Plant Breeding

WILEY

Plant Breeding

ORIGINAL ARTICLE OPEN ACCESS

Evaluation of Deviation-Type Indices for the Joint Selection of Yield and Seed Composition Traits in Soybean (*Glycine max* (L.) Merr.)

Vincent Braun[1] | Clemens Hacke[2] | Tobias Würschum[3,4] | Karl Schmid[2,4] | Volker Hahn[1] | Patrick Thorwarth[1,4]

[1]State Plant Breeding Institute, University of Hohenheim, Stuttgart, Germany | [2]Institute of Crop Biodiversity and Breeding Informatics, University of Hohenheim, Stuttgart, Germany | [3]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany | [4]Cluster of Excellence GreenRobust, University of Hohenheim, Stuttgart, Germany

Correspondence: Patrick Thorwarth (patrick.thorwarth@uni-hohenheim.de)

Received: 23 January 2026 | **Revised:** 6 April 2026 | **Accepted:** 20 May 2026

Keywords: genome wide association studies | grain protein deviation | selection indices | soybean breeding

ABSTRACT

Seed yield and composition are major targets in soybean (*Glycine max* (L.) Merr.) breeding programmes, but simultaneous improvement is constrained by unfavourable correlations among target traits. In this study, we evaluated the potential of different selection indices to simultaneously improve seed yield and quality traits in soybean. Indices were derived from seed yield, protein content, oil content, and combined protein and oil content. For each quality trait, we calculated both yield-based indices (protein yield, oil yield and protein-plus-oil yield) and deviation-type indices (grain protein deviation [GPD], grain oil deviation [GOD] and grain content deviation [GCD]). In addition, three different methods were used to derive the deviation-type indices: plot-based regression, regression based on genotype BLUEs and a bivariate mixed model approach. Our analysis utilized data from 467 soybean genotypes evaluated in advanced multi-environment yield trials. All quality traits and deviation-type indices exhibited high heritability ($h^2 > 0.86$), whereas yield and yield-based indices showed moderate heritability ($h^2 > 0.68$). Genome-wide association studies (GWASs) were performed for all traits and indices using genotypic data from 333 lines. Results showed that selection on deviation-type indices is primarily driven by the underlying quality trait, as indicated by very high correlations and overlapping significant marker-trait associations. In contrast, yield-based indices reflected the combined influence of seed yield and the corresponding quality trait. Notably, GWAS on the bivariate-derived GPD index identified additional significant loci on chromosome 18 that were not detected for the univariate traits, suggesting that the bivariate approach captured aspects of the covariance structure between protein content and seed yield. Overall, our results suggest that deviation-type indices in soybean offer little practical advantage over direct selection for corresponding quality traits in materials with weak or environment-dependent trait correlations. At the same time, comparing index derivation methods shows that constructing phenotypes influences the interpretation of genetics downstream and that modelling multiple traits can reveal additional loci relevant to yield–quality trade-offs.

1 | Introduction

Soybean (*Glycine max* (L.) Merr.) is one of the most important crops worldwide, and by far the most important legume, due to its unique seed composition. Seeds contain approximately 40% protein and 20% oil content (OC; Sedibe et al. 2023; Shea et al. 2020). In the early 2010s, the European Union recognized its dependence on protein imports (Häusling 2011), and

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Plant Breeding* published by Wiley-VCH GmbH.

since then, the area dedicated to soybean cultivation has doubled. Breeding programmes have also been established and expanded with the aim of developing high yielding varieties with high protein contents (PCs) adapted to central European climates (Eurostat 2025; Fogelberg and Recknagel 2017; Hahn and Würschum 2014; Panthee 2010; Sato et al. 2014). A central challenge in soybean breeding is that target traits are not independent. Specifically, PC and OC are often negatively correlated, as is PC with seed yield (SY; Diers et al. 2023; Jin et al. 2023; Kambhampati et al. 2019; Kurasch et al. 2017; Wilcox 1998). These relationships complicate simultaneous genetic improvement and necessitate selection strategies that go beyond simple single-trait maximization.

Negative associations among SY, PC and OC have been described repeatedly in soybean, both at the phenotypic and genetic level (Hwang et al. 2014; Shi et al. 2025; Yu et al. 2025). Previous studies have identified shared and trait-specific loci for seed composition, and several authors have emphasized the difficulty of increasing PC without compromising either SY or oil concentration (Shi et al. 2025). At the same time, the magnitude and even direction of these relationships can vary across germplasm and environments, indicating that both genetic background and environmental context influence the effectiveness of different selection approaches (Assefa, Purcell, et al. 2019; Cober and Voldeng 2000; Wilcox and Cavins 1995).

One way to address antagonistic trait relationships is using selection indices. Selection theory provides a mathematical framework for the simultaneous improvement of two or more traits. Classical multi-trait indices, such as weighted linear indices (e.g., Smith–Hazel), combine several traits into a single score based on predefined economic or statistical weights (Hazel 1943; Smith 1936). Another approach is the use of deviation-type indices. These indices quantify whether a genotype's performance for target quality traits is better or worse than expected at a given yield level. In cereal breeding, grain protein deviation (GPD) and related indices have been proposed to identify genotypes that partially break the negative relationship between SY and PC. GPD is here defined as the residual of the linear regression of PC on SY (Koekemoer et al. 1999; Monaghan et al. 2001; Neuweiler et al. 2021; Rapp et al. 2018; Thorwarth et al. 2018). However, the usefulness of such indices depends on how they are calculated and on the strength and stability of the underlying trait relationships.

The yield-protein trade-off in soybean has been extensively discussed, and protein yield (PY) is often considered a practical breeding target (Kurasch et al. 2017). However, to our knowledge, deviation-type indices such as GPD have not been systematically evaluated in soybean. Additionally, it is unclear whether different statistical methods for deriving these indices result in similar phenotypic rankings and genetic signals. The objectives of this study were therefore to (i) evaluate the relationships among SY, PC, OC and protein-plus-oil content (POC) in a soybean breeding population, (ii) compare three approaches for calculating deviation-type indices and (iii) assess the phenotypic behaviour and genetic architecture of these traits and indices using GWAS.

2 | Material and Methods

2.1 | Plant Material and Field Trials

This study was based on plant material from the soybean (*G. max* (L.) Merr.) breeding programme at the State Plant Breeding Institute of the University of Hohenheim. The dataset included 467 genotypes, 23 of which were registered varieties, and 444 of which were advanced F6 to F10 lines developed using a single-seed descent approach. The data covered several field trials from 2022 to 2024. The genotypes were divided into trials and experiments according to their maturity group from the previous year (LP601 to LP604). Additionally, two field trials, independent of the maturity group, were included in this study (LP605 and LP606). Note that although trials are named identically, they refer to different experiments each year. Each trial was designed as an α-design with 50 varieties, 20 incomplete blocks and two replicates, except for LP605, which consisted of only 12 varieties, eight incomplete blocks and two replicates. Each trial was planted in a different number of locations. Further information on the sites and weather conditions during the growing season can be found in Tables S1 and S2. Depending on the location, all yield plots ranged in area from 8.4 to 13.5 m^2. SY was measured in decitonnes per hectare ($dt\,ha^{-1}$) and adjusted to a dry matter basis (0% moisture). In 2022 and 2023, the quality traits seed PC and seed OC were measured as percentages based on dry matter using a Polytec 2120 spectrometer (Polytec GmbH, Waldenbronn, Germany) and near infrared spectroscopy. In 2024, the same quality traits were measured using a Zeiss Corona Extreme (Zeiss AG, Oberkochen, Germany).

2.2 | Phenotypic Data Analysis

We estimated the adjusted means (best linear unbiased estimates [BLUEs]) for each trait across all environments using the following mixed model in a one-stage approach:

$$y_{ijklm} = \mu + g_i + l_j + gl_{ij} + lt_{jk} + r_{jkl} + b_{jklm} + e_{ijklm} \quad (1)$$

where y_{ijklm} denotes the observed value of the ith genotype at the jth environment, in the kth trial, the lth replication and the mth incomplete block; μ is the overall mean; g_i stands for the ith genotype; l_j is the jth environment (year-location combination); gl_{ij} is the genotype-by-environment interaction; lt_{jk} is the trial effect nested within environment; r_{jkl} is the replication within kth trial and jth environment; b_{jklm} is the mth incomplete block within the lth replication, kth trial and jth environment; and e_{ijklm} is the residual. All effects except g were treated as random, and residual variances were assumed to be homogenous across all environments. Outliers were detected for each trait separately using Bonferroni–Holms test (Bernal-Vasquez et al. 2016).

To estimate variance components the same model was used but all effects were considered to be random and the significance of each variance component was determined by a likelihood ratio test (Stram and Lee 1994). Due to the highly unbalanced data, the broad-sense heritability was calculated according to the following formula:

$$h^2 = 1 - \frac{\overline{\vartheta}_{BLUP}}{2\sigma_g^2} \quad (2)$$

where $\overline{\vartheta}_{BLUP}$ is the mean variance of a difference of two best linear unbiased predictors, and σ_g^2 is the genetic variance (Cullis et al. 2006; Schmidt et al. 2019). All statistical analysis were performed and figures build with the *R* software Version 4.3.1 (R Core Team 2023). The linear mixed models were handled with the *R* package ASReml-R Version 4.2.0.302 (Butler 2023).

2.3 | Calculation of Indices

In addition to PC and OC, we computed their sum—and OC (POC)—as a third base quality trait. For each of these traits, two indices combining quality and yield were calculated:

1. Yield based indices: These include PY, oil yield (OY), and protein- (POY), representing the yield of the respective components.
2. Deviation-type indices: These indices represent the deviation from expected quality content, given SY. Specifically, GPD, grain oil deviation (GOD) and grain content deviation (GCD) were calculated as the residuals from a linear regression of the quality trait on SY as follows:

$$y = \alpha + x\beta + \varepsilon, \quad (3)$$

where y is the dependent variable (PC, OC or POC); α is the intercept; x is the predictor variable (SY); β is the regression coefficient; and ε is the residual. The deviation index is defined as the residuals, which are the differences between the observed values (y) and the predicted values ($\hat{\alpha} + x\hat{\beta}$, Equation 4):

$$D = \hat{\varepsilon} = y - \left(\hat{\alpha} + x\hat{\beta}\right), \quad (4)$$

where $\hat{\varepsilon}$ stands for the respective deviation index, and $\hat{\alpha}$ and $\hat{\beta}$ are estimates of the intercept and the regression coefficient, respectively. A positive value of D indicates that a genotype shows a higher value of the respective quality trait, than expected based on the regression on SY.

Because the deviation depends on how the regression of quality on yield is estimated, the deviation-type indices were calculated with three different methods:

a. Plot-based: Linear regression fitted to each plot, calculated within each environment (Monaghan et al. 2001).
b. BLUE-based: Regression based on the BLUEs of genotypes across all trials (Latshaw et al. 2016). In this approach, genotypes are adjusted for environmental and design effects before calculating the regression.
c. Bivariate mixed model: A bivariate linear mixed model following Thorwarth et al. (2018), accounting for the genetic covariance between the quality trait and yield. The deviation was calculated as follows:

$$D = y_T - \rho\frac{\sigma_T}{\sigma_{SY}}y_{SY} \quad (5)$$

where D stands for the respective deviation, y_T and y_{SY} represent the values of the quality trait and SY on a plot basis, σ_T and σ_{SY} are the estimates of the genetic variances, and ρ is the genetic correlation. Unlike the phenotypic regressions (a and b), the bivariate mixed model (c) explicitly partitions the genetic covariance between both traits. Accordingly, the resulting deviation index can be interpreted as a genetically adjusted deviation from SY.

We compared all three approaches to assess how strongly the resulting indices, genotype rankings and GWAS results depend on the method used for phenotype construction. Table 1 is giving an overview of all evaluated traits.

2.4 | Genetic Material

The genetic material used in this study will be published in a companion study. We sequenced 344 genotypes from the 2023 and 2024 trials using whole-genome sequencing with an average sequencing depth of three. The raw reads were mapped to Version 4 of the Williams 82 reference genome (Valliyodan et al. 2019). We conducted variant calling using the bcftools mpileup function. The VCF file was strictly filtered. We removed all genotypes with more than 35% missing SNPs. This reduced the number of genotypes to 333. Then, we removed all indels and non-biallelic SNPs and set the minimum sequencing depth of each SNP to three, the minimum quality to 30 and the maximum missing data for SNPs to 25%. This resulted in a dataset of 333 genotypes and 377,314 SNPs. We then imputed the dataset using Beagle Version 5.4 (Browning 2021).

2.5 | Genetic Data Analysis

Genome-wide association studies (GWASs) were conducted using REGENIE Version 4.1 (Mbatchou et al. 2021). Genotypic data in VCF format were converted to PLINK format using PLINK Version 1.9 (Chang et al. 2015; Purcell and Chang 2015). Principal component analysis (PCA) was performed on the genetic data to assess population structure, and the first four principal components were included as covariates in each GWAS model to control for confounding due to population stratification.

GWAS analyses were performed separately for all traits and derived indices. Manhattan and quantile–quantile (QQ) plots illustrating GWAS results were generated using the R package 'qqman' (Stephen Turner 2018). To account for multiple testing, SNP-trait associations with p-values below the Bonferroni-corrected significance threshold ($\alpha = 0.05$) were considered statistically significant.

For traits sharing significant SNPs, we plotted scatterplots of the respective effect sizes for shared SNPs to visualize pleiotropic effects. Additionally, for each set of three deviation-type indices, we visualized pairwise similarity of SNP effects using heatmaps.

We grouped the significant SNPs into separate loci based on LD decay estimates. We calculated LD decay using PopLDdecay

TABLE 1 | Overview of measured traits and the indices derived from them.

Trait	Abbreviation	Description/formula	Unit
Seed yield	SY	Seed yield	dt/ha
Protein content	PC	Protein content	%
Oil content	OC	Oil content	%
Protein and oil content	POC	PC + OC	%
Protein yield	PY	PC × SY	dt/ha
Oil yield	OY	OC × SY	dt/ha
Protein and oil yield	POY	POC × SY	dt/ha
Grain protein deviation	GPD	Residuals of linear regression of PC on SY at environment level	—
Grain protein deviation bivariate	GPD_{bivar}	Bivariate model for PC and SY used to calculate GPD	—
Grain protein deviation BLUEs	GPD_{blu}	Residuals of linear regression of PC on SY, BLUEs across series were used	—
Grain oil deviation	GOD	Residuals of linear regression of OC on SY at environment level	—
Grain oil deviation bivariate	GOD_{bivar}	Bivariate model for OC and SY used to calculate GOD	—
Grain oil deviation BLUEs	GOD_{blu}	Residuals of linear regression of OC on SY, BLUEs across series were used	—
Grain content deviation	GCD	Residuals of linear regression of POC on SY at environment level	—
Grain content deviation bivariate	GCD_{bivar}	Bivariate model for POC and SY used to calculate GCD	—
Grain content deviation BLUEs	GCD_{blu}	Residuals of linear regression of POC on SY, BLUEs across series were used	—

(C. Zhang et al. 2019). We then chose the SNP with the highest effect size for each locus-trait combination and calculated the explained phenotypic and genotypic variance using the previously calculated phenotypic and genetic variances of the traits and the effect size of the identified significant SNPs as input.

3 | Results

3.1 | Population Structure and Heritability

A total of 467 soybean genotypes, comprising 23 registered varieties and 444 advanced breeding lines, were evaluated across multiple environments. The PCA analysis of the 333 sequenced genotypes did not show strong population structure. The first two principal components explained 24.4% and 17.5% of the variance calculated across the eight computed principal components. No clustering of genotypes was evident, and the variance was distributed continuously, indicating differences in the ancestry of the genotypes as the main source of variation.

Substantial genotypic variation was present for SY, PC and OC, as well as for derived indices, as indicated by the broad range of BLUEs (Table 2). All traits and indices exhibited significant genotypic and genotype-by-environment variance components ($p < 0.001$). For most traits, environmental variance exceeded genetic variance; however, the reverse was observed for PC and deviation-type indices calculated based on the plot values (Method a). Heritabilities ranged from moderate to very high (from 0.66 for PY to 0.95 for POC), with quality traits and deviation-type indices generally outperforming yield-based traits.

3.2 | Phenotypic Correlations and Selection Indices

Correlations among traits reflected expected genetic relationships. SY was moderately negatively correlated with PC ($r = -0.28$) and moderately positively correlated with OC ($r = 0.25$), whereas the sum of PC and OC (POC) was also negatively correlated with SY ($r = -0.27$). By contrast, PC and OC exhibited a strong negative correlation ($r = -0.90$), consistent with antagonistic pleiotropy. Yield-based indices (PY, OY, POY) correlated very strongly with SY ($r = 0.89$–0.98).

Deviation-type indices (GPD, GOD, GCD) showed correlations with SY similar in magnitude to the base quality trait, but with a sign switch when using the bivariate calculation method—reflecting the explicit modelling of genetic covariance (see Section 2.3). Notably, when deviation indices were calculated from genotype BLUEs, their correlation with SY vanished. For all methods, however, deviation-type indices remained highly correlated with their respective underlying quality trait ($r = 0.86$–0.97; Figure 1).

Thus, in the evaluated material, deviation-type indices largely preserved the ranking of genotypes for the underlying quality traits, whereas yield-based indices more strongly reflected the influence of SY.

TABLE 2 | Summary statistics, estimates of variance components and heritabilities for the evaluated traits seed yield (SY), protein content (PC), oil content (OC), sum of protein and oil content (POC) and the derived protein yield (PY), oil yield (OY), protein and oil yield (POY), grain protein deviation (GPD/GPD_{bivar}), grain oil deviation (GOD/GOD_{bivar}) and grain content deviation (GCD/GCD_{bivar}).

	SY	PC	OC	POC	PY	OY	POY	GPD	GPD_{bivar}	GOD	GOD_{bivar}	GCD	GCD_{bivar}
Min	20.76	37.53	16.39	58.51	9.20	3.98	13.17	−4.77	−1.33	−4.80	−2.37	−1.88	−2.46
Mean	32.15	42.63	18.71	61.36	13.76	6.09	19.84	0.08	0.31	0.00	−0.03	−0.24	0.00
Max	41.49	48.40	21.26	64.92	17.59	8.60	25.26	5.62	2.59	6.27	2.42	1.45	2.47
σ^2_g	4.49***	3.39***	0.91***	1.12***	0.73***	0.34***	1.46***	3.09***	0.30***	0.85***	0.30***	1.04***	0.12***
σ^2_l	96.33***	2.99***	1.42***	4.78***	19.80***	4.2***	42.14***	0.00	2.17***	0.00	0.85***	0.00	1.62***
$\sigma^2_{g\times l}$	5.66***	0.40***	0.11***	0.16***	1.00***	0.21***	1.99***	0.44***	0.11***	0.12***	0.11***	0.16***	0.04***
σ^2_e	7.67	0.68	0.16	0.35	1.49	0.29	2.98	0.72	0.19	0.17	0.16	0.37	0.08
h^2	0.68	0.95	0.95	0.93	0.66	0.81	0.66	0.94	0.86	0.95	0.87	0.93	0.86

Note: *, ** and *** significantly different from zero at the 0.05, 0.01 and 0.001 level of probability, respectively. σ^2_g, genotypic variance; σ^2_l, environment variance; $\sigma^2_{g\times l}$, variance of genotype-by-environment interaction; $\sigma^2_{l\times l}$, variance of the trial-by-environment interaction; σ^2_e, error variance; h^2, heritability.

3.3 | Visualization of Selection Outcomes

Figure 2 illustrates the relationship between traits and how different index-based selection methods prioritize genotypes. Deviation-based indices consistently identified the top 10% of high-quality performers, many of whom also ranked highly for yield. PY and OY, as well as POY, predominantly selected for the highest-yielding genotypes, resulting in more overlap with yield per se, particularly regarding OC. Notably, the bivariate deviation method shifted the selection of individuals toward higher SY (but lower PC and POC, respectively) for protein-related traits and toward lower OC and SY for oil compared with simple regression-based indices.

3.4 | GWAS Results

3.4.1 | Main Trait GWAS

Analysis of 333 genotypes yielded significant associations for SY on chromosome 13 (14 SNPs, positions 28.8–29.8 Mb; Figure 3). For PC, three associations were detected on chromosome 6 (46.3–46.8 Mb) and 103 on chromosome 20 (1.1–34.8 Mb, with the majority at 31.3–34.8 Mb; Figure 4). OC was associated exclusively with loci on chromosome 20, which also overlapped most PC-associated regions (Table 3). Scatterplots of effect sizes for overlapping loci confirmed the antagonistic effects of shared SNPs on PC and OC (Figure S1).

3.4.2 | Index Trait GWAS

Yield-based indices showed varied results. PY yielded no significant SNPs, whereas OY displayed 58 significant associations on chromosome 20 and POY four on chromosome 13, matching the loci found for SY (Figures S2–S4). This pattern aligns with the strong yield-component influence in these indices.

3.4.3 | Deviation-Type GWAS

For GPD, the plot-based and BLUE-based deviation traits largely recapitulated the major associations observed for PC on chromosomes 6 and 20, whereas GWAS on the bivariate-derived GPD trait identified seven additional significant SNPs on chromosome 18 (56.2–56.7 Mb; Figure S5). These SNPs were not significant for any of the univariate traits, although some showed sub-threshold association with SY, suggesting that this phenotype may capture genetic variation related to the covariance structure between PC and SY.

For GOD, all three index calculation methods identified significant associations exclusively on chromosome 20, with only minor differences among the plot-based, BLUE-based and bivariate-derived indices (Figures 4 and S6).

For GCD, significant loci were again nearly identical between plot-based and BLUE-based models (chromosomes 6 and 20; Figures S7 and S8). In contrast, GWAS on the bivariate-derived GCD trait shifted the association pattern toward chromosome 6, where 14 SNPs were detected between 46.4 and 46.8 Mb,

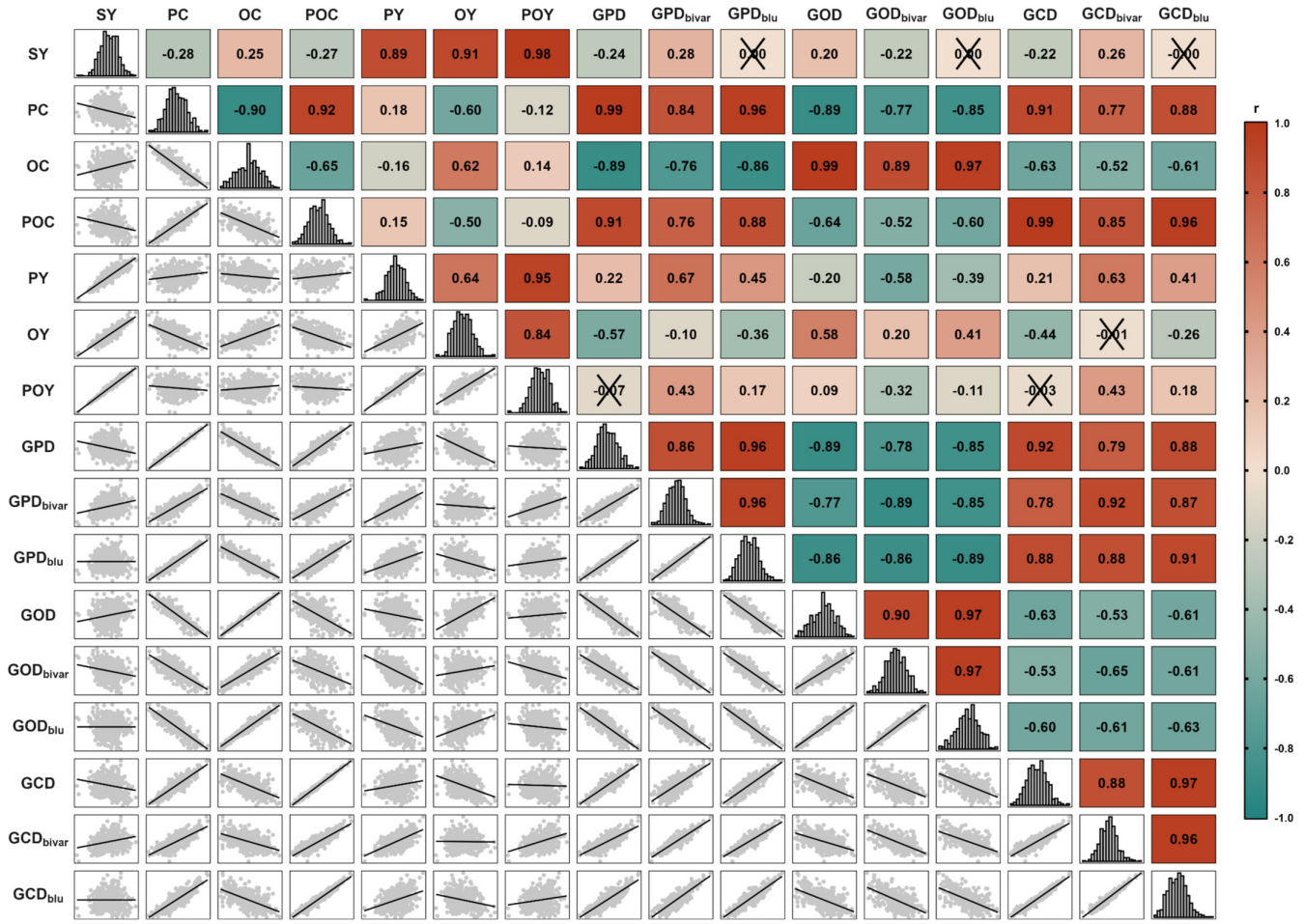


FIGURE 1 | Correlation plot of adjusted means (BLUEs) between all traits and indices (SY, seed yield; PC, protein content; OC, oil content; POC, protein plus oil content; PY, protein yield; OY, oil yield; POY, protein plus oil yield; GPD, grain protein deviation; GPD_{biv}, GPD calculated using the bivariate method; GPD_{blu}, GPD based on BLUEs; GOD, grain oil deviation; GOD_{biv}, GOD calculated using the bivariate method; GOD_{blu}, GOD based on BLUEs; GCD, grain content deviation; GCD_{biv}, GCD calculated using the bivariate method; GCD_{blu}, GCD based on BLUEs). The diagonal shows the histogram of each trait, the lower panel the regression plots and the upper panel the numeric and coloured correlation values. Non-significant correlations at $p < 0.05$ are crossed out.

whereas only a single SNP remained significant on chromosome 20 at 34.8 Mb (Figure S9). Comparisons of effect sizes for shared SNPs showed consistent directions across all three traits; however, the bivariate-derived trait generally yielded smaller effect sizes (see Figures S10–S12).

Overall, the three approaches produced highly similar results for GOD and, to a lesser extent, GCD. In contrast, the bivariate-derived GPD trait showed the clearest deviation from the association patterns observed for the underlying univariate traits. This suggests that the method used to construct phenotypes can affect subsequent genetic inferences.

3.5 | Linkage Blocks

The LD decay analysis indicated a drop of LD below r^2 of 0.2 at a distance of ~30,000 bp. We used this distance to group the significant SNPs into linked loci. This resulted in 53 loci being distributed across chromosomes 6, 13, 18 and 20. As indicated above, we observed strong overlap between the loci identified using the individual traits (Figure S13). For each locus-trait combination, we then identified the SNP with the largest effect and reported the explained phenotypic variance and explained genotypic variance (Table S3). On average, the identified loci explained only 3.4% of the phenotypic variance and 8.7% of the genotypic variance. We identified locus 45 on chromosome 20 between positions 31,892,840 and 31,929,316 as the most important locus for PC, OC, GPD and GOD explaining between 9.43% and 19.44% of the phenotypic variance. The locus exerted a positive effect on protein-related traits but a negative effect on oil-related traits. We then checked all 53 loci for annotated genes using the genome browser tool on SoyBase (Brown et al. 2021, Table S4). Many of the annotated genes were proteins of unknown or uncharacterized function. Apart from that, many genes with known functions in breaking proteins into amino acids, creating proteins from amino acids, stabilizing proteins or transporting proteins were identified. Of particular interest were the loci 8, 9 and 10 on chromosome 6 for all of which we identified a SAUR-like auxin-responsive protein family. All three loci were significant for all three GCD models, as well as POC. Locus 8 was in addition significant for PC. The loci explained between 4% and 9% of the genetic variance for their respective

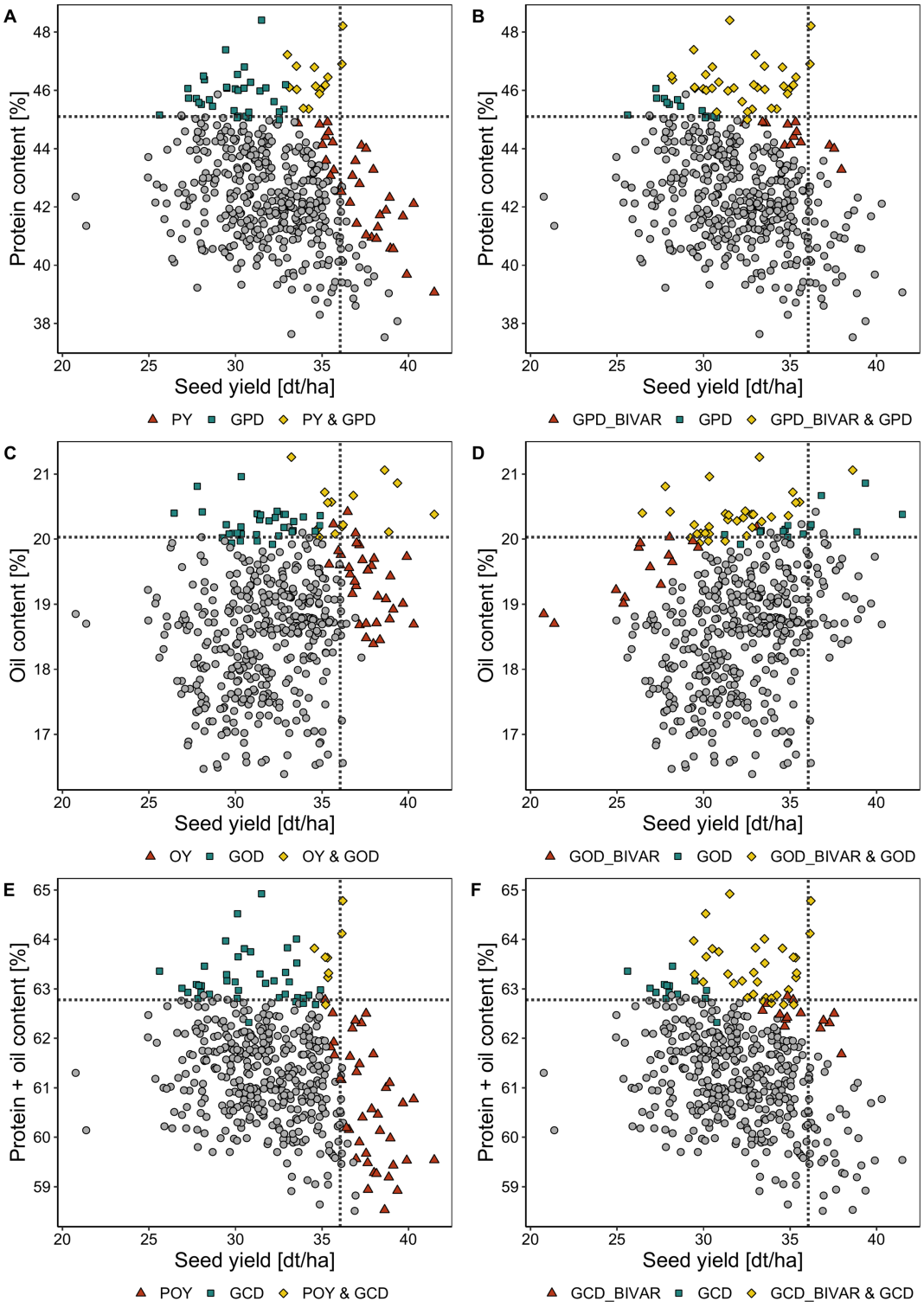


FIGURE 2 | Scatterplots of the BLUEs for protein content (A + B), oil content (C + D) and the sum of protein and oil content (E + F) on seed yield. In each plot, the dotted lines show the top 10% threshold for seed yield and the respective quality trait. The best 10% of the genotypes according to the different indices are coloured and shaped. In A, C and E the yield indices of the respective quality trait (▲; PY: protein yield, OY: oil yield, POY: protein plus oil yield) are compared with the deviation-type index calculated by the standard method (■; GPD: grain protein deviation, GOD: grain oil deviation, GCD: grain content deviation). In B, D and F the standard method and the bivariate method for calculating the deviation-type index are compared. Genotypes that are in the best 10% for both indices are shown as yellow diamonds (◆).

traits. The SAUR-family genes are known to be involved in cell elongation and play a crucial role in plant development and especially root development (Markakis et al. 2013). Another relevant gene was found for locus 44, which together with loci 43 and 45 explained most of the variance for PC, OC, GPD and GOD. For this locus, we identified a 60S ribosomal

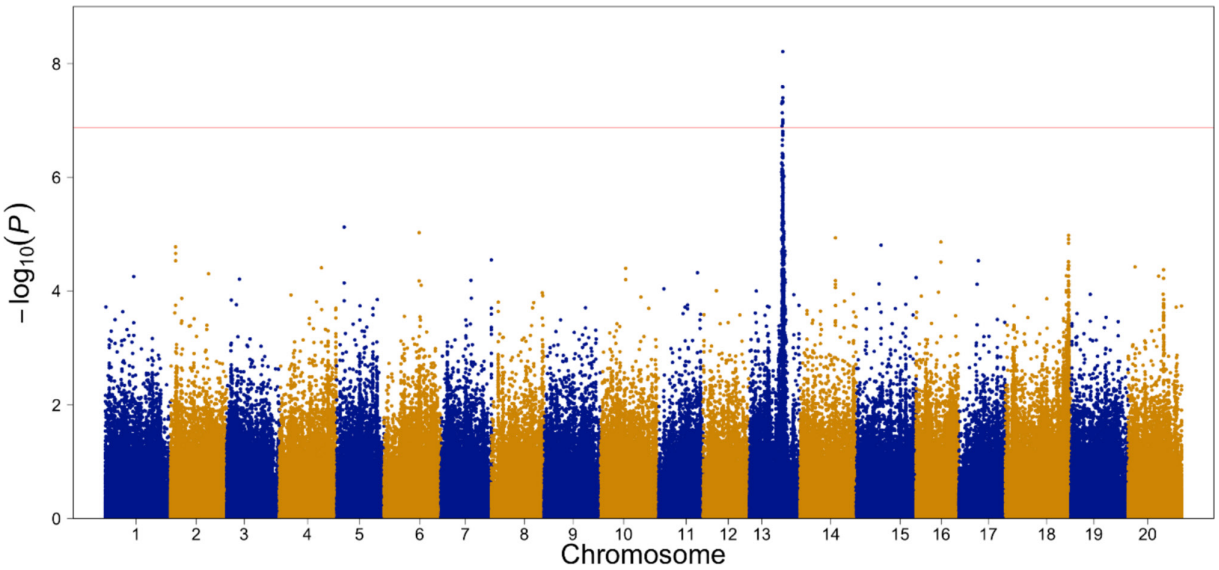


FIGURE 3 | Manhattan plot of the GWAS analysis for SY. The red line indicates the Bonferroni corrected significance threshold of $\alpha = 0.05$.

L23-like protein that plays a crucial role in protein synthesis (Degenhardt and Bonham-Smith 2008).

4 | Discussion

4.1 | Trait Relationships and Their Environmental Dependence

Negative correlations between SY and PC, as well as between PC and OC, are a long-standing challenge in soybean breeding. In agreement with previous studies, we observed a strong negative correlation between PC and OC and a weak negative correlation between SY and PC (Y. Assefa, Purcell, et al. 2019; Cober and Voldeng 2000; Duhnen et al. 2017; Helms and Orf 1998; Johnson et al. 1955; Kurasch et al. 2017; Wilcox and Cavins 1995). These relationships are consistent with the general view that soybean seed composition is shaped by physiological and biochemical trade-offs during seed filling, particularly in the allocation of carbon and nitrogen to storage compounds. The prevailing relationship reflects underlying biochemical and physiological trade-offs, which are not fully understood (Li et al. 2024). Although these antagonisms have been documented repeatedly, their magnitude is known to depend on the germplasm under study and on environmental conditions (Recker et al. 2014). This is also reflected in our data, where all major traits showed significant genotype-by-environment interaction and where environmental variance was substantial, especially for SY. This environmental sensitivity is important when interpreting the usefulness of selection indices. Our dataset revealed that SY was more strongly influenced by environmental variation than PC and OC. Consequently, the relationship between yield and quality was population- and environment-specific. This finding is consistent with previous reports indicating that temperature, water availability and related environmental factors can substantially impact both SY and soybean seed composition (Hou et al. 2006; Sobko et al. 2020; Wijewardana et al. 2018). Therefore, when applying any index that combines yield and quality traits, it is important to consider the practical value in the context of the population and target environment.

4.2 | Practical Value of Deviation-Type Indices in Soybean Breeding

The central aim of this study was to evaluate whether deviation-type indices could facilitate the simultaneous improvement of both SY and seed composition in soybean breeding. Such indices have attracted interest because they attempt to identify genotypes with a quality that exceeds expectations at a given yield level. In principle, they offer a way to target genotypes that partially escape unfavourable trait correlations. Although GPD has been used for this purpose in cereal crops (Rapp et al. 2018; Thorwarth et al. 2018), comparable evaluations in soybean have been lacking.

In the present study, deviation-type indices showed a high degree of correlation with their respective underlying quality traits, often selecting essentially the same genotypes. This was particularly evident for GPD, GOD and GCD, the correlations of which with the corresponding base traits remained very high regardless of the derivation method. Furthermore, the GWAS signatures of the standard and BLUE-based deviation indices largely mirrored those of the underlying quality traits. Overall, these results suggest that, in the soybean material examined in this study, deviation-type indices were primarily influenced by the associated quality trait, offering limited additional benefit over direct selection for that trait. This finding aligns with prior literature (Neuweiler et al. 2021; Rapp et al. 2018; Thorwarth et al. 2018).

This result is important because it clarifies the practical scope of these indices. Our findings do not support the notion that deviation-type indices are always better than direct selection in soybean breeding. Rather, their value appears to depend

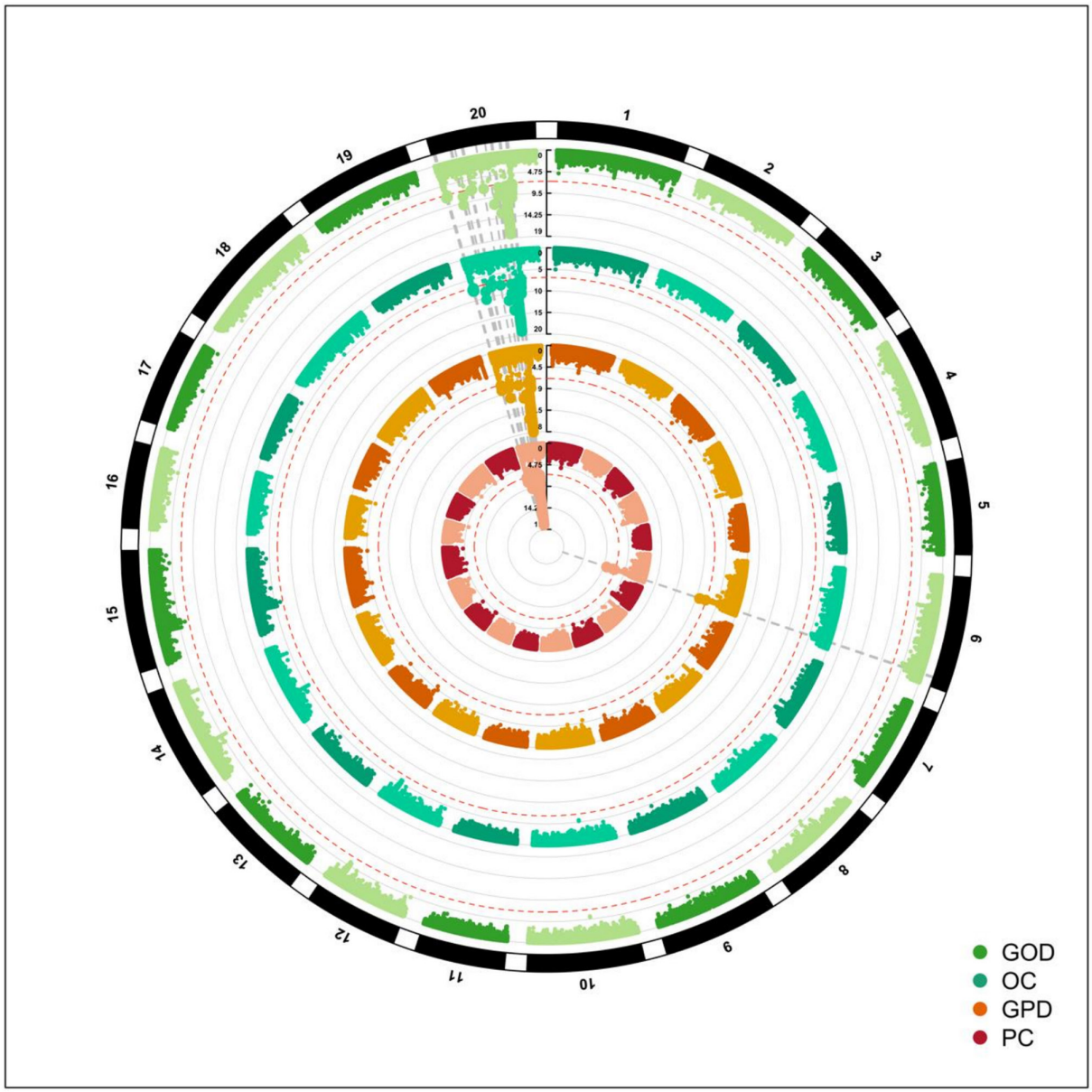


FIGURE 4 | Circular Manhattan plot comparing the GWAS results of GOD, OC, GPD and PC. The indices displayed were calculated using the standard method. The dashed red lines indicate the shared Bonferroni corrected significance threshold of $\alpha = 0.05$. The vertical grey dashed lines indicate positions with significant results for at least one of the traits.

on the context. In populations with weak, unstable or environmentally sensitive correlations between yield and quality, deviation-type indices may offer little additional benefit over direct selection for the quality trait (Table S1). This is also reflected in the high environmental variance of SY and composition in our study (Table 2), as well as in the environmental dependence observed in the work of others (Hou et al. 2006; Vollmann et al. 2000). In such cases, direct selection is likely to be equally effective and more straightforward from an operational breeding perspective. Nevertheless, deviation-type indices may be useful in specific scenarios where the explicit breeding objective is to identify genotypes that combine competitive yield with above-expected quality performance, especially when the underlying antagonism is stronger and more stable than in the present dataset.

4.3 | Yield-Based Indices and Their Interpretation

The yield-based indices behaved differently from the deviation-type indices. Although both classes of indices remained related to their component traits, the dominant driver differed. Yield-based indices were dominated by SY, whereas deviation-type indices were dominated by the corresponding quality trait. This is evident from their high correlations with yield, as well as the close overlap between several of their association signals and those identified for SY itself. This confirms that these indices do not isolate deviation from the yield–quality relationship, but instead represent the combined outcome of both components, with yield often dominant.

From a breeding perspective, these indices remain useful when the goal is to maximize the output of a valuable component

TABLE 3 | Overview of the identified loci for all traits, indicating the main chromosomes associated with the traits and shared associations with the three base traits: PC, OC and SY.

Trait/index	Main chromosome(s)	Shared SNPs (with OC/PC/SY)	Major model difference	Heritability
PC	6, 20	OC (chr 20)	Significant overlap with OC	High
OC	20	PC (chr 20)	Significant overlap with PC	High
POC	6, 20	OC (chr 20), PC (chr 6, 20)	Significant overlap with PC and OC	High
SY	13	—	Independent of PC and OC	Moderate
OY	20	OC (chr 20), PC (chr 20)	No overlap with SY	High
POY	13	SY (chr 13)	Significant overlap with SY	Moderate
GPD	6, 18, 20	OC (chr 20), PC (chr 6, 20)	Bivariate model identifies chr 18	High
GOD	20	OC (chr 20), PC (chr 20)	No meaningful model differences	High
GCD	6, 20	OC (chr 20), PC (chr 6, 20)	Bivariate model shifts the importance of the loci toward chr 6	High

per unit area rather than its concentration. For instance, PY could still be a relevant target for soybeans intended for feed or processing, provided breeders recognize that this index is heavily influenced by yield and is therefore sensitive to environmental factors. In our dataset, however, POY and the other yield-based indices did not resolve the underlying antagonistic relationships. Instead, they largely tracked the performance of SY itself. Therefore, although they may serve practical, product-oriented breeding goals, they should not be considered tools that inherently overcome the trade-off between yield and seed composition.

4.4 | Methodological Implications for Deviation-Type Indices

One of the key findings of this study is that the calculation method used for deviation-type indices affects their phenotypic behaviour and downstream genetic interpretation. In the present dataset, the plot-based and BLUE-based approaches produced highly similar results in terms of both phenotypic correlations (Figure 1) and GWAS signals (Table 3). This suggests that adjusting genotypes for design and environmental effects before regression did not fundamentally alter the information captured in the individual environments studied here.

In contrast, the deviation index derived from the bivariate mixed model differed more clearly from the other two regression-based indices. This was reflected in the altered correlations with SY, the smaller effect sizes observed for shared loci and, most notably, the additional signal detected on chromosome 18 for GPD. As this phenotype was derived from the estimated genetic covariance structure between SY and PC, it probably captured a different aspect of the trait relationship to the purely phenotypic, residual-based approaches. This finding demonstrates that the construction of indices is not merely a technical detail. When antagonistically related traits are genetically studied, the statistical definition of the index can influence the detection of loci.

This is perhaps the most important positive result regarding deviation-type indices in the present study. Although the indices were often of limited additional value for direct phenotypic selection, the bivariate-derived trait provided extra resolution for the genetic dissection of the yield–protein relationship. This suggests that multivariate indices may be useful not only for selection but also for improving our understanding of the genomic basis of adverse trait correlations.

4.5 | Genetic Architecture of Seed Composition and Yield-Related Traits

The GWAS results confirmed the central role of chromosome 20 in determining the composition of soybean seeds. Strong and overlapping association signals for both protein- and oil-related traits were observed in this region, and the major locus identified there explained a considerable proportion of the phenotypic and genotypic variance relative to the other loci detected. This is consistent with earlier reports that chromosome 20 harbours important loci affecting soybean seed composition and substantially contributing to the antagonistic relationship between PC and OC (Hwang et al. 2014). The opposite directions of SNP effects for protein- and oil-related traits further support the view that this region is involved in the biological basis of their trade-off (Brown et al. 2021; Li et al. 2024).

However, not all previously reported major loci were recovered in our material. *GmSWEET39*, for example, has been described as a gene with pleiotropic effects on both PC and OC (H. Zhang et al. 2020), but it was not detected in this population. This is not unexpected as the genetic architecture of these traits varies across populations, and the ability to detect specific loci depends on allele frequencies, linkage disequilibrium structure, statistical power and environmental context.

The seed-yield associations detected on chromosome 13 were largely separate from the major chromosome 20 composition signals. This is consistent with previous studies that have

suggested that at least some of the genetic architecture of SY is independent of the major loci that control seed composition (T. Assefa, Otyama, et al. 2019; Ayalew et al. 2022; Priyanatha et al. 2022). This partial independence is important for breeding as it suggests that unfavourable phenotypic correlations do not necessarily imply complete genetic overlap among the underlying causal loci. Nevertheless, the relatively modest variance explained by most loci in our study also highlights that SY and seed composition are complex, polygenic traits influenced by many loci of small to moderate effect, as well as substantial environmental variation.

The association identified chromosome 18 for the bivariate-derived GPD phenotype did not coincide exactly with loci reported in previous studies. However, earlier studies indicate that chromosome 18 contains genomic regions relevant to both soybean seed composition and yield-related traits. A recent GWAS detected the protein-content locus qPC18.1 on chromosome 18 across multiple environments (Jin et al. 2023). Meanwhile, Yang et al. (2022) reported a nearby region at 55.53 Mb that is associated with plant architecture and seed weight per plant. Although these earlier findings do not validate our specific region, they do support the biological plausibility of chromosome 18 as a region that contributes to the relationship between SY and PC. Although these interpretations are still speculative, the results emphasize the potential value of covariance-based indices in revealing further components of complex trait architecture.

4.6 | Biological Interpretation of the Observed Trade-Offs

The negative relationships between SY, PC and OC probably result from a combination of developmental, physiological and genetic constraints. Protein accumulation requires nitrogen assimilation and storage, whereas oil synthesis depends heavily on the supply of carbon and the activity of lipid biosynthetic pathways (Kambhampati et al. 2019). As developing seeds have limited metabolic capacity and assimilate availability, increases in one storage fraction may be associated with decreases in another (Bianculli et al. 2016). Furthermore, total SY depends on processes such as canopy development, source strength, seed number and filling duration of the seeds, all of which are influenced by environmental factors and may not align with the physiological requirements for increased protein concentration (Liu et al. 2023).

Our results fit this general framework. The strong negative relationship between PC and OC, as well as the overlap of their major loci, suggests substantial antagonistic pleiotropy or closely linked genetic effects. In contrast, the weaker relationship between SY and PC implies that this trade-off is present, but less deterministic and more susceptible to modification by environmental conditions and population-specific allele combinations. This may explain why deviation-type indices provided only limited additional phenotypic discrimination in the present material. When the underlying relationship is relatively weak and unstable, the derived deviation phenotype remains dominated by the quality trait itself.

4.7 | Implications for Soybean Breeding

From a breeding perspective, the main practical implication of this study is that deviation-type indices should be used with caution. In the soybean populations and environments analysed here, they did not generally outperform direct selection for the corresponding quality traits. Therefore, their usefulness appears to be restricted to breeding situations in which the simultaneous improvement of yield and quality is clearly defined as the objective, and where the underlying trait relationships are strong and stable enough to make the deviation concept informative.

This does not mean that the indices are without value. They may still be useful as analytical tools for identifying rare genotypes with favourable trait combinations, for comparing breeding material across selection cycles or for constructing covariance-aware traits for genetic analysis. Additionally, the bivariate-derived GPD trait indicates that multivariate approaches could enhance the identification of loci relevant to yield–quality trade-offs. These approaches could be particularly relevant in genomic prediction or in breeding programmes targeting combined performance rather than individual traits in isolation. However, our results suggest that further research is needed before broad practical recommendations can be made.

4.8 | Limitations and Outlook

Several limitations of the present study should be considered. Firstly, the usefulness of deviation-type indices is expected to depend on the specific population, trait architecture and target environments. Therefore, the conclusions drawn here may not be applicable to all soybean breeding materials. Secondly, because SY was strongly influenced by environmental conditions, the stability of deviation-type indices across environments was limited. Thirdly, the genomic regions identified in this study represent associated intervals rather than confirmed causal loci, and further validation will be necessary before they can be used in breeding programmes.

Future work could build on these findings in several ways. It would be valuable to test deviation-type indices in additional soybean populations exhibiting stronger yield-protein antagonism to evaluate their stability more explicitly across environments and to compare them directly with other multi-trait selection strategies. Additionally, multivariate genomic prediction approaches could be promising, particularly in breeding stages where direct phenotypic assessment of yield remains limited. Finally, investigating loci such as the chromosome 18 region identified in this study could improve our understanding of the biological basis of yield–quality trade-offs in soybean.

Author Contributions

Vincent Braun: data analysis, data curation, visualization, writing – original draft, writing – review and editing, literature research, conceptualization. **Clemens Hacke:** data analysis, data curation, visualization, writing – original draft, writing – review and editing, providing genotypic data, literature research. **Tobias Würschum:** writing – review and editing. **Karl Schmid:** writing – review and

editing, providing genotypic data, funding acquisition. **Volker Hahn:** funding acquisition, project administration, collection of phenotyping data. **Patrick Thorwarth:** data analysis, conceptualization, writing – review and editing, supervision.

Acknowledgements

Open Access funding enabled and organized by Projekt DEAL.

Funding

This paper was written as part of the project "SENSOJA-Sensor-assisted breeding of high-performance soybean varieties with increased tolerance to abiotic stresses" (FKZ 2820EPS006). The project is supported by funds of the German Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the Protein Crop Strategy.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data underlying this study are available from the corresponding author upon reasonable request.

References

Assefa, T., P. I. Otyama, A. V. Brown, S. R. Kalberer, R. S. Kulkarni, and S. B. Cannon. 2019. "Genome-Wide Associations and Epistatic Interactions for Internode Number, Plant Height, Seed Weight and Seed Yield in Soybean." *BMC Genomics* 20, no. 1: 527. https://doi.org/10.1186/s12864-019-5907-7.

Assefa, Y., L. C. Purcell, M. Salmeron, et al. 2019. "Assessing Variation in US Soybean Seed Composition (Protein and Oil)." *Frontiers in Plant Science* 10: 298. https://doi.org/10.3389/fpls.2019.00298.

Ayalew, H., W. Schapaugh, T. Vuong, and H. T. Nguyen. 2022. "Genome-Wide Association Analysis Identified Consistent QTL for Seed Yield in a Soybean Diversity Panel Tested Across Multiple Environments." *Plant Genome* 15, no. 4: e20268. https://doi.org/10.1002/tpg2.20268.

Bernal-Vasquez, A.-M., H.-F. Utz, and H.-P. Piepho. 2016. "Outlier Detection Methods for Generalized Lattices: A Case Study on the Transition From ANOVA to REML." *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik* 129, no. 4: 787–804. https://doi.org/10.1007/s00122-016-2666-6.

Bianculli, M. L., L. A. Aguirrezábal, G. A. Pereyra Irujo, and M. M. Echarte. 2016. "Contribution of Incident Solar Radiation on Leaves and Pods to Soybean Seed Weight and Composition." *European Journal of Agronomy* 77: 1–9. https://doi.org/10.1016/j.eja.2016.03.002.

Brown, A. V., S. I. Conners, W. Huang, et al. 2021. "A New Decade and New Data at SoyBase, the USDA-ARS Soybean Genetics and Genomics Database." *Nucleic Acids Research* 49, no. D1: D1496–D1501. https://doi.org/10.1093/nar/gkaa1107.

Browning, S. R. 2021. "Fast Two-Stage Phasing of Large-Scale Sequence Data." *American Journal of Human Genetics* 108, no. 10: 1880–1890.

Butler, D. 2023. "ASReml: Fits the Linear Mixed Model." www.vsni.co.uk.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4: 7. https://doi.org/10.1186/s13742-015-0047-8.

Cober, E. R., and H. D. Voldeng. 2000. "Developing High-Protein, High-Yield Soybean Populations and Lines." *Crop Science* 40, no. 1: 39–42. https://doi.org/10.2135/cropsci2000.40139x.

Cullis, B. R., A. B. Smith, and N. E. Coombes. 2006. "On the Design of Early Generation Variety Trials With Correlated Data." *Journal of Agricultural, Biological, and Environmental Statistics* 11, no. 4: 381–393. https://doi.org/10.1198/108571106X154443.

Degenhardt, R. F., and P. C. Bonham-Smith. 2008. "Arabidopsis Ribosomal Proteins RPL23aA and RPL23aB Are Differentially Targeted to the Nucleolus and Are Disparately Required for Normal Development." *Plant Physiology* 147, no. 1: 128–142. https://doi.org/10.1104/pp.107.111799.

Diers, B. W., J. E. Specht, G. L. Graef, et al. 2023. "Genetic Architecture of Protein and Oil Content in Soybean Seed and Meal." *Plant Genome* 16, no. 1: e20308. https://doi.org/10.1002/tpg2.20308.

Duhnen, A., A. Gras, S. Teyssèdre, et al. 2017. "Genomic Selection for Yield and Seed Protein Content in Soybean: A Study of Breeding Program Data and Assessment of Prediction Accuracy." *Crop Science* 57, no. 3: 1325–1337. https://doi.org/10.2135/cropsci2016.06.0496.

Eurostat. 2025. "Rape, Turnip Rape, Sunflower Seeds and Soya by Area." Eurostat. https://doi.org/10.2908/TAG00100.

Fogelberg, F., and J. Recknagel. 2017. "Developing Soy Production in Central and Northern Europe." In *Legumes in Cropping Systems*, 109–124. CAB International. https://doi.org/10.1079/9781780644981.0109.

Hahn, V., and T. Würschum. 2014. "Molecular Genetic Characterization of Central European Soybean Breeding Germplasm." *Plant Breeding* 133, no. 6: 748–755. https://doi.org/10.1111/pbr.12212.

Häusling, M. 2011. "The EU Protein Deficit: What Solution for a Long-Standing Problem?(2010/2111 (INI))." Brussels, Belgium.

Hazel, L. N. 1943. "The Genetic Basis for Constructing Selection Indexes." *Genetics* 28, no. 6: 476–490. https://doi.org/10.1093/genetics/28.6.476.

Helms, T. C., and J. H. Orf. 1998. "Protein, Oil, and Yield of Soybean Lines Selected for Increased Protein." *Crop Science* 38, no. 3: 707–711. https://doi.org/10.2135/cropsci1998.0011183X003800030015x.

Hou, G., G. R. Ablett, K. P. Pauls, and I. Rajcan. 2006. "Environmental Effects on Fatty Acid Levels in Soybean Seed Oil." *Journal of the American Oil Chemists' Society* 83, no. 9: 759–763. https://doi.org/10.1007/s11746-006-5011-4.

Hwang, E.-Y., Q. Song, G. Jia, et al. 2014. "A Genome-Wide Association Study of Seed Protein and Oil Content in Soybean." *BMC Genomics* 15, no. 1: 1. https://doi.org/10.1186/1471-2164-15-1.

Jin, H., X. Yang, H. Zhao, et al. 2023. "Genetic Analysis of Protein Content and Oil Content in Soybean by Genome-Wide Association Study." *Frontiers in Plant Science* 14: 1182771. https://doi.org/10.3389/fpls.2023.1182771.

Johnson, H. W., H. F. Robinson, and R. E. Comstock. 1955. "Genotypic and Phenotypic Correlations in Soybeans and Their Implications in Selection." *Agronomy Journal* 47, no. 10: 477–483. https://doi.org/10.2134/agronj1955.00021962004700100008x.

Kambhampati, S., J. A. Aznar-Moreno, C. Hostetler, et al. 2019. "On the Inverse Correlation of Protein and Oil: Examining the Effects of Altered Central Carbon Metabolism on Seed Composition Using Soybean Fast Neutron Mutants." *Metabolites* 10, no. 1: 18. https://doi.org/10.3390/metabo10010018.

Koekemoer, F. P., M. T. Labuschagne, and C. S. van Deventer. 1999. "A Selection Strategy for Combining High Grain Yield and High Protein Content in South African Wheat Cultivars." *Cereal Research Communications* 27, no. 1–2: 107–114. https://doi.org/10.1007/BF03543926.

Kurasch, A. K., V. Hahn, W. L. Leiser, N. Starck, and T. Würschum. 2017. "Phenotypic Analysis of Major Agronomic Traits in 1008 RILs From a Diallel of Early European Soybean Varieties." *Crop Science* 57, no. 2: 726–738. https://doi.org/10.2135/cropsci2016.05.0318.

Latshaw, S. P., M. F. Vigil, and S. D. Haley. 2016. "Genotypic Differences for Nitrogen Use Efficiency and Grain Protein Deviation in Hard Winter Wheat." *Agronomy Journal* 108, no. 6: 2201–2213. https://doi.org/10.2134/agronj2016.02.0070.

Li, H., J. Sun, Y. Zhang, et al. 2024. "Soybean Oil and Protein: Biosynthesis, Regulation and Strategies for Genetic Improvement." *Plant, Cell and Environment* 49: 3730–3746. https://doi.org/10.1111/pce.15272.

Liu, S., Z. Liu, X. Hou, and X. Li. 2023. "Genetic Mapping and Functional Genomics of Soybean Seed Protein." *Molecular Breeding: New Strategies in Plant Improvement* 43, no. 4: 29. https://doi.org/10.1007/s11032-023-01373-5.

Markakis, M. N., A. K. Boron, B. van Loock, et al. 2013. "Characterization of a Small Auxin-Up RNA (SAUR)-Like Gene Involved in *Arabidopsis thaliana* Development." *PLoS ONE* 8, no. 11: e82596. https://doi.org/10.1371/journal.pone.0082596.

Mbatchou, J., L. Barnard, J. Backman, et al. 2021. "Computationally Efficient Whole-Genome Regression for Quantitative and Binary Traits." *Nature Genetics* 53, no. 7: 1097–1103. https://doi.org/10.1038/s41588-021-00870-7.

Monaghan, J. M., J. W. Snape, A. J. S. Chojecki, and P. S. Kettlewell. 2001. "The Use of Grain Protein Deviation for Identifying Wheat Cultivars With High Grain Protein Concentration and Yield." *Euphytica* 122, no. 2: 309–317. https://doi.org/10.1023/A:1012961703208.

Neuweiler, J. E., H. P. Maurer, and T. Würschum. 2021. "Genetic Architecture of Phenotypic Indices for Simultaneous Improvement of Protein Content and Grain Yield in Triticale (×*Triticosecale*)." *Plant Breeding* 140, no. 2: 232–245. https://doi.org/10.1111/pbr.12894.

Panthee, D. R. 2010. "Varietal Improvement in Soybean." In *The Soybean: Botany, Production and Uses*, edited by G. Singh, 92–112. CABI. https://doi.org/10.1079/9781845936440.0092.

Priyanatha, C., D. Torkamaneh, and I. Rajcan. 2022. "Genome-Wide Association Study of Soybean Germplasm Derived From Canadian × Chinese Crosses to Mine for Novel Alleles to Improve Seed Yield and Seed Quality Traits." *Frontiers in Plant Science* 13: 866300. https://doi.org/10.3389/fpls.2022.866300.

Purcell, S., and C. Chang. 2015. "PLINK [1.9]".

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." Vienna, Austria. Retrieved from https://www.R-project.org/.

Rapp, M., V. Lein, F. Lacoudre, et al. 2018. "Simultaneous Improvement of Grain Yield and Protein Content in Durum Wheat by Different Phenotypic Indices and Genomic Selection." *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik* 131, no. 6: 1315–1329. https://doi.org/10.1007/s00122-018-3080-z.

Recker, J. R., J. W. Burton, A. Cardinal, and L. Miranda. 2014. "Genetic and Phenotypic Correlations of Quantitative Traits in Two Long-Term, Randomly Mated Soybean Populations." *Crop Science* 54, no. 3: 939–943. https://doi.org/10.2135/cropsci2013.07.0447.

Sato, T., M. van Schoote, H. Wagentristl, and J. Vollmann. 2014. "Effects of Divergent Selection for Seed Protein Content in High-Protein vs. Food-Grade Populations of Early Maturity Soybean." *Plant Breeding* 133, no. 1: 74–79. https://doi.org/10.1111/pbr.12138.

Schmidt, P., J. Hartung, J. Rath, and H.-P. Piepho. 2019. "Estimating Broad-Sense Heritability With Unbalanced Data From Agricultural Cultivar Trials." *Crop Science* 59, no. 2: 525–536. https://doi.org/10.2135/cropsci2018.06.0376.

Sedibe, M. M., A. M. Mofokeng, and D. R. Masvodza. 2023. "Soybean Production, Constraints, and Future Prospects in Poorer Countries: A Review." In *Production and Utilization of Legumes—Progress and Prospects*, edited by M. Hasanuzzaman. IntechOpen. https://doi.org/10.5772/intechopen.109516.

Shea, Z., W. M. Singer, and B. Zhang. 2020. "Soybean Production, Versatility, and Improvement." In *Legume Crops*. IntechOpen. https://doi.org/10.5772/intechopen.91778.

Shi, Q., W. Mo, X. Zheng, et al. 2025. "Balancing Act: Progress and Prospects in Breeding Soybean Varieties With High Oil and Seed Protein Content." *Frontiers in Plant Science* 16: 1560845. https://doi.org/10.3389/fpls.2025.1560845.

Smith, H. F. 1936. "A Discriminant Function for Plant Selection." *Annals of Eugenics* 7, no. 3: 240–250. https://doi.org/10.1111/j.1469-1809.1936.tb02143.x.

Sobko, O., A. Stahl, V. Hahn, S. Zikeli, W. Claupein, and S. Gruber. 2020. "Environmental Effects on Soybean (*Glycine max* (L.) Merr) Production in Central and South Germany." *Agronomy* 10, no. 12: 1847. https://doi.org/10.3390/agronomy10121847.

Stram, D. O., and J. W. Lee. 1994. "Variance Components Testing in the Longitudinal Mixed Effects Model." *Biometrics* 50, no. 4: 1171–1177. https://doi.org/10.2307/2533455.

Thorwarth, P., H. P. Piepho, Y. Zhao, et al. 2018. "Higher Grain Yield and Higher Grain Protein Deviation Underline the Potential of Hybrid Wheat for a Sustainable Agriculture." *Plant Breeding* 137, no. 3: 326–337. https://doi.org/10.1111/pbr.12588.

Turner, S. 2018. "qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots." *Journal of Open Source Software* 3, no. 25: 731. https://doi.org/10.21105/joss.00731.

Valliyodan, B., S. B. Cannon, P. E. Bayer, et al. 2019. "Construction and Comparison of Three Reference-Quality Genome Assemblies for Soybean." *Plant Journal* 100, no. 5: 1066–1082. https://doi.org/10.1111/tpj.14500.

Vollmann, J., C. N. Fritz, H. Wagentristl, and P. Ruckenbauer. 2000. "Environmental and Genetic Variation of Soybean Seed Protein Content Under Central European Growing Conditions." *Journal of the Science of Food and Agriculture* 80, no. 9: 1300–1306. https://doi.org/10.1002/1097-0010(200007)80:9<1300::AID-JSFA640>3.0.CO;2-I.

Wijewardana, C., K. R. Reddy, F. A. Alsajri, J. T. Irby, J. Krutz, and B. Golden. 2018. "Quantifying Soil Moisture Deficit Effects on Soybean Yield and Yield Component Distribution Patterns." *Irrigation Science* 36, no. 4–5: 241–255. https://doi.org/10.1007/s00271-018-0580-1.

Wilcox, J. R. 1998. "Increasing Seed Protein in Soybean With Eight Cycles of Recurrent Selection." *Crop Science* 38, no. 6: 1536–1540. https://doi.org/10.2135/cropsci1998.0011183X003800060021x.

Wilcox, J. R., and J. F. Cavins. 1995. "Backcrossing High Seed Protein to a Soybean Cultivar." *Crop Science* 35, no. 4: 1036–1041. https://doi.org/10.2135/cropsci1995.0011183X003500040019x.

Yang, C., J. Yan, S. Jiang, et al. 2022. "Resequencing 250 Soybean Accessions: New Insights Into Genes Associated With Agronomic Traits and Genetic Networks." *Genomics, Proteomics & Bioinformatics* 20, no. 1: 29–41. https://doi.org/10.1016/j.gpb.2021.02.009.

Yu, Z., B. Hu, H. Ning, and W.-X. Li. 2025. "Detection of Genes Associated With Soybean Protein Content Using a Genome-Wide Association Study." *Plant Molecular Biology* 115, no. 2: 49. https://doi.org/10.1007/s11103-025-01576-8.

Zhang, C., S.-S. Dong, J.-Y. Xu, W.-M. He, and T.-L. Yang. 2019. "PopLDdecay: A Fast and Effective Tool for Linkage Disequilibrium Decay Analysis Based on Variant Call Format Files." *Bioinformatics (Oxford, England)* 35, no. 10: 1786–1788. https://doi.org/10.1093/bioinformatics/bty875.

Zhang, H., W. Goettel, Q. Song, et al. 2020. "Selection of GmSWEET39 for Oil and Protein Improvement in Soybean." *PLoS Genetics* 16: e1009114.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Years, locations, trials which were planted, raw plot correlations between seed yield (SY) and the traits protein content (PC), oil content (OC) and protein and OC (POC) and OC and PC and temperature and precipitation sum from begin of april till end of september in each environment (year × location combination). **Table S2:** Coordinates and altitudes of all 13 locations. **Table S3:** Explained phenotypic and genotypic variance for each trait of the most significant SNPs within each locus defined based on LD decay. **Table S4:** Genes types and families identified for the loci of the GWAS results. **Figure S1:** Comparison of the effect size of shared significant GWAS SNPs between PC and OC. **Figure S2:** Manhattan and qq-plot for the PY GWAS. **Figure S3:** Manhattan and qq-plot for the OY GWAS. **Figure S4:** Manhattan and qq-plot for the POY GWAS. **Figure S5:** Manhattan and qq-plot for the GPD_{bivar} GWAS. **Figure S6:** Manhattan and qq-plot for the GOD_{bivar} GWAS. **Figure S7:** Manhattan and qq-plot for the GCD GWAS. **Figure S8:** Manhattan and qq-plot for the GCD_{blu} GWAS. **Figure S9:** Manhattan and qq-plot for the GCD_{bivar} GWAS. **Figure S10:** Heatmap comparing the effect size of shared significant GWAS SNPs between the methods of calculating GPD. **Figure S11:** Heatmap comparing the effect size of shared significant GWAS SNPs between the methods of calculating GOD. **Figure S12:** Heatmap comparing the effect size of shared significant GWAS SNPs between the methods of calculating GCD. **Figure S13:** Scatter plot of the identified loci across their positions on chromosomes 6, 13, 18 and 20. Highlighting the distribution of the significant loci across all traits.