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:
† Current address: Updated affiliation
‡ These authors contributed equally to this work.

Version June 5, 2022 submitted to Journal Not Specified

check for updates

Simple Summary: A Simple summary goes here.

Abstract: In current markets, the primary uses for soybean seed are in products derived from it's oil or protein content. However, growers are compensated based on seed oil so a more valuable crop is one with both high yield, and high protein, oil, or a combination of the two. A negative correlation between seed yield and seed protein makes improving these traits simultaneously difficult but not impossible through conventional breeding. A selection of lines with exceptional yield and seed composition qualities was made from two recombinant inbred line (RIL) soybean mapping populations to identify soybean varieties with yield comparable to existing cultivars, and seed composition traits superior to existing cultivars. The performance of these RILs were evaluated in multiple environments and several genotypes were identified with yield comparable to that of existing high-performance cultivars with seed protein and or oil that is equal to or superior to the high-performing cultivars. These genotypes will provide breeders with additional sources of germplasm for continuing efforts to improve seed composition traits, and provide growers with valuable genotype options.

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

1. 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 yield comparable to existing check cultivars and protein and or oil composition superior to that of the check cultivars.

2. Materials and Methods

2.1. 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. These yield tests were named test 1 and test 2 for RILs derived from mapping population 201 and test 3 and 4 for RILs derived from mapping population

202. Yield check cultivars Dunphy, Osage[11], and Roy were used in tests 1 and 2, while Dunphy, Dilday, and NC-Raleigh[12] 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.

2.2. Experimental design

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 collected from the 2019 season 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 based on 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. RILs were grown in a randomized complete block design with four replications in each location. 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.

2.3. Statistical Analysis

Phenotypic traits were analysed with a mixed effects model with the form:

$$y_{ijk} = \mu + E_i + B(E_i) + G_k + GE_{ik} + \epsilon_{ijk}$$

Where y_{ijk} is the phenotypic measurement for rep$j$ of genotype k in environment i, E_i is the effect of environment i, $B(E_i)$ is the effect of replication nested within environment, G_k is the effect of genotype G, GE_{ik} is the interaction effect of environment E and genotype G, and ϵ_{ijk} is the measurement error. The genotype effect was treated as fixed and all other factors were treated as random.

Models were fit using the gamem_met function from the metan package[13]. Least-square means (LS Means) for each genotype and trait were calculated using the above model using the emmeans package[14] in R. The emmeans package was also used to calculate contrasts as a post-hoc test to compare RIL phenotype means to check means for all collected phenotypes. A sidak adjustment was used to account for multiple comparisons in the calculation of contrasts.Pearson correlation coefficients between each phenotype were calculated with the metan package. Pearson correlation coefficients between each phenotype were calculated with the metan package as well.

The 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.

3. Results and discussion

3.1. Phenotypic correlations

A strong negative correlation was observed in both tests between seed protein and seed oil content. A coefficient of r = -0.93 was observed in both populations. A negative correlation was observed between seed protein and yield with correlation coefficient of r = -0.29 was observed in both tests. This correlation was not statistically significant in either population. A weak positive correlation was observed between seed oil and seed yield in both populations. A correlation coefficient of r = 0.24 was observed in Test 1 while a coefficient of r = 0.28 was observed in Test 2. These correlation coefficients were not statistically significant in either population. Visualizations of the pairwise correlation coefficients for Tests 1 and 2 can be seen in supplementary figures 2 and 3, respectively.

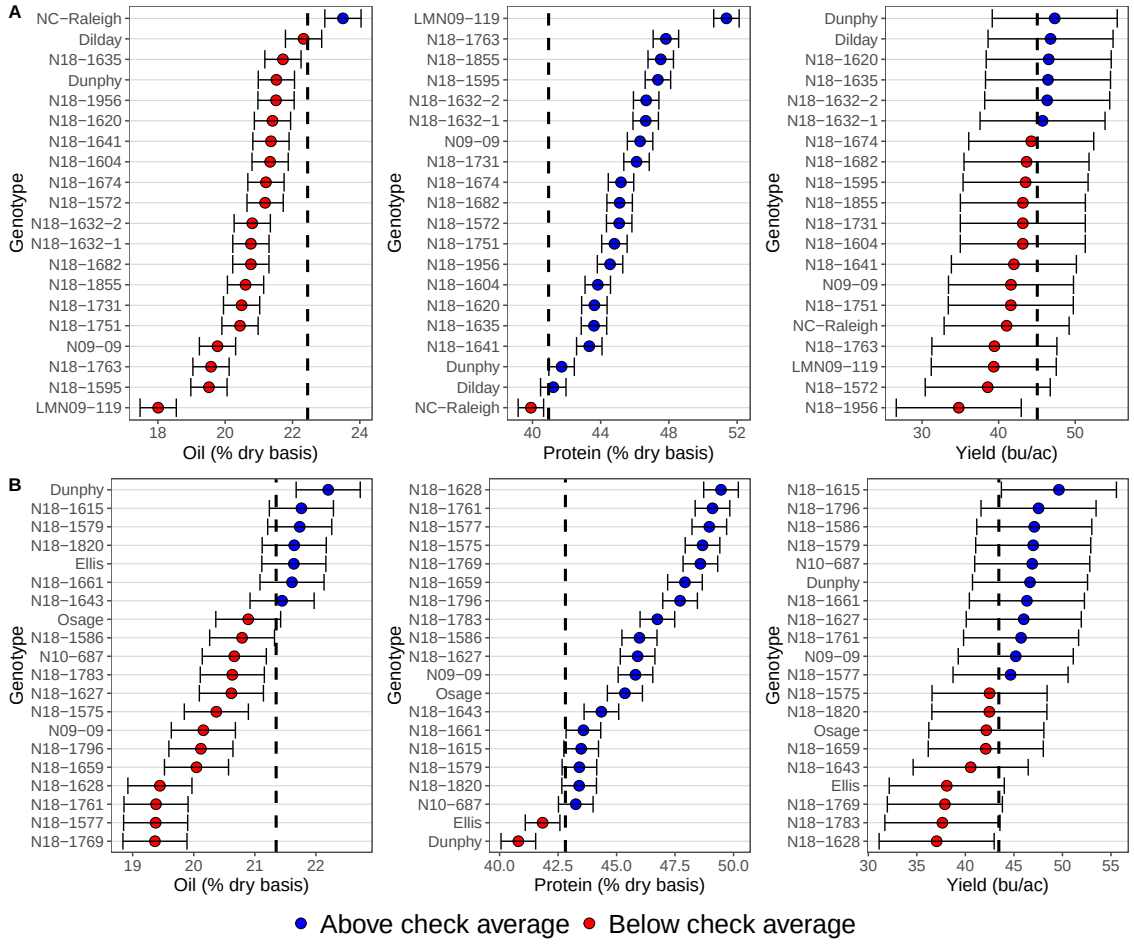

Figure 1. Genotype Least square means for seed oil, seed protein, and seed yield for soybean RILs in Test 1 (**A**) and Test 2 (**B**). Points indicate the least square means and error bars indicate the standard errors in the estimation of least square means. Blue dots indicate that a RIL has a least square mean above the average of the checks in the test and a color of red indicates that the RIL has a mean value below the check average. The check average is shown with the vertical dashed line.

Genotype least square means were calculated for all RILs in both tests. A visualization of these least square means and the standard error on the estimation of these means for seed protein, seed oil, and seed yield can be seen in figure 1. Both tests tended to have more RILs with high protein rather than high oil when compared to the average of the checks, and a moderate number had a yield average greater than the average of the checks. Qualitatively, many RILs had a comparable yield to the checks following the relatively large standard error associated with the estimating the yield, and a

qualitatively larger seed protein content given the relatively small standard error associated with the estimation of seed protein content. To separate means quantitatively, we performed contrasts between each RIL mean and the average of the checks for each test.

3.2. Yield contrasts

Several RILs in each of the tests had yield that was comparable to that of their check cultivars as per contrasts between RIL means and the average of the check cultivars for each test. No genotypes had yield that was quantitatively higher than that of the yield checks, however many had comparable yield. Many of these genotypes with comparable yield also had protein content that was superior to the check average in each test. No genotypes had oil that was superior to the check cultivars, but some had comparable yield and oil content as well as superior protein content.

RILs were also evaluated for lodging and seed quality. All RILs had seed quality that was on par with that of the high-yielding checks however, however 11 genotypes had lodging scores that were statistically significantly greater than the checks on the basis of contrasts. These RILs were removed from consideration for recommendation on the basis of their relatively poor agronomic performance.

A detailed summary of all contrasts can be seen in supplementary table 1.

3.3. Genotypes with comparable seed yield and seed oil and superior seed protein content

Four genotypes had yield and oil that was similar to that of the yield checks, and seed protein that was greater than the average of the checks. These genotypes are N18-1635 from Test 1 and genotypes N18-1627, N18-1643, and N18-1783 from Test 2. Summary data for these lines is given in table 1.

Table 1. Soybean genotypes with yield and seed oil comparable to check cultivars, and seed protein superior to check cultivars.

		Protein				Yield				Oil			
Genotype	Test Name	Value	Rank	Test Average	Check Average	Value	Rank	Test Average	Check Average	Value	Rank	Test Average	Check Average
N18-1635	Jay Test 1	43.61 (106.5%)	16	45.09	40.94	46.45 (103.1%)	4	42.92	45.05	21.72 (96.7%)	3	20.89	22.45
N18-1783	Jay Test 2	46.74 (109.3%)	8	45.7	42.77	37.63 (86.6%)	19	43.65	43.44	20.63 (96.5%)	11	20.7	21.37
N18-1627		45.9 (107.3%)	9			46.01 (105.9%)	8			20.62 (96.5%)	12		
N18-1643		44.35 (103.7%)	13			40.54 (93.3%)	16			21.45 (100.4%)	7		

* The genotype name.
† The genotype marginal mean for the phenotype (value divided by check average).
‡ The ranking of this genotype for the phenotype within its test.
§ The average phenotype value for all genotypes in the test.
¶ The average value of the checks in the test.

A strong negative correlation was observed in both populations between seed oil and seed protein, and a weak inverse correlation was observed between seed protein and seed yield. This observation matches well with previous findings on the inverse correlation between seed protein and seed oil, and between seed protein and seed yield. Many RILs had high protein when compared with the check cultivars so it was not surprising that relatively few had both a superior seed protein content, and a comparable seed oil content.

These RILs can provide valuable germplasm that can be used to simultaneously improve both seed oil and seed protein content without compromising seed yield or agronomics. This can be of use both to breeders that are looking to improve seed composition traits in soybean and for growers who are seeking options for novel soybean varieties with seed composition that is superior to that of cultivars that are already used in production.

3.4. Genotypes with comparable yield and superior protein

Seventeen RILs had a similar yield and superior protein content to the check cultivars. Summary data for the seed protein and seed yield of these lines is given in table 2. From among these RILs, N18-1627 had the highest seed yield relative to the checks included in the test with an average yield that was 105.9% that of the check average of Test 2. Genotype N18-1763 had the highest seed protein content on average with a protein content that was 116.8% that of the check average of Test 1. We observed

a negative correlation between seed protein and seed yield in both tests so RILs with seed protein content that was above the check average tended to have a seed yield content that was comparatively lower.

Table 2. Soybean RILs with superior protein content and comparable yield performance to high-yielding check cultivars.

		Protein				Yield			
Genotype	Test Name	Value	Rank	Test Average	Check Average	Value	Rank	Test Average	Check Average
N18-1763	Jay Test 1	47.82 (116.8%)	2	45.09	40.94	39.45 (87.6%)	17	42.92	45.05
N18-1855		47.52 (116.1%)	3			43.17 (95.8%)	10		
N18-1595		47.36 (115.7%)	4			43.52 (96.6%)	9		
N18-1632-1		46.64 (113.9%)	6			45.73 (101.5%)	6		
N18-1620		43.63 (106.6%)	15			46.54 (103.3%)	3		
N18-1635		43.61 (106.5%)	16			46.45 (103.1%)	4		
N18-1731		46.1 (112.6%)	8			43.16 (95.8%)	11		
N18-1674		45.19 (110.4%)	9			44.26 (98.2%)	7		
N18-1682		45.11 (110.2%)	10			43.64 (96.9%)	8		
N18-1572		45.08 (110.1%)	11			38.58 (85.6%)	19		
N18-1751		44.8 (109.4%)	12			41.59 (92.3%)	15		
N18-1761	Jay Test 2	49.1 (114.8%)	2	45.7	42.77	45.74 (105.3%)	9	43.65	43.44
N18-1575		48.67 (113.8%)	4			42.49 (97.8%)	12		
N18-1769		48.58 (113.6%)	5			37.88 (87.2%)	18		
N18-1783		46.74 (109.3%)	8			37.63 (86.6%)	19		
N18-1627		45.9 (107.3%)	9			46.01 (105.9%)	8		
N18-1643		44.35 (103.7%)	13			40.54 (93.3%)	16		

* The genotype name.
† The genotype marginal mean for the phenotype (value divided by check average).
‡ The ranking of this genotype for the phenotype within its test.
§ The average phenotype value for all genotypes in the test.
¶ The average value of the checks in the test.

However, among these genotypes are several with both average seed protein and average seed yield that were greater than the averages of the checks for each test. These N18-1632-1, N18-1620, N18-1635 from test 1 and RILs N18-1761 and N18-1627 from test 2. Of particular note are genotypes N18-1632-1 from test 1 and N18-1761 from test 2. These two RILs have a seed protein content that is substantially above that of the average protein content of he checks while also maintaining an average yield that is above the average yield of the checks. These RILs are ideal candidates for use in breeding programs which seek to improve seed protein content in particular without compromising seed yield or argonomics.

4. Conclusion

We have identified several genotypes with seed yield that is comparable to existing cultivars that are commonly used in production. Many of these genotypes have seed protein content that exceeds that of the check cultivars, and some of these with superior protein have a seed oil content that is comparable to these check cultivars as well. These genotypes have good agronomic qualities as well and will provide both breeders and growers with new options for genotypes that can be used in production or used in breeding programs that seek to improve valuable soybean seed composition traits.

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: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RIL	Recombinant inbred line
RCBD	Randomized complete block design
CLA	Central crops research station
CAS	Caswell research farm
PLY	Tidewater research station

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. Burton, J.; Carter Jr, T.; Fountain, M.; Bowman, D. Registration of'NC-Raleigh'soybean. *46*, 2710.
13. Olivoto, T.; Lucio, A. Metan: An R Package for Multi-Environment Trial Analysis. *11*, 783–789. doi:10.1111/2041-210X.13384.
14. Lenth, R.V. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*.

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

© 2022 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).