












# SNP marker association for incrementing soybean seed protein content

Arthur Bernardeli<sup>1,\*</sup>, Aluizio Borém<sup>1</sup>, Rodrigo Lorenzoni<sup>2</sup>, Rafael Aguiar<sup>2</sup>, Jéssica Nayara Basilio Silva<sup>2</sup>, Rafael Delmond Bueno<sup>2</sup>, Cléberon Ribeiro<sup>3</sup>, Newton Piovesan<sup>4</sup> and Maximiller Dal-Bianco Lamas Costa<sup>2</sup>

<sup>1</sup>Department of Agronomy, Universidade Federal de Viçosa, Av. PH Rolfs, s/n, Viçosa-MG, 36570900, Brazil. <sup>2</sup>Department of Biochemistry and Molecular Biology, Universidade Federal de Viçosa, Av. PH Rolfs, s/n, Viçosa-MG, 36570900, Brazil. <sup>3</sup>Department of Biology, Universidade Federal de Viçosa, Av. PH Rolfs, s/n, Viçosa-MG, 36570900, Brazil. <sup>4</sup>BioAgro, Universidade Federal de Viçosa, Av. PH Rolfs, s/n, Viçosa-MG, 36570900, Brazil. \*Corresponding author, E-mail: arthurbernardeli@gmail.com

## ABSTRACT

Soybean seed protein content (SPC) has been decreasing throughout last decades and DNA marker association has shown its usefulness to improve this trait even in soybean breeding programs that focus primarily on soybean yield and seed oil content (SOC). Aiming to elucidate the association of two SNP markers (ss715630650 and ss715636852) to the SPC, a soybean population of 264  $F_5$ -derived recombinant inbred lines (RILs) from a bi-parental cross was tested in four environments. Through the single-marker analysis, the additive effect ( $a$ ) and the portion of SPC variation due to the SNPs ( $r^2$ ) for single and multi-environment data were assessed, and transgressive RILs for SPC were observed. The estimates revealed the association of both markers to SPC in most of environments. The marker ss715636852 was more frequently associated to SPC, including multi-environment data, and contributed up to  $a = 1.30\%$  for overall SPC, whereas ss715630650 had significant association just in two locations, with contributions of  $a = 0.76\%$  and  $a = 0.74\%$  to overall SPC in Vic1 and Cap1, respectively. The RILs 84-13 was classified as an elite genotype due to its favorable alleles and high SPC means, which reached 53.78% in Cap1, and 46.33% in MET analysis. Thus, these results confirm the usefulness of the SNP marker ss715636852 in a soybean breeding program for SPC.

**Keywords:** additive effect, allelic polymorphism, favorable alleles, *Glycine max*, quantitative trait loci, transgressive genotypes.

## INTRODUCTION

Soybean [*Glycine max* (L.) Merr.] is a major crop widely cultivated worldwide, and its importance is mainly assigned to its seed protein content (SPC), denoting the relevance of this crop for human and animal nutrition, as well as economy and world food security. SPC is quantitatively inherited and correlates negatively with most of the main traits taken into account in a soybean-breeding program (Kwon & Torrie, 1964). It is complex to elevate the percentages of SPC, once the selection towards grain yield and seed oil content (SOC) has been prioritized when compared to protein increment (Bandillo et al., 2015; Patil et al., 2018). For example, SPC of ancestral soybean cultivars and early releases declined from 40% to 37% during the period of 1924 to 2004 at the United States, whereas grain yield was incremented steadily during this same time (Mahmoud et al., 2006).

Promising methodologies are being proposed to increase soybean SPC without compromising grain yield, and efforts concentrate in gathering adequate specific allelic combinations for performing marker assisted selection. By using the marker information within or in proximity to important protein-related QTLs (Quantitative Trait Loci), strategies involving the validation of molecular markers in structured populations have been combined with traditional breeding methods in order to deliver more rapid genetic gains in many soybean traits (Jun et al., 2008).

In view of the soybean protein demand, several markers have been described to have an important effect over SPC on the 20 soybean chromosomes (<https://www.soybase.org/>). Zhang et al. (2015) validated SSR (simple sequence repeats) markers followed by two cycles of marker-assisted selection, where parent lines were outperformed in 9% regarding SPC of the selected progenies. Rodrigues et al. (2010) performed a single-marker association analysis in two SSR markers located at D1a soybean linkage group, where 5.57%

of the SPC phenotypic variation was attributed to the marker *satt408*.

Gains in SPC are complex for soybean breeding programs and fundamental to supply the global demand for vegetable protein. Therefore, this study aimed to evaluate the effects of two single nucleotide polymorphisms (SNPs) over soybean SPC in a single and multi-environment (MET) approach.

## MATERIALS AND METHODS

### Plant material and field trials

In this study, 264  $F_5$ -derived soybean recombinant inbred lines (RILs) were obtained from a single cross between PMQS12 and PMQS80. The parent material belonged to BioAgro soybean breeding program germplasm of Universidade Federal de Viçosa. SPC of PMQS12 and PMQS80 averaged 45.7 and 46.4% under greenhouse conditions, respectively. The population was advanced through the single seed descent (SSD) method under greenhouse environment up to  $F_5$  generation and their derived seeds were consistently used for all phenotyping trials. The RILs population, parent lines and the checks ANSC83022, M7739 and M8372 were tested in four environments in southeastern Brazil, Viçosa (20°45'14"S - 42°52'55"O, 649 m of altitude), and Capinópolis (18°40'55"S - 49°34'12"O, 530 m of altitude), both at Minas Gerais state, in 2017 and 2018 crop years. The field trials were set under randomized complete block design with two replicates. Each plot consisted in a 1m row spaced 0.5m apart, with 15 plants as final plant density per plot. The crop management from planting to harvest was performed in accordance to Sedyama et al. (2015).

The soil of the experimental sites comprised red-yellow dystrophic latosol with clayey texture, and dark red eutrophic latosol with medium texture, at Viçosa and Capinópolis, respectively (Santana & Moura Filho, 1978). The climatic conditions within the experimental seasons (i.e. average temperatures and rainfall) for both locations can be checked at <http://www.inmet.gov.br/>. The environments Viçosa in 2017, Viçosa in 2018, Capinópolis in 2017, and Capinópolis in 2018 were referred as Cap1, Cap2, Vic1 and Vic2, respectively.

### Phenotyping

After manual harvesting of all plants per plot, a 30g random seed sample was collected, milled and the SPC was assessed through a near infrared spectrometry equipment (Thermo Fisher Antaris II FT-NIR) similarly to the studies by Rodrigues et al. (2014). The SPC values were converted to dry basis, as it follows:

$$SPC_{\%} = \frac{100 \times SPC_{\%}}{100 - moisture_{\%}}$$

### Genotyping

A leaf disk was sampled on a single plant at V4 stage of every RIL, and the DNA was extracted according to Dellaporta et al. (1983). After nucleic acid quantification by NanoDrop spectrophotometer, the DNA was diluted to 10  $ng \cdot \mu L^{-1}$  for SNP genotyping. SNPs located in proximity to SSRs, which were previously used in the BioAgro-UFV program for marker assisted SPC breeding, were chosen. Those polymorphic between PMQS12 and PMQS80 were selected to carry out this study (Table 1). The SNP genotyping followed the KASP methodology developed by Biosearch Technology (<https://www.biosearchtech.com>), and was performed using the Applied Biosciences 7500 equipment. The amplification reaction comprised 1 cycle of 94 °C for 15 minutes; 10 cycles of 94 °C for 20 seconds, with a gradient of 61-55 °C, decreasing 0.6 °C every 60 seconds; 30 cycles of 94 °C for 20 seconds and 55 °C for 60 seconds; and a 37 °C cycle for 60 seconds. Each reaction consisted of 2.5  $\mu L$  of DNA at 10  $ng \cdot \mu L^{-1}$ , 2.5  $\mu L$  of 2x Master Mix and 0.14  $\mu L$  of Primer Mix. Allelic discrimination was performed in the AB 7500 v. 2.3 software.

### Statistics

From the phenotypic data, the individual and MET analyses were performed according to the following models:

$$y = \mu + Z_1g + Z_2b + e, \quad [\text{model 1 for individual analysis}]$$

**Table 1.** Summary of SNP markers used for the association study based on the information available at <https://www.soybase.org/>. These SNPs are part of the SoySNP50K array (Song et al., 2013).

SNP marker	Chromosome	Linkage group	Position (bp) <sup>1</sup>	SNP alleles <sup>2</sup>	Close QTLs	Previous researches
ss715630650 (56)	18	G	41887139 (41887079-41887199)	G/A	Seed protein 26-8	Reinprecht et al. (2006), Rodrigues et al. (2010).
ss715636852 (62)	20	I	1897580 (1897519-1897640)	A/G	Seedprotein 3-12 Seedprotein 10-1	Brummer et al. (1997), Sebolt et al. (2000), Rodrigues et al. (2010).

<sup>1</sup>*Glycine max* genome assembly version Glyma.Wm82.a2 (Gmax2.0) map version 4.0.

<sup>2</sup>The former base at the left of the dash represents the predominant allele in the mapping populations.

where  $y$  is the vector of observed data within environment;  $\mu$  is the mean;  $g$  is the random effect of genotypes (RILs),  $g \sim \text{NID}(0, \sigma_g^2)$ , where  $\sigma_g^2$  is the genetic variance;  $b$  is the random vector of blocks,  $b \sim \text{NID}(0, \sigma_b^2)$ , where  $\sigma_b^2$  is the block variance.  $Z_1$  and  $Z_2$  are the design matrices for  $g$  and  $b$ , respectively; and

$$y = \mu + Xt + Z_1g + Z_2ge + e, \text{ [model 2 for MET analysis]}$$

where  $y$  is the vector of observed data for the  $t$  trials;  $\mu$  is the mean;  $t$  is the fixed vector of trials;  $g$  is the random effect of genotypes (RILs),  $g \sim \text{NID}(0, \sigma_g^2)$ , where  $\sigma_g^2$  is the genetic variance;  $ge$  is the random effect of genotype by environment interaction,  $ge \sim \text{NID}(0, \sigma_{ge}^2)$ , where  $\sigma_{ge}^2$  is the variance of genotype by environment interaction; and  $e$  is the random vector of residuals,  $e \sim \text{NID}(0, \sigma^2)$ , where  $\sigma^2$  is the residual variance.  $X$ ,  $Z_1$  and  $Z_2$  are the design matrices for  $t$ ,  $g$  and  $ge$ , respectively. From both models, broad sense heritability was assessed as  $h_g^2 = 1 - \frac{\bar{v}_{BLUP}}{2\sigma_g^2}$ , where  $\bar{v}_{BLUP}$  is the average variance of pairwise differences between the best linear unbiased predictions (BLUPs) of  $g$  effects. SPC generated from RILs, parent material and checks were compared. The marker analysis was based on the single-marker association (Schuster & Cruz, 2008) through the following linear regression model:

$$y_j = \beta_0 + \beta_1 X_{1j} + \varepsilon_j,$$

where  $y_j$  is the predicted mean for the genotype  $j$  from individual and MET data (models 1 and 2);  $\beta_0$  is the intercept of regression or the mean,  $\beta_1$  is the marker additive effect, and  $\varepsilon_j$  is the random vector or residuals.  $X_{1j}$  is the design matrix for coding genotypes as  $A_1A_1 = 2$ ,  $A_1A_2 = 1$ ,  $A_2A_2 = 0$ . The coefficient of determination ( $r^2$ ) obtained from the regression analysis denotes the proportion of the SPC variance due to SNP. The additive effect ( $a$ ) of an associated marker was obtained as the difference between the average of the individuals with the favorable allele and the average of the individuals with the unfavorable allele. The genetic effects were tested by the Likelihood Ratio Test (LRT; Rao, 1973) and the normal distribution of the data was verified through Lilliefors test (Razali & Wah, 2011). The precision of the experiments was assessed in concordance with Rodrigues et al. (2010), by using the coefficient of variation, presented in percentages, as it follows:

$$CV\% = \frac{\sigma}{\bar{Y}_i} \cdot 100,$$

where  $\sigma$  represents the standard deviation of residuals for each experiment, and  $\bar{Y}_i$  as the average calculated from the observed values for single-environment analyses or average from the predicted genetic

values from multi-environment analysis. The statistical analysis were performed at GENES (Cruz, 2013) and R (R Core Team, 2019) softwares.

## RESULTS AND DISCUSSION

### Soybean SPC field data

Figure 1 shows that SPC followed a normal distribution for all single-environment and MET analyses ( $p > 0.05$ ), presenting frequencies of the data according to the expected pattern for a quantitative inherited trait. The occurrence of genotypes with SPC values above 50% and below 40% was observed. However, the RILs average SPC are barely the same as their parent material. Cap1 and Cap2 presented the highest values for protein overall, being 47.80 and 47.60% respectively. In order to compare the environments, the SPC means of 10% superior RILs were 52.84, 51.49, 49.39, 48.63, and 47.84% for Cap 1, Cap2, Vic1, Vic2 and MET data, respectively. Cap1 also presented the highest frequency (18.18%) of transgressive RILs above 50% of SPC, in contrary to Vic2 and MET data that presented 1.13 and 0%.

The means of parent lines and RILs outperformed the checks in all trials, as previously expected. The LRT analysis showed significant differences ( $p < 0.01$ ) for the variance component of genotype effect ( $\hat{\sigma}_g^2$ ) for both single environment and MET data (Table 2). The broad sense heritability values were similar and ranged from 70.79 to 76.82% for individual trial data, with the superior and inferior estimates for Vic1 and Cap2, respectively. For the MET analysis, broad heritability estimate was moderate, corresponding to 53%. All  $CV_{\%}$  values were below 10% for all field trials, similar to the results obtained by Rodrigues et al. (2010) in a study of QTL mapping for seed protein content in soybean.

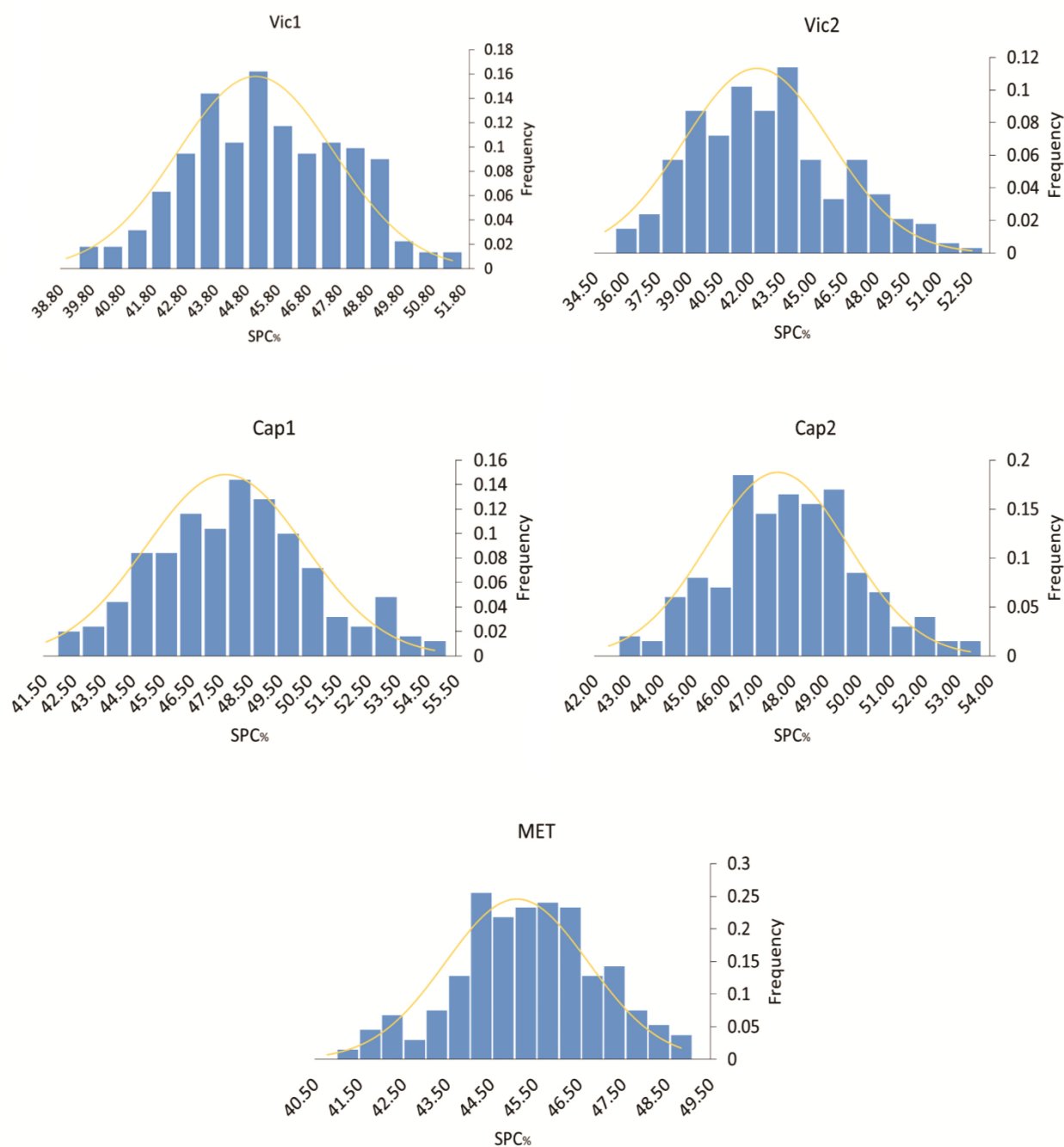
The data analyses denote a good field data quality, and reveal a satisfactory precision in terms of  $CV_{\%}$  and normality of data, as well as in the findings by Rodrigues et al. (2010) and Rodrigues et al. (2014). The  $F_5$ -derived RILs population used in this study had enough recombination cycles that resulted in transgressive genotypes for both superior and inferior SPC means, especially towards superior ones for Capinópolis location. Genotypes with SPC means above 50% were more frequent in Cap1 and Cap2 environments. Viçosa and Capinópolis experimental sites contrast in terms of weather conditions. Capinópolis is warmer especially during soybean growing season and SPC is favored by elevated temperatures. This same fact has been addressed by Piper and Boote (1999) and Patil et al. (2017), where they revealed increases in soybean SPC under elevated temperatures. The weather data from these locations can be downloaded as above-mentioned in the materials and methods section.

In Vic2, the SPC transgressive segregation range observed in our research surpassed the one by Zhang et al. (2015), which ranged from 35.89% to 49.10%, irrespective to the fact of one more recombination cycle in the NJRSXG population of  $F_6$ -derived RILs in their study. The same remark is done when comparing to the study from Rodrigues et al. (2010), where SPC varied from 32.2% to 44.5%. Despite the use of contrasting SPC parent lines, Rodrigues et al. (2010) evaluated a  $F_{2:3}$  soybean population, in which the lack of recombination cycles may have costed the appearances of more transgressive genotypes. Thus, the superior transgressive genotypes means of this study suggest that PMQS12 x PMQS80  $F_5$ -derived RILs population is an elite germplasm for SPC selection. Furthermore, checks were outperformed in an average of 5.40% in all trial scenarios, which is inferior to the results reported by Warrington et al. (2015), where  $F_5$ -derived RILs mapping population exceeded the cultivars in a rate of 7.50% for SPC.

The three checks used in the present study are modern cultivars widely adopted by soybean growers in Brazil and correspond to a diverse range of maturity group. Their SPC estimates bellow to the ones from RILs population are due to the plant improvement process that prioritized higher yields and may have lost favorable alleles for SPC through breeding. The genetic correlation between SPC and yield was -0.58 in the research by Kwon and Torrie (1964). Likewise, Yesudas et al. (2013) reported a positive correlation between yield and SOC, and a negative correlation between yield and SPC. SOC means were assessed and will be mentioned briefly in the next section.

Genetic variation is responsible for the largest portion of the phenotypic variance in the population, and the significant variance estimate for genotype by environment interaction depicts the changes in performance of genotypes in terms of SPC depending on the environment (Table 2). All SPC broad sense heritability estimates were lower than the ones presented by Patil et al. (2018). In their research, SPC data from four environments were presented and the broad sense heritability values were around 90%. Instead, our individual trial means and broad sense heritability results are similar to the ones presented by Zhang et al. (2015) and fits perfectly to those from Hwang et al. (2014) that associated molecular markers with

soybean quality traits, including SPC. Our MET analysis delivered lower broad sense heritability estimates than the ones from single analyses, exactly the contrary as Zhang et al. (2015) reported in the MET data. In the present study, the genotype by environment interaction consumes the overall genetic variance and results in decreases in broad sense heritability, as described by Kang (1997).



1. Seed protein content (SPC) distribution for single-environment data (Vic1, Vic2, Cap1, and Cap2) and multi-environment data (MET).

### SNP marker association

The two markers tested in this association study showed their usefulness for increasing SPC in most of the studied environments (Table 3). The SNP marker *ss715630650* was significantly associated with SPC in Vic1 and Cap1 environments. However, *ss715636852* was associated in Cap1, Cap2 and Vic1 environments, and MET data. The total variation in SPC explained individually by these markers ranged from 1.65% (*ss715630650*) to 6.11% (*ss715636852*). The direct contribution of the additive effect of these markers to the SPC the additive effect varied from 0.74% to 1.30%, which corresponds to 1.54 and 2.77% of the overall SPC mean for *ss715630650* and *ss715636852*, respectively. The only cases in which the two markers had

significant influence over SPC was at Vic1 and Cap1 environments. The simultaneous additive effect of the favorable alleles of these markers over SPC in Vic1 and Cap1 were similar and accounted for 2.01% and 2.04, respectively.

**Table 2.** Mean, statistics and heritability of seed protein content (SPC) of F<sub>5</sub>-derived RILs obtained from a single cross between PMQS80 and PMQS12, analyzed through single-environment (Vic1, Vic2, Cap1, and Cap2) and multi-environment (MET) approaches. Ranges are included inside parenthesis.

Environment	SPC% mean and standard deviation			$\hat{\sigma}_g^2$	$\hat{\sigma}_{ge}^2$	CV%	$h_g^2$ <sup>3</sup>
	Parent lines <sup>1</sup>	Checks <sup>2</sup>	RILs				
Vic1	45.34 ± 1.94	39.93 ± 2.51	45.10 ± 2.79 (38.94 - 51.46)	4.88**	-	3.80	76.82
Vic2	41.24 ± 4.02	36.41 ± 3.83	42.10 ± 3.25 (34.93 - 52.63)	9.22**	-	5.40	76.16
Cap1	47.45 ± 1.81	43.39 ± 1.46	47.80 ± 2.94 (41.53 - 54.78)	5.43**	-	3.76	75.75
Cap2	47.16 ± 2.20	40.9 ± 1.37	47.6 ± 2.49 (42.67 - 53.40)	3.50**	-	3.42	70.79
MET	45.27 ± 0.76	40.66 ± 1.28	45.7 ± 1.29 (39.48 - 48.83)	3.42**	2.95**	5.55	53.00

<sup>1</sup>Average between PMQS12 and PMQS80 means.

<sup>2</sup>Average between ANSC83022, M7739 and M8372 means.

<sup>3</sup>Broad sense heritability in % basis.

$\hat{\sigma}_g^2$  - variance component for genotypic effect.

$\hat{\sigma}_{ge}^2$  - variance component for genotype by environment interaction

\*\* Significant at p≤0.01.

Table 4 presents the genotypes with superior and inferior SPC means when applying 2% of selection intensity, as well as their respective the alleles for *ss715630650* and *ss715636852*. Among the selected genotypes from each environment, eight appeared at least in two environments. The RIL 78-43 was selected at Vic1, Cap2 and MET data. Out of the five superior selected genotypes in each data set, at least four of them had the favorable allele for *ss715636852*.

In Cap1, all of the genotypes had that superior allele. Among the lower ones, the unfavorable allele (guanine) appearances became more frequent. For Vic1 and Cap1, where the marker *ss715630650* was associated, the unfavorable allele (guanine) was present regardless the magnitude of the means. However, not all the superior genotypes presented the two favorable alleles, in which the absence of *ss715630650* was remarkable.

Although this research involves a very low amount of associated markers, it is possible to observe the importance of the complementarity of parents to result in high SPC populations. The favorable alleles of *ss715630650* and *ss715636852* were donated by PMQS12 and PMQS80, respectively. Parent lines that donated favorable alleles with higher effects over SPC than the ones addressed in this study were used by Warrington et al. (2015). In their research, Danbaekong was the parent line that contributed with most of the favorable alleles. This line accounted for 55% of SPC variation and was able to increment 13.64% of SPC in terms of additive effect. The QTL responsible to for this occurrence is *qProt\_Gm20* at chromosome 20, located at least 35 Mbp apart of *ss715636852*. The relevance of chromosome 20 in soybean seed content traits is once again highlighted by this research, as well as described by Patil et al. (2018), who mapped haplotypes in the region comprised between 28.5 and 33.5 Mbp of chromosome 20, being considered a hotspot for protein content. This interval overlaps haplotypes identified in the QTL mapping and genome wide association studies by Bolon et al. (2010), Hwang et al. (2014), Vaughn et al. (2014), and Bandillo et al. (2015).

**Table 3.** SNP marker association estimates for seed protein content (SPC) obtained through the single marker analysis for single-environment (Vic1, Vic2, Cap1, and Cap2) and multi-environment (MET) approaches.

SNP	Environment	p-value	$r_{\%}^2$ <sup>1</sup>	$a^2$	Allele <sup>3</sup>	Source <sup>4</sup>
ss715630650	Vic1	0.012*	2.47	0.76	A	PMQS12
	Vic2	1.000	0.39	0.50		
	Cap1	0.043*	1.65	0.74		
	Cap2	1.000	0.11	0.13		
	MET	0.100	1.08	0.33		
ss715636852	Vic1	0.001**	4.11	1.25	A	PMQS80
	Vic2	0.185	0.73	0.77		
	Cap1	0.001**	4.15	1.30		
	Cap2	0.001**	6.11	1.19		
	MET	0.004**	4.98	0.86		

<sup>1</sup>Coefficient of determination: proportion of RILs population means due to the SNP contribution.

<sup>2</sup>Additive effect of SNP marker over SPC%.

<sup>3</sup>Favorable allele that leads to the increase of SPC% [A: adenine; G: guanine; T: thymine; C: cytosine].

<sup>4</sup>Parent line source of favorable allele.

\*\*\* Significant at  $p \leq 0.05$  and  $p \leq 0.01$ .

Nonetheless, the stability of  $r_{\%}^2$  and  $a$  estimates throughout the environments and MET data suggests that a special attention should be given to ss715636852 for its consistent results. In environments where the two SNPs had significant association, the increment in SPC was greater than in the presence of either one of them alone. From that, we can infer that ss715630650 and ss715636852 may have distinct functions in the SPC metabolic pathway. Therefore, the marker association study to SPC should be performed previously in order to discard markers that show redundant contributions to the trait.

The lack of association of any marker to Vic2 environment can be assigned to the sowing date in late December 2018, in contrast to the sowing in early October for the other environments. Then, the genotype by environment interaction can lead to differential response of genes and its respective markers to sowing date.

The minor effect of ss715630650 can be verified in the Table 4, where its favorable allele at Vic1 and Cap1 was present at the inferior genotypes but absent in the superior ones. Zhang et al. (2015) also obtained high SPC progenies with minor effect favorable alleles when compared to other progenies in the population. This fact is justified by the interaction of the associated markers with the environment, as well as the performance of other markers associated to SPC, including those with irrelevant effect. That is to say, the additive effect of each associated marker can change according to different environments. Then, the changes in ranking genotypes according to SPC means and the lack of association of specific markers for some of the environments confirms an interaction between RILs with environments (significant  $\hat{\sigma}_{ge}^2$  estimates, Table 2), and suggests the interaction of the SPC associated markers with environments. Patil et al. (2018) justified a similar fact in their study due to the complex nature of a quantitative trait, in which its interaction with the environment implies instability of QTLs controlling soybean seed composition.

The top ranked RILs for SPC that carried favorable alleles had about 4% less SOC than the checks. In addition, the ones that had no favorable allele for SPC had about 2% less SOC. It can be considered that ss715630650 and mainly ss715636852 are not to recommended simultaneous selection as suggested by Singh (2017) and Li et al. (2017). Hence, for breeding programs seeking to increase soybean SPC and SOC simultaneously, ss715630650 and ss715636852 are not recommended.

The co-segregation and complexity in separating high SPC from low SOC evidence the effects of major pleiotropic genes (Hwang et al., 2014; Wang et al., 2014; Patil et al., 2018). However, it is possible to achieve a range of 41-43% and 20-22% for SPC and SOC respectively, especially in cases where the genes coding for each trait are tightly linked (Zhang et al., 2018). In such case, the recombination cycles for obtaining RILs, as well as further marker association studies are essential in identifying favorable alleles that can ease the breeding process. Beyond those values, pleiotropic genes play a large role over these traits

hampering the simultaneous selection. It reinforces the fact that simultaneous selection for both traits is complex and limited to specific situations (Zhang et al., 2018).

**Table 4.** Seed protein content (SPC) means of transgressive genotypes under 2% of selection intensity for single-environment (Vic1, Cap1, and Cap2) and multi-environment (MET) data. The environment V2 is absent due to lack of significant marker association. The colored cells represent the favorable allele for increasing SPC (A: adenine, G: guanine). The blank cells represent the lack of association between marker and SPC at the environment. The superior genotypes are presented in the first five rows, and the inferior ones in the last five rows.

Vic1				Cap1			
RIL #	SPC%	ss715630650	ss715636852	RIL #	SPC%	ss715630650	ss715636852
64-25	51.46	A	A	63-36	54.78	A	A
81-20	51.21	G	A	61-20	54.14	A	A
78-38	51.10	A	G	61-18	54.12	G	A
78-43	50.63	A	A	84-13	53.79	A	A
64-02	50.24	G	A	81-20	53.78	G	A
84-19	39.83	G	G	63-07	41.53	A	A
78-30	39.63	A	G	78-24	41.72	G	A
63-08	39.52	G	A	78-48	42.09	A	A
83-21	39.37	G	A	64-08	42.36	A	A
83-20	38.94	A	A	78-36	42.55	G	G
Cap2				MET			
RIL #	SPC%	ss715630650	ss715636852	RIL #	SPC%	ss715630650	ss715636852
61-30	53.40		A	61-30	48.84		A
84-17	53.19		A	61-01	48.81		A
78-43	52.99		A	78-43	48.73		A
79-28	52.69		G	79-01	48.68		A
61-01	52.66		A	64-02	48.39		A
62-12	42.67		G	83-21	40.75		A
63-07	42.97		A	63-20	40.89		A
79-54	43.52		G	63-11	41.38		G
78-31	43.79		G	61-03	41.47		A
63-11	43.92		G	79-75	41.49		G

Among the 264 RILs, the RIL 84-13 was considered an elite genotype once it presented a good field performance for SPC (53.78, 49.01, 47.56, 46.61, and 46.33% for Cap1, Cap2, Vic1, Vic2, and MET, respectively) and can serve as parent material to be destined to crosses for obtaining transgressive lines. In contrary, the remaining SPC lines containing both favorable alleles must undergo to a pre-breeding process to become more adapted to possess the essential agronomic traits. Similarly, ss715636852 must have its immediate use recommended in soybean breeding programs for increasing SPC. This marker can contribute to higher genetic gains.

## CONCLUSIONS

This work was useful for the association of markers that have impact on SPC and for the identification of superior RILs. The significant association of ss715636852 and its additive effect contribute to incrementing SPC in the process of marker-assisted breeding. On the other hand, the use of ss715630650 did not result in statistically detectable increment in SPC. The results showed the differential performance of genotypes in



respect to the different environments, which hinder selection in breeding programs for SPC. This research do not intend to be dogmatic and recommends a previous association of ss715630650, ss715636852 or any other marker prior the marker assisted breeding process. Therefore, the use of ss715636852 was beneficial for increasing SPC.

## ACKNOWLEDGEMENTS

Many thanks to colleagues from Plant Genetic Biochemistry Lab at Universidade Federal de Viçosa for the support. We also appreciate the assistance from CEPET-UFV and Horta Nova-UFV experimental breeding stations, as well as the efforts of this Journal's Editorial Committee and reviewers. The authors declare there is no conflict of interest in publishing this manuscript.

## FUNDING

CAPES, CNPq, FAPEMIG (grants APQ-01416-16 and PPM-640-18) and Caramuru Alimentos supported this work.

## REFERENCES

- Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., & Lorenz, A. (2015). A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. *The Plant Genome*, 8(3), plantgenome2015.04.0024. <https://doi.org/10.3835/plantgenome2015.04.0024>
- Bolon, Y., Joseph, B., Cannon, S. B., Graham, M. A., Diers, B. W., Farmer, A. D., May, G. D., Muehlbauer, G. J., Specht, J. E., Tu, Z. J., Weeks, N., Xu, W. W., Shoemaker, R. C., & Vance, C. P. (2010). Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean.
- Cruz, C. D. (2013). Genes: a software package for analysis in experimental statistics and quantitative genetics. *Acta Scientiarum. Agronomy*, 35(3), 271–276.
- Dellaporta, S. L., Wood, J., & Hicks, J. B. (1983). A plant DNA miniprep: Version II. *Plant Molecular Biology Reporter*, 1(4), 19–21. <https://doi.org/10.1007/BF02712670>
- Graef, G. L., Orf, J., Wilcox, J. R., & Shoemaker, R. C. (1997). Mapping QTL for Seed Protein and Oil Content in Eight Soybean Populations. 370–378.
- Hwang, E., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., & Cregan, P. B. (2014). A genome-wide association study of seed protein and oil content in soybean. 1–12.
- Jun, T. H., Van, K., Kim, M. Y., Lee, S. H., & Walker, D. R. (2008). Association analysis using SSR markers to find QTL for seed protein content in soybean. *Euphytica*, 162(2), 179–191. <https://doi.org/10.1007/s10681-007-9491-6>
- Kang, M. S. (1997). Using Genotype-by-Environment Interaction for Crop Cultivar Development. *Advances in Agronomy*, 62(C), 199–252. [https://doi.org/10.1016/S0065-2113\(08\)60569-6](https://doi.org/10.1016/S0065-2113(08)60569-6)
- Kwon, S. H., & Torrie, J. H. (1964). Heritability and interrelationship among traits of two soybean populations. *Crop Sci*, 4(2), 196–198.
- Li, Z., Stewart-Brown, B., Steketee, C., & Vaughn, J. (2017). Impact of Genomic Research on Soybean Breeding. In *The Soybean Genome* (pp. 111–129). Springer.

- Mahmoud, A. A., Natarajan, S. S., Bennett, J. O., Mawhinney, T. P., Wiebold, W. J., & Krishnan, H. B. (2006). Effect of six decades of selective breeding on soybean protein composition and quality: A biochemical and molecular analysis. *Journal of Agricultural and Food Chemistry*, 54(11), 3916–3922. <https://doi.org/10.1021/jf060391m>
- Patil, G., Mian, R., Vuong, T., Pantalone, V., Song, Q., Chen, P., Shannon, G. J., Carter, T. C., & Nguyen, H. T. (2017). Molecular mapping and genomics of soybean seed protein : a review and perspective for the future. *Theoretical and Applied Genetics*, 130(10), 1975–1991. <https://doi.org/10.1007/s00122-017-2955-8>
- Patil, G., Vuong, T. D., Kale, S., Valliyodan, B., Deshmukh, R., Zhu, C., Wu, X., Bai, Y., Yungbluth, D., Lu, F., Kumpatla, S., Shannon, J. G., Varshney, R. K., & Nguyen, H. T. (2018). Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an interspecific mapping population of soybean using high-density linkage mapping. *Plant Biotechnology Journal*, 16(11), 1939–1953. <https://doi.org/10.1111/pbi.12929>
- Piper, E. L., & Boote, K. J. (1999). Temperature and Cultivar Effects on Soybean Seed Oil and Protein Concentrations. 76(10). <https://doi.org/10.1007/s11746-999-0099-y>
- Rao, C. R. (1973). Linear statistical inference and its applications. In *Zeitschrift Angewandte Mathematik und Mechanik* (XX, Vol. 57). John Wiley & Sons.
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21–33.
- Reinprecht, Y., Poysa, V. W., Yu, K., Rajcan, I., Ablett, G. R., & Pauls, K. P. (2006). Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (*Glycine max* (L.) Merrill) germplasm. *Genome*, 49(12), 1510–1527. <https://doi.org/10.1139/G06-112>
- Rodrigues, J. I. da S., Arruda, K. M. A., Cruz, C. D., Piovesan, N. D., de Barros, E. G., & Moreira, M. A. (2014). Biometric analysis of protein and oil contents of soybean genotypes in different environments. *Pesquisa Agropecuaria Brasileira*, 49(6), 475–482. <https://doi.org/10.1590/S0100-204X2014000600009>
- Rodrigues, J. I. S., Miranda, F. D., Ferreira, A., Borges, L. L., Ferreira, M. F. da S., Good-God, P. I. V., Piovesan, N. D., de Barros, E. G., Cruz, C. D., & Moreira, M. A. (2010). Mapeamento de QTL para conteúdos de proteína e óleo em soja. *Pesquisa Agropecuaria Brasileira*, 45(5), 472–480. <https://doi.org/10.1590/S0100-204X2010000500006>
- Santana, D. P., & Moura Filho, W. (1978). Estudos de solos do Triângulo Mineiro e de Viçosa. I. Mineralogia. *Embrapa Milho e Sorgo-Artigo Em Periódico Indexado (ALICE)*.
- Schuster, I., & Cruz, C. D. (2008). *Estatística Genômica* (2nd ed.). Editora UFV.
- Sebolt, A. M., Shoemaker, R. C., & Diers, B. W. (2000). Analysis of a Quantitative Trait Locus Allele from Wild Soybean That Increases Seed Protein Concentration in Soybean. 468916, 1438–1444.
- Sediyama, T., Silva, F., & Borém, A. (2015). *Soja: do plantio à colheita*. Editora UFV.
- Singh, R. J. (2017). The Soybean Genome. *The Soybean Genome, Compendium of Plant Genomes*, 11–38. <https://doi.org/10.1007/978-3-319-64198-0>
- Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B. (2013). Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. *PLoS ONE*, 8(1), 1–12. <https://doi.org/10.1371/journal.pone.0054985>

- R Core Team. (2019). R: A Language and Environment for Statistical Computing Version 3.5.2, R Foundation for Statistical Computing, Vienna, Austria.
- Vaughn, J. N., Nelson, R. L., Song, Q., Cregan, P. B., & Li, Z. (2014). The Genetic Architecture of Seed Composition in Soybean Is Refined by Genome-Wide Association Scans Across Multiple Populations. 4(November), 2283–2294. <https://doi.org/10.1534/g3.114.013433>
- Wang, X., Liang, G., Marci, J., Scott, R. A., Song, Q., Hyten, D. L., & Cregan, P. B. (2014). Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. <https://doi.org/10.1007/s00438-014-0865-x>
- Warrington, C. V., Abdel-Haleem, H., Hyten, D. L., Cregan, P. B., Orf, J. H., Killam, A. S., Bajjalieh, N., Li, Z., & Boerma, H. R. (2015). QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. *Theoretical and Applied Genetics*, 128(5), 839–850. <https://doi.org/10.1007/s00122-015-2474-4>
- Yesudas, C. R., Bashir, R., Geisler, M. B., & Lightfoot, D. A. (2013). Identification of germplasm with stacked QTL underlying seed traits in an inbred soybean population from cultivars Essex and Forrest. 693–703. <https://doi.org/10.1007/s11032-012-9827-3>
- Zhang, J., Wang, X., Lu, Y., Bhusal, S. J., Song, Q., Cregan, P. B., Yen, Y., Brown, M., & Jiang, G. (2018). Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. *Molecular Plant*, 11(3), 460–472. <https://doi.org/10.1016/j.molp.2017.12.016>
- Zhang, Y. H., Liu, M. F., He, J. B., Wang, Y. F., Xing, G. N., Li, Y., Yang, S. P., Zhao, T. J., & Gai, J. Y. (2015). Marker-assisted breeding for transgressive seed protein content in soybean [*Glycine max* (L.) Merr.]. *Theoretical and Applied Genetics*, 128(6), 1061–1072. <https://doi.org/10.1007/s00122-015-2490>

**Received: May 31, 2020.**

**Accepted: June 29, 2020.**

**Published: July 17, 2020.**

**English by: Arthur Bernadeli.**