Genome-wide association for mapping QTLs linked to protein and oil contents in soybean

Douglas Antônio Dias[1], Leandra Regina Texeira Polo[2], Fabiane Lazzari[2], Glacy Jaqueline da Silva[3] and Ivan Schuster[3]

[1]Universidade Federal da Fronteira Sul, Departamento de Agronomia, Campus Erechim, ERS 135, Km 72, Caixa Postal 764, CEP 99700-970 Erechim, RS, Brazil. E-mail: douglasadias66@hotmail.com [2]Cooperativa Central de Pesquisa Agrícola, BR 467, Km 98, Caixa Postal 89, CEP 85813-450 Cascavel, PR, Brazil. E-mail: leandra.texeira@coodetec.com.br, flazzari1@dow.com [3]Universidade Paranaense, Campus I, Praça Mascarenhas de Moraes, nº 4.282, Zona III, CEP 87502-210 Umuarama, PR, Brazil. E-mail: glacy.jaqueline@gmail.com, ivanschuster.ivan@gmail.com

Abstract – The objective of this work was to identify single-nucleotide polymorphism (SNP) markers linked with quantitative trait loci (QTLs) associated with increased contents of protein and oil in soybean. A total of 169 Brazilian soybean varieties, genotyped with 6,000 SNP markers, were evaluated. Protein and oil contents were obtained with the near-infrared reflectance method. Correlation and multiple linear regression analyses were used to identify linkage disequilibrium between SNP markers and the QTLs associated with the two characteristics. Seven QTLs were found to be associated with protein content, on six chromosomes (2, 6, 11, 12, 13, and 16), explaining 60.9% of the variation in this trait. For oil content, eight QTLs were identified on six chromosomes (1, 4, 5, 6, 17, and 19), explaining 78.3% of the variation in the trait. The correlation between the number of loci containing favorable alleles and the evaluated characteristics was 0.49 for protein content and 0.60 for oil content. The molecular markers identified are mapped in genomic regions containing QTLs previously mapped for both characteristics, which reinforces the association between these regions and the genetic control of oil and protein contents in soybean.

Index terms: *Glycine max*, association genetics, genomic selection, grain quality, linkage disequilibrium, marker-assisted selection.

Associação genômica ampla para mapeamento de QTLs ligados a conteúdos de proteína e óleo em soja

Resumo – O objetivo deste trabalho foi identificar marcadores de polimorfismos de nucleotídeo único (SNPs) ligados a locos controladores de características quantitativas (QTLs) associados a maiores conteúdos de proteína e óleo em soja. Foram avaliadas 169 cultivares brasileiras de soja, genotipadas com 6 mil marcadores SNPs. Os conteúdos de proteína e óleo foram obtidos pelo método de refletância no infravermelho próximo. As análises de correlação e regressão linear múltipla foram utilizadas para a identificação de desequilíbrio de ligação entre os marcadores SNPs e os QTLs associados às duas características. Foram identificados sete QTLs associados ao conteúdo de proteína, em seis cromossomos (2, 6, 11, 12, 13 e 16), o que explicou 60,9% da variação nesta característica. Para o conteúdo de óleo, foram identificados oito QTLs em seis cromossomos (1, 4, 5, 6, 17 e 19), o que explicou 78,3% da variação na característica. A correlação entre o número de locos com alelos favoráveis e as características avaliadas foi de 0,49 para o conteúdo de proteína e de 0,60 para o conteúdo de óleo. Os marcadores moleculares identificados estão mapeados em regiões genômicas que contêm QTLs previamente mapeados para as duas características, o que reforça a associação dessas regiões com o controle genético dos conteúdos de óleo e proteína em soja.

Termos para indexação: *Glycine max*, genética de associação, seleção genômica, qualidade do grão, desequilíbrio de ligação, seleção assistida por marcadores.

Introduction

Soybean (*Glycine max* L.) is the world's main source of plant protein and oil, providing approximately 20 to 24% of the oil and fat consumed worldwide (Cavalcante et al., 2009). This legume also has the greatest concentration of protein of all food crops. The soybean production technology used in Brazil is among the best in the world, as are Brazilian soybean yields (Masuda & Goldsmith, 2009). Commercial soybean varieties in the country contain about 40% protein and 20% oil,

which vary according to genetic and environmental factors (Soares et al., 2004; Moraes et al., 2006).

Quantitative trait locus (QTL) mapping strategies that use populations derived from two-parent crossing reveal only a small fraction of all possible alleles in a target species. Therefore, molecular markers associated to QTLs can generally be used only in populations for which the markers were specifically developed (Cahill & Shmidt, 2004; Holland, 2004; Schuster, 2011). QTL mapping using linkage disequilibrium infers the association between genotypes (or haplotypes) and phenotypes by evaluating the genetic polymorphism generated in different genetic backgrounds, throughout many recombination generations (Dekkers & Hospital, 2002; Nordborg & Tavaré, 2002).

Association mapping detects and locates QTLs based on the correlation intensity between molecular markers and phenotypic characteristics. Linkage disequilibrium between two loci is a function both of the time (number of generations) passed since the recombination generations began and of the recombination frequency between the loci. After many recombination generations in an unstructured population, only the correlations between QTLs and markers closely linked should remain, facilitating a more precise mapping (Mackay & Powell, 2007). For linkage disequilibrium mapping, however, there is no need to prepare a mapping population, and the entire genome is evaluated to identify regions associated with a particular phenotype. The greater the association between marker alleles and phenotype variants, the greater the probability that the phenotype is physically linked to the marker (Hwang et al., 2014).

The objective of this work was to identify SNP markers linked with QTLs associated with increased contents of protein and oil in soybean.

Materials and Methods

The experiment was carried out at the facilities of Cooperativa Central de Pesquisa Agrícola, located in the municipality of Cascavel, in the state of Paraná, Brazil. A total of 169 Brazilian soybean varieties were evaluated in the 2011/2012 crop year, and the field trial was performed in a 13x13 lattice design. The plots contained four 5.0-m lines, where the agronomic characteristics of the varieties were assessed. At harvest, a sample of grains was taken from each variety for the evaluation of protein and oil contents.

Approximately 20 g of grains from each sample were ground in a cyclone-type grinder to obtain uniform, thin powder. Protein, oil, and moisture contents were determined using the Instalab 600 near-infrared reflectance (NIR) device (Dickey-John, Auburn, IL, USA), which was previously calibrated. The data were converted to express the contents of oil and protein on a dry matter basis.

A sample of 100 seeds from each variety was ground, and 50 mg of homogenized powder were used for DNA extraction, as described by Schuster et al. (2004).

The genotyping with SNP markers was carried out at Deoxi Biotecnologia Ltda, located in the municipality of Araçatuba, in the state of São Paulo, Brazil, using the iScan platform and the 6k Infinium iSelect HD Custom Genotyping BeadChip panel (Illumina, Inc., San Diego, CA, USA), customized for soybean. The process followed the instructions of the manufacturer (Illumina, 2014). Markers with more than 10% lost data or minor allele frequency (MAF) less than 5% were removed from the analysis.

The association between markers and phenotypes was assessed by correlation and multiple regression analyses. The correlation analysis was carried out using the following expression:

$$r = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left(\sum x_i^2 - \frac{1}{n}(x_i)^2\right)\left(\sum y_i^2 - \frac{1}{n}(y_i)^2\right)}};$$

where x_i is the score of the genotype with marker x for individual i; y_i is the phenotype of characteristic y in individual i; and n is the number of samples. The markers received score 0, for one homozygote genotype; and 2, for the alternate homozygote genotype. The few heterozygote genotypes were considered lost data.

The square of the correlation value (R^2), weighted by n - 1, has a chi-squared distribution with one degree of freedom: $\chi^2 = (n - 1)r^2$; where n is the number of individuals and r is the correlation between the marker and the phenotype.

Correlation significance was determined using a chi-squared distribution, expressed as -log10(p), where p is the probability value. The significance levels were corrected using the false discovery rate (FDR) method (Benjamini & Hochberg, 1995). The correlation

analyses and the FDR correction were carried out in an Excel sheet. Markers with p-values lower than 0.0001 after the correction with the FDR method were considered associated with the studied phenotype.

The JMP software (SAS Institute Inc., Cary, NC, USA) was used for the multiple regression analysis. Input and output probability was 5%. The Stepwise procedure for variable selection was adopted. For the multiple regression analysis, only markers that were significant in the correlation analysis ($p<0.0001$) were used.

In addition, a correlation analysis was performed between the number of loci containing favorable alleles with the selected marker and the protein and oil contents of the soybean variety.

Results and Discussion

The contents of oil and protein of the 169 varieties varied from 37.2 to 48.3% for protein and from 18.2 to 27.5% for oil, based on the dry matter of the grains (Figure 1). However, protein contents between 42 and 43%, and oil contents between 23 and 24% were more frequent.

After filtering the markers for MAF higher than 5% and lost data lower than 10%, 4,962 SNPs were used to analyze genome association. In the correlation analysis, a significant association was observed for markers in all of the 20 chromosomes, for protein content, and in all chromosomes, except 7 and 16, for oil content (Figure 2). In the soybean consensus map (USDA, 2016), 125 QTLs have been mapped for protein and 148 QTLs for oil, and all of the 20 chromosomes contain QTLs for oil and protein contents.

In the multiple regression analysis, seven significant markers were associated with protein contents, in the six following chromosomes: 2, 6, 11, 12, 13, and 16 (Table 1). These markers explained 60.93% of the variation in the protein content in the evaluated set of varieties. Redundant markers, associated with the same QTL, were eliminated from the model, leaving just one marker associated with each QTL. Markers linked to QTLs with minor effects were also eliminated. Therefore, although 95 markers were significant in the correlation analysis, on all 20 chromosomes, only 7 markers, on 6 chromosomes, explained more than 60% of the variation in protein content. The regression coefficient values of these seven markers varied from 0.41 to 1.45, meaning that the substitution of one allele in the SNP markers associated with protein resulted in an increase in protein content between 0.41 and 1.45%.

On chromosome 6, there are two significant SNPs according to the multiple regression model. These SNPs are separated by more than 7.8 Mb on the soybean genome and are in linkage equilibrium. This means that each marker was associated with a different QTL.

Jun et al. (2008) identified 11 QTLs for protein content in soybean using the association analysis. Csanádi et al. (2001) reported QTLs for protein content on chromosomes 1, 6, 7, and 9. In Brazil, Soares et al. (2008) detected QTLs on chromosomes 3, 6, 15, 18, and 19, explaining from 38.84 to 55.53% of the variation in protein content in soybean, depending on the cultivation site. QTLs for protein content in soybean have also been found on chromosome 20 (Chung et al., 2003; Nichols et al., 2006), chromosome 18 (Panthee

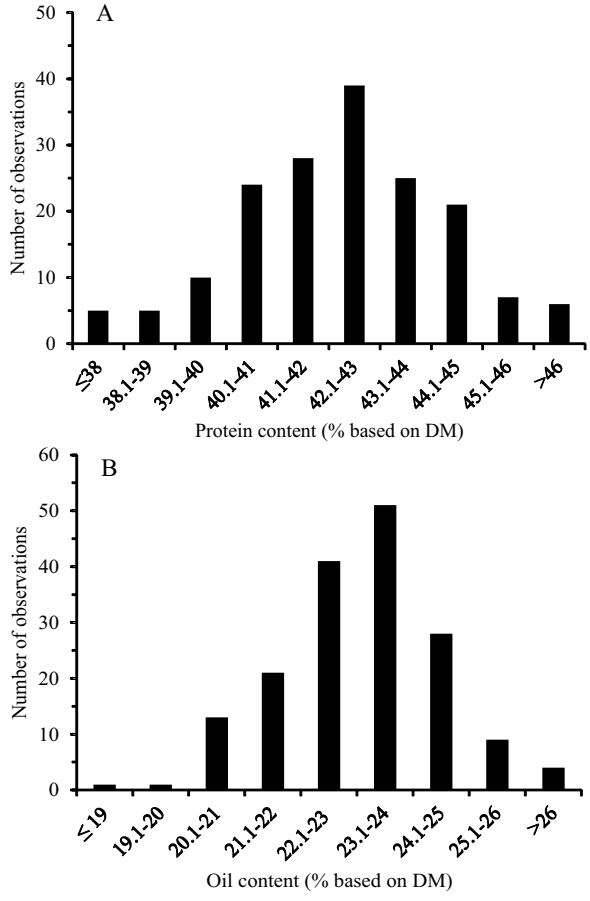

Figure 1. Distribution frequency of protein (A) and oil (B) contents in the 169 Brazilian soybean (*Glycine max*) varieties evaluated. DM, dry matter.

et al., 2005; Leamy et al., 2017), and chromosome 14 (Zhang et al., 2004; Leamy et al., 2017).

In the region of chromosome 13 where marker Gm13_33637077_T_C was mapped, three QTLs were associated with protein in the soybean consensus map: *Seed Protein 6-1*, *Seed Protein 24-2*, and *Seed Protein 26-11*. The *Seed Protein 5-2* QTL is mapped on the same region of chromosome 12 where

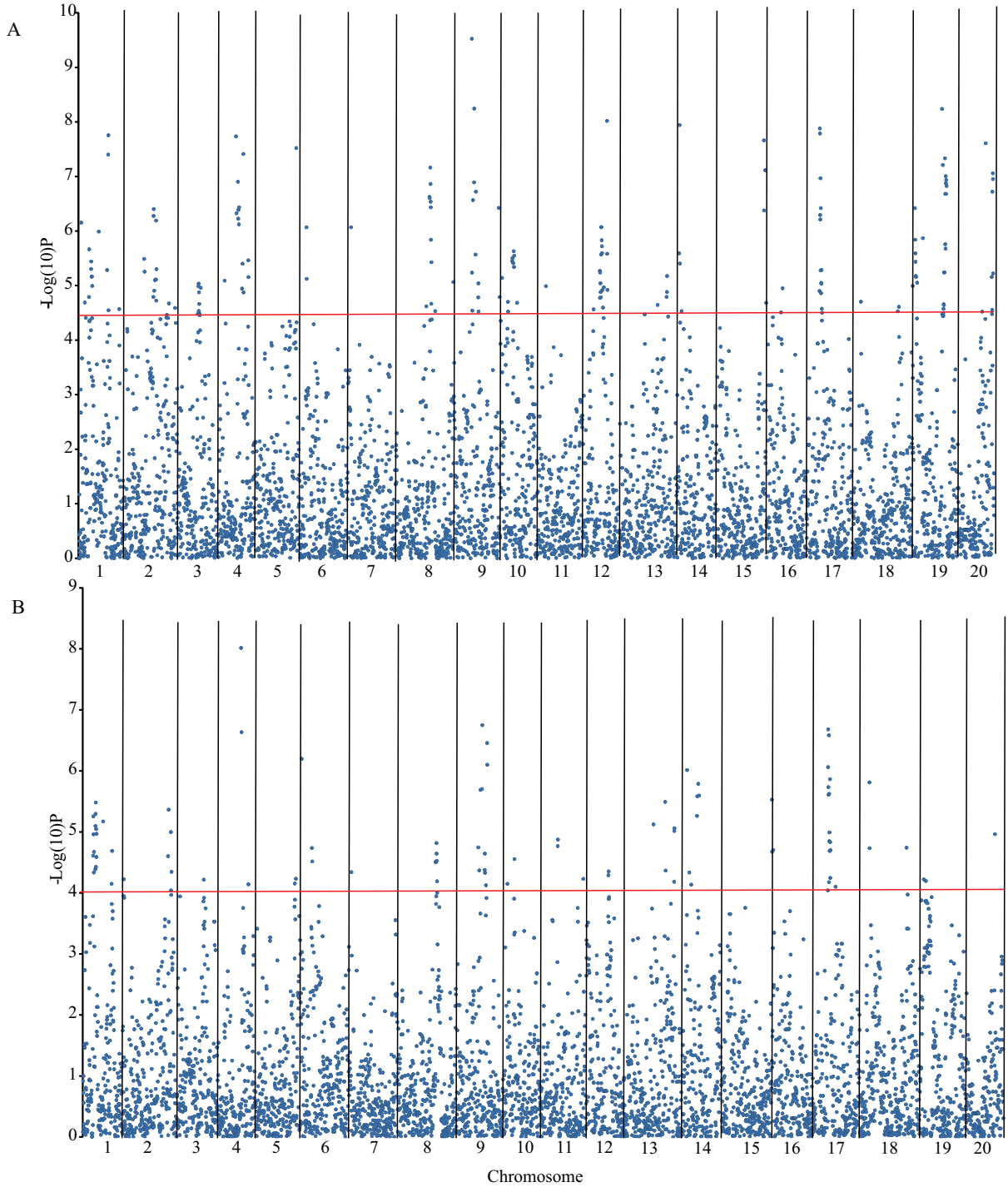

Figure 2. Manhattan plot of the association probability between single-nucleotide polymorphism markers and the protein (A) and oil (B) contents of the 169 Brazilian soybean (*Glycine max*) varieties evaluated. The horizontal line indicates the cut-off point for significance, at 0.01% probability.

marker Gm12_35525603_A_G is mapped. Marker Gm11_5536036_C_A is mapped on chromosome 11, in the same region as the *Seed Protein 3-2* and *Seed Protein 34-7* QTLs. On chromosome 6, marker Gm06_12914255_A_G is mapped in the same region as the *Seed Protein 34-2* QTL, while marker Gm06_49106015_C_T is located on the edge of chromosome 6. In this region, there are no protein QTLs mapped (USDA, 2016).

None of the assessed varieties had alleles associated with high or low protein contents on all of the seven markers. The ten varieties with the highest protein content had 4–6 loci with alleles associated with high protein content on the seven markers considered. Of these, the four varieties with the highest protein content had at least five markers with alleles associated with high protein content (Table 2). Among the ten varieties with the lowest protein content, the number of loci associated with high protein content varied from one to four.

Among the set of 20 varieties containing the ten higher and ten lower protein contents (Table 2), the correlation between protein content and the number of loci containing alleles associated with the trait was 0.84. Out of all 169 varieties, the correlation was 0.49, significant at 1% probability. Although there is a close association between protein content and the number of loci containing favorable alleles, this association is even more important in identifying extreme-high and extreme-low protein contents in soybean.

Of the seven markers, the best at discriminating genotypes into a group of high or low protein content were Gm11_5536036_C_A (chromosome 11) and Gm13_33637077_T_C (chromosome 13). In the group of varieties with the highest protein contents containing data for the marker, seven had alleles associated with high protein content on the SNP Gm11_5536036_C_A. In the group of nine varieties with the lowest protein content containing data for the marker, none had any allele associated with high protein content on this locus.

On the locus SNP Gm13_33637077_T_C, all ten of the varieties with the highest protein content had an allele associated with high protein content, whereas, in the group of eight varieties with the lowest protein content, only two had this allele.

Oil content had eight significant SNP markers in the stepwise multiple regression analysis, on the following six chromosomes: 1, 4, 5, 6, 17, and 19 (Table 1).

Table 1. Stepwise multiple regression analysis of the protein and oil contents of the 169 Brazilian soybean (*Glycine max*) varieties evaluated, averages of the genotypes containing alleles associated with high and low protein contents, and number of individuals found for each genotype.

Marker	SNP	Chromosome	b[1]	High protein (allele A)[2]	Average±SD for allele A	Homozygous allele A (No.)	Average for allele B[2]	Allele B (No.)
					Protein content			
Gm13_33637077_T_C	T/C	13	0.77	C	42.83 (1.87)	123	40.94 (1.82)	26
Gm12_35525603_A_G	A/G	12	1.45	G	43.41 (2.67)	11	42.32 (1.87)	151
Gm11_5536036_C_A	C/A	11	0.65	C	43.12 (1.81)	61	41.93 (1.92)	87
Gm06_12914255_A_G	A/G	6	1.35	A	42.47 (1.93)	160	39.34 (1.80)	4
Gm02_7987834_A_G	A/G	2	0.80	G	43.27 (1.84)	13	42.34 (1.95)	144
Gm06_49106015_C_T	C/T	6	0.67	C	42.67 (1.86)	124	41.57 (2.22)	26
Gm16_37209075_T_G	T/G	16	0.41	G	42.56 (2.16)	70	42.26 (1.85)	74
R^2					60.93%			
					Oil content			
Gm17_8270421_A_G	A/G	17	0.63	A	23.66 (1.32)	90	22.48 (1.16)	49
Gm04_40811025_C_A	C/A	4	0.45	C	23.81 (1.21)	73	22.46 (1.44)	74
Gm05_32361439_C_A	C/A	5	0.60	C	23.25 (1.47)	137	22.40 (1.17)	22
Gm19_1420943_T_C	T/C	19	0.31	C	23.49 (1.26)	62	22.92 (1.64)	82
Gm05_8656389_T_C	T/C	5	0.40	T	23.37 (1.59)	75	22.89 (1.18)	68
Gm01_51330200_C_A	C/A	1	0.43	C	24.24 (1.30)	37	22.74 (1.37)	119
Gm17_18586619_C_T	C/T	17	0.69	C	23.23 (1.40)	157	22.04 (1.83)	9
Gm06_13237698_A_G	A/G	6	0.33	A	23.47 (1.54)	75	22.80 (1.40)	73
R^2					78.30%			

[1]Angular coefficient of the regression equation. [2]The allele associated with increased protein content was codified as A, and the allele associated with reduced protein content was codified as B. R^2, correlation value; SNP, single-nucleotide polymorphism; and SD, standard deviation.

These markers explained 78.30% of the variation in oil content, in the set of varieties studied. Two SNPs were identified on chromosome 5 and two on chromosome 17. Because the markers were in linkage equilibrium on the same chromosome, they were probably linked to different QTLs. The regression coefficients of the eight markers varied from 0.31 to 0.69.

Csanádi et al. (2001) identified QTLs for oil content on chromosomes 14, 9, and 20. Panthee et al. (2005) reported QTLs on chromosomes 2, 10, and 18. The QTL of chromosome 2 was also detected by Zhang et al. (2004). Leamy et al. (2017), in turn, found QTLs on chromosome 3 and 20 in wild soybean. Although none of these QTLs were observed in the present study, the eight QTLs identified were located in regions that have QTLs mapped for oil and fatty acid contents in the consensus map (USDA, 2016).

The *Seed Oil 23-3* QTL is mapped in the region of chromosome 17, near the marker Gm17_8270421_A_G. The *Seed Oil 6-1*, *Seed Protein 3-3*, *cqSeed Oil-001*, and *Seed oleic 6-5* QTLs are mapped near marker Gm04_40811025_C_A on chromosome 4. Marker Gm05_32361439_C_A is located in a region that contains both the *Seed Oil 4-1* and *Seed oil to protein ratio 1-1* QTLs, on chromosome 5. Marker Gm05_8656389_T_C is also on chromosome 5, near the *Seed palmitic 2-1*, *Seed linolenic 7-5*, *Seed Oil 24-16*, *Seed Oil 36-1*, *Seed palmitic 7-1*, *Seed stearic 6-1*, and *Seed Oil 38-1* QTLs. Marker Gm19_1420943_T_C is located on chromosome 19 near the *Seed Oil 27-5*, *Seed linolenic 7-1*, *Seed linolenic 8-3*, and *Seed oil 37-6* QTLs. Marker Gm01_51330200_C_A is on chromosome 1, near the *Seed Oil* 24-21 QTL. Marker Gm17_18586619_C_T is located on chromosome 17 in the region that contains the *Seed Oil 5-6*, *Seed Oil 5-4*, *Seed Oil 5-5*, *Seed Oil 24-22*, *Seed linoleic 6-6*,

Table 2. Single-nucleotide polymorphism (SNP) genotypes in the groups of ten soybean (*Glycine max*) varieties with the highest (allele A associated with increased protein contents) and the lowest protein contents (allele B associated with decreased protein contents).

Variety	Gm02_7987834_A_G	Gm06_12914255_A_G	Gm06_49106015_C_T	Gm11_5536036_C_A	Gm12_35525603_A_G	Gm13_33637077_T_C	Gm16_37209075_T_G	Protein	Favorable loci[1]
CAC 1	BB	AA	AA	AA	AA	AA	AA	48.32	6
BRSMT Pintado	BB	AA	AA	AA	BB	AA	AA	46.61	5
CD 206	AA	AA	AA	AA	BB	AA	BB	46.28	5
Capinópolis	BB	AA	AA	BB	AA	AA	AA	46.18	5
FT Abyara	-[2]	AA	AA	AB	BB	AA	BB	46.16	5
CD 251RR	BB	AA	AA	AA	BB	AA	BB	46.08	4
CD 2792RR	BB	AA	AA	AA	AA	AA	AA	45.79	6
CD 2800	BB	AA	AA	AA	BB	AA	AB	45.64	5
CD 246	BB	AA	AA	-	AB	AA	AA	45.44	5
IAS 5	BB	AA	AA	AA	BB	AA	BB	45.26	4
SPRING 5.3	BB	AA	BB	BB	BB	BB	BB	39.15	1
ANTA	BB	AA	AB	BB	BB	BB	BB	39.01	2
A 6001RR	BB	AA	AB	BB	BB	BB	BB	38.65	2
FUNDACEP 55RR	BB	AA	AA	BB	AA	-	AA	38.64	4
CD 202	AB	AA	BB	-	AB	-	AA	38.11	4
BRS 283	BB	AA	AA	BB	BB	BB	BB	37.99	2
TMG 7161RR	BB	BB	AA	BB	BB	BB	AB	37.88	2
BMX ATIVA RR	BB	BB	AA	BB	BB	BB	AA	37.69	2
CD 233RR	BB	AA	BB	BB	BB	AA	AA	37.54	3
CD 235RR	BB	AA	BB	BB	BB	AA	AA	37.18	3

[1]Number of loci containing alleles associated with high protein content, in the group of seven SNP markers being considered. [2]Data lost in genotyping.

Seed palmitic 8-1, *Seed linoleic 8-2*, and *Seed oleic 8-4* QTLs. Finally, marker Gm06_13237698_A_G is near the *Seed Oil 9-2* QTL, on chromosome 6 (USDA, 2016).

In the group of ten varieties with the highest oil contents, the number of loci with alleles associated with the trait varied from 5 to 8, considering the eight significant SNP markers; however, in the group of ten varieties with the lowest oil content, the number varied from 2 to 4 (Table 3). In the group with these 20 varieties, the correlation between oil content and the number of loci was 0.88, and, when all 169 varieties were considered, the correlation was 0.60, significant at 1% probability.

Markers Gm01_51330200_C_A (chromosome 1) and Gm04_40811025_C_A (chromosome 4) were the most common in the group with high and low oil contents, respectively. The SNP Gm01_51330200_C_A had alleles associated with high oil content in nine of the ten varieties with the highest oil contents, but only in one of the ten varieties with the lowest (Table 3). The SNP Gm01_51330200_C_A had alleles associated with high oil content in eight of the ten varieties with the highest oil contents, but in none of the ten varieties of the group with the lowest.

The SNPs Gm05_32361439_C_A and Gm05_8656389_T_C (chromosome 5) and Gm17_8270421_A_G (chromosome 17) were also well defined, with alleles associated with high oil content in nine of the ten varieties with the highest contents and in five of the nine with the lowest (Table 3).

In the studied population, the correlation between oil and protein contents was -0.68. Note that the significant markers in the multiple regression models for both characteristics were found on different chromosomes, except for chromosome 6, which had QTLs for protein and oil. Chromosomes 2, 11, 12, 13, and 16 have QTLs for protein content, while chromosomes 1, 4, 5, 17, and 19 have QTLs for oil content.

Table 3. Single-nucleotide polymorphism (SNP) genotypes in the groups of ten soybean (*Glycine max*) varieties with the highest (allele A associated with increased oil contents) and the lowest oil contents (allele B associated with decreased oil contents).

Variety	Gm04_40811025_C_A	Gm05_32361439_C_A	Gm17_8270421_A_G	Gm19_1420943_T_C	Gm05_8656389_T_C	Gm01_51330200_C_A	Gm17_18586619_C_T	Gm06_13237698_A_G	Oil	Favorable loci[1]
TMG 7161RR	AA	AA	AA	BB	AA	BB	AA	BB	27.5	5
SPRING 5.3	AA	AA	AA	AA	AA	AA	AA	AA	26.9	8
BRS 283	BB	AA	AA	BB	AA	AA	AA	AA	26.7	6
CD 235RR	AA	AA	AA	BB	AB	AA	AA	AA	26.3	7
NIDERA A4725RG	AA	AA	AA	BB	AA	AA	AA	AA	26.0	7
FUNDACEP 63RR	AA	AA	AA	BB	AA	AA	AA	AA	26.0	7
TMG 1067RR	AA	AA	AA	AA	BB	AA	AA	BB	26.0	6
CD 2585RR	AA	AA	AA	BB	AB	AA	AA	-[2]	25.6	6
NA 5909RR	AA	AA	AA	AA	AA	AA	AA	AA	25.6	8
CD 252	AB	AB	AB	AA	AA	BB	AA	BB	25.2	6
BRS 256RR	BB	-	AB	AA	BB	BB	AA	AA	20.7	4
CD 251RR	AB	BB	BB	BB	AA	BB	AA	BB	20.7	3
CD 246	BB	BB	AB	AA	-	BB	AB	AA	20.6	4
R7	BB	BB	AA	AA	BB	BB	AA	BB	20.5	3
EMGOPA 304	BB	AA	BB	BB	BB	BB	AA	BB	20.4	2
BRS 245RR	BB	BB	AA	BB	BB	BB	AA	BB	20.4	2
CAC-1	BB	AA	BB	BB	AA	BB	AA	BB	20.3	3
CD 247RR	BB	AA	BB	AA	AA	BB	AA	BB	20.3	4
BRSMT Crixás	BB	AA	BB	BB	BB	BB	AA	AA	19.7	3
BRSMT Pintado	BB	AA	BB	BB	AA	BB	BB	AA	18.2	3

[1]Number of loci containing alleles associated with high oil content, in the group of eight SNP markers being considered. [2]Data lost in genotyping.

The correlation between the two characteristics might occur due to population structure, genetic linkage, or pleiotropy. Because the QTLs for protein and oil contents were found on different chromosomes, there were QTLs that were not linked. While there are genetic and physiological limits to the simultaneous increase in both oil and protein contents in the grains, the presence of independent QTLs for these two characteristics indicates that it should be possible to simultaneously select for QTLs associated with oil and protein contents.

All SNP markers significant in the multiple regression analysis were located in regions with QTLs for oil and protein contents previously identified, except Gm06_49106015_C_T. Since these traits are complex, the identification of the same QTLs in independent studies increases the consistency of their mapping.

Mapping QTLs associated with protein and oil contents in soybean has generally been carried out on structured populations, which is not the case of the present study. However, the genomic regions associated with QTLs for protein and oil contents found here are consistent with those of a structured population. This is not surprising because the genetic bases of cultivated soybean are relatively narrow.

In the marker-assisted selection for quantitative traits, plants with the greatest number of favorable alleles in QTL loci should be selected, since this greatly increases the chance of adding desirable characteristics to the plants. The high correlation between the number of loci containing favorable alleles associated with the studied phenotypes, observed in the present study, can certainly help Brazilian soybean genetic breeding programs to generate superior genotypes as to the contents of protein and oil in grains.

This is the first known study that identifies QTLs associated with protein and oil contents using Brazilian germplasm subjected to genome-wide association analysis, and it should be a starting point for understanding the population structure of this germplasm as to protein and oil contents.

Conclusions

1. The single-nucleotide polymorphism markers associated with oil and protein contents in soybean (*Glycine max*), reported here, can enhance the mapping consistency of the quantitative trait loci (QTLs) identified in other studies for these characteristics in the same genomic regions.

2. Selecting plants or lines containing the greatest number of favorable QTLs on the identified markers can increase their oil or protein content.

3. Most of the QTLs for protein content are found on different chromosomes than those for oil content, which allows marker-assisted selection to provide gains for both characteristics simultaneously.

Acknowledgments

To Conselho Nacional de Pesquisa e Desenvolvimento (CNPq), for financial support (process No. 482198/2010-9) and for the scientific productivity fellowships (process No. 3059000/2011-0) granted to the third author.

References

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society. Series B (Methodological)**, v.57, p.289-300, 1995.

CAHILL, D.J.; SCHMIDT, D.H. Use of marker assisted selection in a product development breeding program. In: INTERNATIONAL CROP SCIENCE CONGRESS, 4., 2004, Brisbane, Australia. **New Directions For a Diverse Planet**: proceedings. Queensland: The Regional Institute Ltd., 2004. Editors Tony Fisher, Neil Turner, John Angus, Lynne McIntyre, Michael Robertson, Andrew Borrell, and David Lloyd. Available at: <www.cropscience.org.au/icsc2004/>. Accessed on: Apr. 30 2008.

CAVALCANTE, A.K.; ESPINDOLA, S.M.C.G.; HAMAWAKI, T.O.; BISINOTTO, F.F.; COSTA, E.G.; GONÇALVES, F.A. Avaliação e seleção de linhagens de soja quanto ao teor de óleo para a produção de biodiesel. **FAZU em Revista**, n.6, p.11-52, 2009.

CHUNG, J.; BABKA, H.L.; GRAEF, G.L.; STASWICK, P.E.; LEE, D.J.; CREGAN, P.B.; SHOEMAKER, R.C.; SPECHT, J.E. The seed protein, oil, and yield QTL on soybean linkage group I. **Crop Sciences**, v.43, p.1053-1067, 2003. DOI: 10.2135/cropsci2003.1053.

CSANÁDI, G.; VOLLMANN, J.; STIFT, G.; LELLEY, T. Seed quality QTLs identified in a molecular map of early maturing soybean. **Theoretical and Applied Genetics**, v.103, p.912-919, 2001. DOI: 10.1007/s001220100621.

DEKKERS, J.C.M.; HOSPITAL, F. The use of molecular genetics in the improvement of agricultural populations. **Nature Reviews Genetics**, v.3, p.22-32, 2002. DOI: 10.1038/nrg701.

HOLLAND, J. Implementation of molecular markers for quantitative traits in breeding programs – change and opportunities.

In: INTERNATIONAL CROP SCIENCE CONGRESS, 4., 2004, Brisbane, Australia. **New Directions for a Diverse Planet**: proceedings. Queensland: The Regional Institute Ltd., 2004. Editors Tony Fisher, Neil Turner, John Angus, Lynne McIntyre, Michael Robertson, Andrew Borrell, and David Lloyd. Available at: <www.cropscience.org.au/icsc2004/>. Accessed on: Apr. 30 2008.

HWANG, E.-Y.; SONG, Q.; JIA, G.; SPECHT, J.E.; HYTEN, D.L.; COSTA, J.; CREGAN, P.B. A genome-wide association study of seed protein and oil content in soybean. **BMC Genomics**, v.15, p.1-12, 2014. DOI: 10.1186/1471-2164-15-1.

ILLUMINA, INC. **Sequencing and array-based solutions for genetic research**. Available at: <http://www.illumina.com/>. Accessed on: Sept. 10 2014.

JUN, T.H.; VAN, K.; KIM, M.Y.; LEE, S.-H.; WALKER, D.L. Association analysis using SSR markers to find QTL for seed protein content in soybean. **Euphytica**, v.162, p.179-191, 2008. DOI: 10.1007/s10681-007-9491-6.

LEAMY, L.J.; ZHANG, H.; LI, C.; CHEN, C.Y.; SONG, B.-H. A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). **BMC Genomics**, v.18, p. 1-15, 2017. DOI: 10.1186/s12864-016-3397-4.

MACKAY, I.; POWELL, W. Methods for linkage disequilibrium mapping in crops. **Trends in Plant Science**, v.12, p.57-63, 2007. DOI: 10.1016/j.tplants.2006.12.001.

MASUDA, T.; GOLDSMITH, P.D. World soybean production: area harvested, yield, and long-term projections. **International Food and Agribusiness Management Review**, v.12, p.143-161, 2009.

MORAES, R.M.A. de; JOSÉ, I.C.; RAMOS, F.G.; BARROS, E.G. de; MOREIRA, M.A. Caracterização bioquímica de linhagens de soja com alto teor de proteína. **Pesquisa Agropecuária Brasileira**, v.41, p.725-729, 2006. DOI: 10.1590/S0100-204X2006000500002.

NICHOLS, D.M.; GLOVER, K.D.; CARLSON, S.R.; SPECHT, J.E.; DIERS, B.W. Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. **Crop Science**, v.46, p.834-839, 2006. DOI: 10.2135/cropsci2005.05-0168.

NORDBORG, M.; TAVARÉ, S. Linkage disequilibrium: what history has to tell us. **Trends in Genetics**, v.18, p.83-90, 2002. DOI: 10.1016/S0168-9525(02)02557-X.

PANTHEE, D.R.; PANTALONE, V.R.; WEST, D.R.; SAXTON, A.M.; SAMS, C.E. Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. **Crop Science**, v.45, p.2015-2022, 2005. DOI: 10.2135/cropsci2004.0720.

SCHUSTER, I. Marker-assisted selection for quantitative traits. **Crop Breeding and Applied Biotechnology**, v.11, p.50-55, 2011. Número especial. DOI: 10.1590/S1984-70332011000500008.

SCHUSTER, I.; QUEIROZ, V.T. de; TEIXEIRA, A.I.; BARROS, E.G. de; MOREIRA, M.A. Determinação da pureza varietal de sementes de soja com o auxílio de marcadores moleculares microssatélites. **Pesquisa Agropecuária Brasileira**, v.39, p.247-253, 2004. DOI: 10.1590/S0100-204X2004000300007.

SOARES, T.C.B.; GOOD-GOD; P.I.V.; MIRANDA, F.D. de; SOARES, Y.J.B.; SCHUSTER, I.; PIOVESAN, N.D.; BARROS, E.G. de; MOREIRA, M.A. QTL mapping for protein content in soybean cultivated in two tropical environments. **Pesquisa Agropecuária Brasileira**, v.43, p.1533-1541, 2008. DOI: 10.1590/S0100-204X2008001100012.

SOARES, T.C.B.; PIOVESAN, N.D.; SCHUSTER, I.; CRUZ, C.D.; BARROS, E.G. de; MOREIRA, M.A. Quantitative genetic analysis of storage proteins in soybean. **Crop Breeding and Applied Biotechnology**, v.4, p.317-324, 2004. DOI: 10.12702/1984-7033.v04n03a09.

USDA. United States Department of Agriculture. **SoyBase and the Soybean Breeder's Toolbox**. Available at: <https://www.soybase.org/>. Accessed on: Oct. 24 2016.

ZHANG, W.-K.; WANG, Y.-J.; LUO, G.-Z.; ZHANG, J.-S.; HE, C.-Y.; WU, X.-L.; GAI, J.-Y.; CHEN, S.-Y. QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. **Theoretical and Applied Genetics**, v.108, p.1131-1139, 2004. DOI: 10.1007/s00122-003-1527-2.

Received on October 25, 2016 and accepted on March 29, 2017