trade-off. This study used a simulation-based approach to understand how selection influences genetic divergence by leveraging population genomic metrics like Fst to monitor genetic and phenotypic changes. The three main objectives of this manuscript are to evaluate the genetic differentiation over several election cycles in populations selected for yield and high protein content. How repeated selection cycles for high yield and high protein affect genetic and phenotypic changes. Explore recombination to break the negative correlation between yield and protein content. The preliminary results show that selection divergence creates genetic divergence with validation using phenotypic data. This manuscript aligned with course interest by implementing the population genomic metrics with a selection study to optimize soybeans' yield and protein content. The literature review also supports the study is in the perfect place. However, I have some suggestions to make this work more impactful:

Dear reviewer, thank you for the comments about the scope of our research. The corrections you suggested were addressed accordingly in the manuscript file, and the answers to your comments are included below.

2. In your future report, explaining the specific regions of the genome identified by Fst would be better, and any known QTL identified in those regions. FST must be written as Fst in the formal publication manuscript. I see a clear trend in the Fst and PCA with the hypothesis, but the significance of this difference can be explained more clearly in the future.

   Dear reviewer, we agree that integrating known QTL information with regions showing elevated Fst values would enhance the biological interpretation of the genomic divergence observed. In this initial study, our first goal was to quantify genome-wide divergence patterns over selection cycles and how they could influence gains using distinct selection scenarios, without identifying Fst peaks and overlapping them with known QTLs. We also changed the nomenclature for Fst throughout the revised manuscript.

3. The literature review would help make this study more relevant if you could explain the origin and domestication of soybeans.

   Dear reviewer, this was briefly added at the beginning of the first paragraph.

4. Figures are not self-explanatory. The legends and the axis labels are not clear and could be modified. The titles within the figures are also not clearly explained. Consistency for the groups P and Y (protein and yield) can be maintained (hard to understand at my initial reading through the entire manuscript), for example, Fst vs Founder in figures 3 and 4.

   Dear reviewer, the original figures contained bigger legends and labels since they were generated individually. The figures plotted in the manuscript were after

merging all of them. We decided to merge them to prevent the manuscript being fully populated by numerous figures. We are sorry for that. For publication purposes, we can provide the original individual figures so its dimensios can be reshaped to meet the journal standards. Regarding the “Fst vs founder” comment, this represents the differentiation in terms of Fst measured between the population in the current cycle and the founder population. We corrected the descriptions below each figure in the manuscript.

5. In Figure 5, the wording cannot be easily understood. The meaning of colors (2 clusters) is hard to understand. Similarly, in Figure 6, the clusters do not have clear names.

   Dear reviewer, thank you for pointing this out. Our code to generate plots was incorrect, and colors were not linked to populations. Fin figure 5, yellow dots represent individuals from the improved populations for the current cycle, and gray dots represent individuals from the founder population. In figure 6, yellow dots represent individuals from the improved populations for the current cycle, and green dots represent individuals from the founder population. In figure 7, yellow dots represent individuals from the P group, and green dots represent individuals from the Y group. This was corrected in the manuscript.