MDPI

Article

Identification of high-yielding soybean lines with exceptional seed composition qualities

Jay Gillenwater[1,‡,*]

[1] Department of Crop and Soil Sciences, NCSU, Raleigh, NC, USA; jhgille2@ncsu.edu
* Correspondence: leutnant@fh-muenster.de; Tel.: +XX-000-00-0000.
† Current address: Updated affiliation
‡ These authors contributed equally to this work.

Version April 20, 2022 submitted to Journal Not Specified

Simple Summary: A Simple summary goes here.

Abstract: A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: 1) Background: Place the question addressed in a broad context and highlight the purpose of the study; 2) Methods: Describe briefly the main methods or treatments applied; 3) Results: Summarize the article's main findings; and 4) Conclusion: Indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

Keywords: yield; protein; oil;soybean; protein meal

1. Version

This Rmd-skeleton uses the mdpi Latex template published 2019/02. However, the official template gets more frequently updated than the 'rticles' package. Therefore, please make sure prior to paper submission, that you're using the most recent .cls, .tex and .bst files (available here).

2. Introduction

Seed yield, oil, and protein are all valuable traits in a soybean variety, however breeding lines which have both high yield and protein has been difficult to develop due to the negative correlation between the two traits[1–3]. While considerable efforts have been made to identify loci which control these seed quality traits so that MAS breeding strategies can be utilized for their improvement, to date the applications of such markers have been few. This is largely due to the lack of markers which are uniquely associated with one trait, and are also stable across genetic and environmental backgrounds. While there is still reason to continue this genetic research, it is important that breeders take every opportunity to identify lines with both high yield, and seed composition traits like oil and protein content so that new varieties can be released.

Soybean lines typically contain about 20% oil and 40% protein content on a dry weight basis[4]. The market for soybean meal requires 47.5% protein content in the meal, which corresponds to approximately 41.5% protein content on a dry weight basis[4]. Oil and protein content are two of the most important seed composition traits in soybean so if one is decreased, the other should be correspondingly increased to account for the loss in value. The inverse correlation between protein

and oil contents is well known and is suspected to be due at least partially to the action of pleiotropic genes and competing metabolic pathways which control the expression of each trait[5].

Despite the difficulty in simultaneously breeding for all three of these traits, releases of varieties with elevated protein contents and only moderately reduced yield has shown that it is not impossible. The high protein germplasm lines R05-1415 and R05-1772 were released recently and contain 46.9% and 46.1% protein content with while still producing yields 94% and 91% of that of the high yielding 5002T cultivar[6]. Lines TN03-350 and TN04-5321 contain 43.9% and 43.1% protein content while having superior or comparable performance to yield checks[7]. The Prolina cultivar has a protein content of 46.1% with a yield only 13% reduced from the Centennial check[8]. The Highpro1 cultivar was released in 2016 and has a yield which is greater than or equal to 97% of that of the highest yielding check cultivar, IA3023 with a protein content of 40.1%[9]. Cultivars produced through conventional breeding techniques such as these have shown that it is possible to identify lines with both high seed protein and seed yield. Efforts to find these lines should be continued to provide growers and breeders with additional high value cultivars, and germplasm which can be further used to improve protein and yield traits.

To meet this goal, two recombinant inbred line (RIL) oil mapping populations were screened for lines which showed promising combinations of yield and seed composition traits. Successive rounds of selection were conducted to identify and characterize lines with high values for yield as well as protein and oil composition were performed between 2018 and 2021 to identify soybean lines with elevated yield as well as the valuable seed composition traits when compared with existing check cultivars.

3. Materials and Methods

3.1. Population development

In 2018, oil mapping populations 201 and 202 were grown as plant rows at the Central Crops Research Station in Clayton, NC. These populations consisted of 273 and 237 recombinant inbred lines (RILs) respectively. Several agronomic traits were scored in the field for each population.

The agronomic traits recorded in the field were height, lodging, maturity date, and a composite agronomic score. Lodging was scored on a scale of 1-5 where 5 indicates that all plants in a plot are on the ground, and a score of 1 indicates that all plants are erect[10]. The agronomic score aimed to capture other traits of value such as visual estimation of pod load and plot uniformity to provide a general score of a line's agronomic desirability. Agronomic score was recorded on a scale of 1-5 as well, with 1 identifying the best lines of a population, and 5 the worst. Maturity was recorded at the R8 maturity date and was recorded as the number of days after September 1. Height was measured in inches from the soil to the top of the plant.

Following harvest, yield, seed weight, protein, and oil content were measured after seed was air dried to approximately 7% moisture content in a greenhouse. Protein and oil contents were measured on a dry basis using a Perten DA 7250 NIR®instrument. Yield and seed weight were measured after seed had been sifted and cleaned of debris and cracked seed.

To select lines for the 2019 growing season, lines with abnormally low bulk weights or extreme maturity dates from 2018 were first removed from consideration. Two yield trials were then developed for each mapping population. The maturity dates of RILs were considered when forming tests such that the lines of each test would have a maturity date range approximately half that of the total mapping population from which it was derived. RILs were selected for each test which were also representative of the distribution of seed protein and seed oil traits for each population.

Eighty unique lines were selected from each population which satisfied these criteria, and each yield test was comprised of 40 RILs. Three high-yielding check cultivars and the two parents of the respective population were also included in each test. Yield check cultivars Dunphy, Osage[11], and Roy were used in tests 1 and 2, while Dunphy, Dilday, and N.C. Raleigh were used for tests 3 and 4.

These lines were selected to represent the estimated maturities of the RILs in each test. The parents for tests 1 and 2 were cultivars LMN09-119 and N09-09, and the parents for tests 3 and 4 were LMN09-19 and N13-47.

These four tests were grown in two locations in 2019: the Tidewater Research Station in Plymouth, NC (PLY) and the Caswell Research Farm in Kinston, NC (CAS). The same data was collected for each test in this season that was collected in the previous season.

Using the data follected from this season, further selections were done to identify high-yielding lines from the four tests. This was done by identifying the RILs with a yield within or above a least significant difference (LSD) of the average yield of the checks for each test. Further selection was done using the seed composition traits by identifying the thirty RILs with the highest protein + oil content on a dry basis from among the RILs which had passed the yield selection threshold.

These thirty lines were then grouped into two new tests of 15 RILs each on the basis of maturity date. These two new tests are named Test 1 and Test 2. Yield check cultivars were again assigned to each test to match the maturity dates of the RILs that were in each test. Cultivars Dunphy, DIlday, and NC-Raleigh were used as checks in Test 1 and Dunphy, Ellis, N10-697, and Osage were used as checks in Test 2.

These two tests were grown in both the 2020 and 2021 seasons. These tests were grown in CLA and CAS in 2020 and CAS and PLY in 2021. The same phenotypes were evaluated for each genotype in the 2020 and 2021 seasons using the same methodology that was employed in the 2019 season.

3.2. Statistical Analysis

Phenotypic traits were analysed with a linear model with the form:

$$y_{ijk} = \mu + L_i + B(L_i) + G_k + GL_{ik} + \epsilon_{ijk}$$

Where y_{ijk} is the phenotypic measurement for rep$j$ of genotype k in environment i, L_i is the effect of location i, G_k is the effect of genotype G, GL_{ik} is the interaction effect of location L and genotype G, and ϵ_{ijk} is the measurement error.

Models were fit using the lm function in R[12], and analysis of variance (ANOVA) performed using the anova function. Least-square means (LS Means) for each genotype and trait were calculated using the above model using the emmeans package[13] in R. Least-significant difference values for each trait were calculated using the LSD.test function from the agricolae[14] package. Pearson correlation coefficients between each phenotype were calculated with the cor function in R.

Pearson correlation is calculated for each pair of traits as:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

Where x and y are measurements of the two phenotypes, m_x and m_y are the means of each phenotype, and r is the correlation coefficient.

Materials and Methods should be described with sufficient details to allow others to replicate and build on published results. Please note that publication of your manuscript implicates that you must make all materials, data, computer code, and protocols associated with the publication available to readers. Please disclose at the submission stage any restrictions on the availability of materials or information. New methods and protocols should be described in detail while well-established methods can be briefly described and appropriately cited.

Research manuscripts reporting large datasets that are deposited in a publicly available database should specify where the data have been deposited and provide the relevant accession numbers. If the accession numbers have not yet been obtained at the time of submission, please state that they will be provided during review. They must be provided prior to publication.

Interventionary studies involving animals or humans, and other studies require ethical approval must list the authority that provided approval and the corresponding ethical approval code.

4. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

4.1. Subsection Heading Here

Subsection text here.

4.1.1. Subsubsection Heading Here

Bulleted lists look like this:

- First bullet
- Second bullet
- Third bullet

Numbered lists can be added as follows:

1. First item
2. Second item
3. Third item

The text continues here.
All figures and tables should be cited in the main text as Figure 1, Table 1, etc.

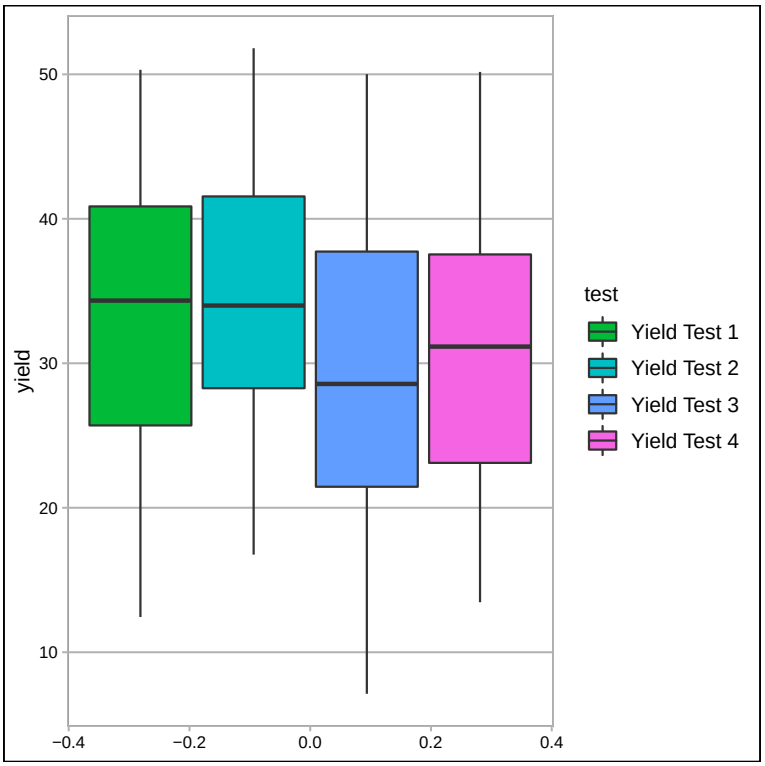

Figure 1. This is a figure, Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

Table 2. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Figure 2. This is a figure, Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

Table 1. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Title 1	Title 2	Title 3
entry 1	data	data
entry 2	data	data

This is an example of an equation:

$$\mathbb{S} \tag{1}$$

Example of a theorem:

Theorem 1. *Example text of a theorem.*

The text continues here. Proofs must be formatted as follows:
Example of a proof:

Proof of Theorem 1. Text of the proof. Note that the phrase 'of Theorem 1' is optional if it is clear which theorem is being referred to. □

The text continues here.

5. Discussion

Authors should discuss the results and how they can be interpreted in perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

6. Conclusion

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

7. Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

Acknowledgments: All sources of funding of the study should be disclosed. Please clearly indicate grants that you have received in support of your research work. Clearly state if you received funds for covering the costs to publish in open access.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "X.X. and Y.Y. conceive and designed the experiments; X.X. performed the experiments; X.X. and Y.Y. analyzed the data; W.W. contributed reagents/materials/analysis tools; Y.Y. wrote the paper.'' Authorship must be limited to those who have contributed substantially to the work reported.

Conflicts of Interest: Declare conflicts of interest or state 'The authors declare no conflict of interest.' Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funding sponsors in the design of the study; in the collection, analyses or interpretation of data in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state 'The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, an in the decision to publish the results'.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	linear dichroism

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text. For example, explanations of experimental details that would disrupt the flow of the main text, but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

Appendix B

All appendix sections must be cited in the main text. In the appendixes, Figures, Tables, etc. should be labeled starting with 'A', e.g., Figure A1, Figure A2, etc.

References

1. Burton, J. Quantitative Genetics: Results Relevant to Soybean Breeding.
2. Chaudhary, J.; Patil, G.B.; Sonah, H.; Deshmukh, R.K.; Vuong, T.D.; Valliyodan, B.; Nguyen, H.T. Expanding Omics Resources for Improvement of Soybean Seed Composition Traits. *6*, 1021.
3. Hartwig, E.E.; Hinson, K. Association between Chemical Composition of Seed and Seed Yield of Soybeans 1. *12*, 829–830.
4. Patil, G.; Mian, R.; Vuong, T.; Pantalone, V.; Song, Q.; Chen, P.; Shannon, G.J.; Carter, T.C.; Nguyen, H.T. Molecular Mapping and Genomics of Soybean Seed Protein: A Review and Perspective for the Future. *130*, 1975–1991.
5. Gupta, M.; Bhaskar, P.B.; Sriram, S.; Wang, P.H. Integration of Omics Approaches to Understand Oil/Protein Content during Seed Development in Oilseed Crops. *36*, 637–652. doi:10.1007/s00299-016-2064-1.

6. Chen, P.; Ishibashi, T.; Dombek, D.; Rupe, J. Registration of R05-1415 and R05-1772 High-Protein Soybean Germplasm Lines. *5*, 410–413.
7. Panthee, D.; Pantalone, V. Registration of Soybean Germplasm Lines TN03-350 and TN04-5321 with Improved Protein Concentration and Quality. *46*, 2328.
8. Burton, J.; WILSON, R.; others. Registration ofProlina'Soybean. *39*, 294–295.
9. Mian, M.R.; McHale, L.; Li, Z.; Dorrance, A.E. Registration of 'Highpro1'soybean with high protein and high yield developed from a North x South cross. *Journal of Plant Registrations* **2017**, *11*, 51–54.
10. Fehr, I. Soybeans: Improvement, Production and Uses.
11. Chen, P.; Sneller, C.; Mozzoni, L.; Rupe, J. Registration of Osage Soybean. *1*, 89–92.
12. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
13. Lenth, R.V. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*.
14. De Mendiburu, F. Agricolae: Statistical Procedures for Agricultural Research. *1*, 1–4.

Sample Availability: Samples of the compounds ..... are available from the authors.