Euphytica (2019) 215:54
https://doi.org/10.1007/s10681-019-2378-5

Genome analysis to identify SNPs associated with oil content and fatty acid components in soybean

R. H. G. Priolli · C. R. L. Carvalho · M. M. Bajay · J. B. Pinheiro · N. A. Vello

Received: 28 May 2018 / Accepted: 13 February 2019 / Published online: 19 February 2019

Abstract The nutritional value, flavor and stability of soybean oil are determined by its five dominant fatty acids: saturated palmitic and stearic, monounsaturated oleic, and polyunsaturated linoleic and linolenic acids. Identifying molecular markers or quantitative trait loci associated with these components has the potential to facilitate the development of improved varieties and thus improve soybean oil content and quality. In this study, we used the BARCSoySNP6K BeadChip array to conduct a genome analysis of diverse soybean accessions evaluated for 2 years under Brazilian field conditions. The results demonstrated high broad-sense heritability, suggesting that the soybean genotype panel could be useful for oil trait breeding programs. Moreover, the range of oil trait variation among the plant introductions (PIs) was superior to that among the Brazilian cultivars in this study, indicating that a PI population could be used to find genes controlling these traits. The genome analysis showed that the genetic structure of the soybean germplasm comprised two main genetic groups, and it revealed linkage disequilibrium decay of approximately 300 kb. A total of 19 single-nucleotide polymorphism (SNP) loci on ten different chromosomes significantly associated with palmitic acid, oleic acid and total oil contents were discovered. Analysis of the SNP annotations revealed enzymes associated with several oil-related physiological metabolisms. Loci and specific alleles in our soybean panel that contributed to lower palmitic acid contents and higher oleic acid and total oil contents were identified. Overall, this genome analysis confirmed previous findings and identified SNP markers that may be useful to rapidly improve oil traits in soybean.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10681-019-2378-5) contains supplementary material, which is available to authorized users.

R. H. G. Priolli (✉)
FIFO, UNISANTA, R. Oswaldo Cruz 266, Santos,
SP 11045-970, Brazil
e-mail: rhpriolli@gmail.com

R. H. G. Priolli
NEPA, UNICAMP, Av. Albert Einstein 291, Campinas,
SP 13083-852, Brazil

C. R. L. Carvalho
IAC, CPDRGV, Av. Barao de Itapura 1481, Campinas,
SP 13001-970, Brazil

M. M. Bajay
CERES, UDESC, R. Cel. Fernandes Martins 270, Laguna,
SC 88790-000, Brazil

J. B. Pinheiro · N. A. Vello
Department of Genetics, ESALQ, USP, Av Padua Dias
11, Piracicaba, SP 13400-970, Brazil

Keywords *Glycine max* · Soybean germplasm · Fatty acids · Oil quality · Molecular markers · Soybean breeding

Introduction

Oil content and quality have drawn much attention in soybean genetics and breeding programs due to the increased demand for vegetable oils. The oil fraction corresponds to 20% of dry mass in the seeds of cultivated soybean (*Glycine max* L. Merr.) and is mainly (95%) directed toward the consumption of edible oil; the remainder is used for industrial products such as fatty acids, soaps and biodiesel (http://www.soyatech.com/soyfats). The nutritional value, flavor and stability of soybean oil are determined by its five dominant fatty acids: saturated palmitic (16:0) and stearic (18:0), monounsaturated oleic (18:1), and polyunsaturated linoleic (18:2) and linolenic (18:3) acids. The average percentages of these five fatty acids in soybean oil are 10%, 4%, 18%, 55%, and 13%, respectively. Previous research has shown that decreasing saturated (16:0 and 18:0) and polyunsaturated fatty acids (18:2 and 18:3) and increasing monounsaturated acids (18:1) improves the health benefits of soybean oil for human consumption (Wilson et al. 2002). For certain industrial applications, such as biodiesel production, the development of an oil high in oleic acid and low in saturated fatty acids has been suggested to simultaneously improve oxidative stability and augment cold flow (Aransiola et al. 2014). Significant efforts have been made to increase the oxidative stability of soybean oil as a means to avoid the trans fats generated through the hydrogenation process, enhance the ω-3 fatty acid content of the oil for use in both food and feed applications and increase the total oil content of the seeds (Graef et al. 2009; Clemente and Cahoon 2009).

Previous studies have shown that the Brazilian soybean germplasm has a narrow genetic base (Hiromoto and Vello 1986) with only five ancestors, representing approximately 60% of the overall genetic base of the soybean (Wysmierski and Vello 2013). In this context, the characterization and introduction of new sources of genes represents a crucial step in fostering efficient breeding strategies and, consequently, the development of new cultivars to improve soybean oil content and quality.

The oil content and fatty acid components (hereafter referred to as oil traits for simplicity) of soybean seeds behave as quantitative traits. Identifying molecular markers or quantitative trait loci (QTLs) associated with oil traits using marker-assisted selection (MAS) has the potential to facilitate the development of improved varieties. Although oil traits show quantitative inheritance, the cited heritability estimates for these traits are moderate to high (Fehr et al. 1991; Panthee et al. 2006; Hyten et al. 2004), highlighting the utility of identifying genetic markers associated with these traits.

Linkage mapping using biparental mapping populations is one approach for the identification of QTLs using molecular markers, and a number of molecular markers associated with oil traits have been reported (Diers and Shoemaker 1992; Spencer et al. 2003; Monteros et al. 2008; Li et al. 2011). However, the numbers of parents that have been used in previous QTL genetic linkage mapping experiments represent only a very small proportion of the total germplasm of soybean, and it is not known how often QTLs can be detected repeatedly in practical breeding (Mackay et al. 2009).

The genome-wide association study represents an alternative approach to association mapping for finding QTLs and has been widely applied in soybean studies (Deshmukh et al. 2014). This type of study requires a high density of single-nucleotide polymorphisms (SNPs) across the genomes of a diverse number of individuals as well as phenotyping of all the individuals in the study. Significant statistical associations are then determined between SNP alleles and trait phenotypes. Compared with the QTL linkage mapping approach, the association study can greatly increase the range of detection of natural variation, the number of genome-wide significant loci, and the QTL resolution for complex agronomic traits. Through application of this approach, many important QTLs can be localized, and candidate genes associated with oil traits can be identified (Hwang et al. 2014; Vaughn et al. 2014; Li et al. 2015; Cao et al. 2017; Leamy et al. 2017; Smallwood et al. 2017).

The number of soybean genome-wide association studies has substantially increased with the availability of next-generation sequencing (NGS). A large number of SNPs have been developed, mainly for diploid organisms (Koboldt et al. 2013). Array-based SNP genotyping platforms, such as Illumina GoldenGate, Infinium, and Affymetrix Axiom, have permitted the assaying of hundreds to thousands of SNPs in a high-throughput and cost-effective manner. In soybean, a Universal Soy Linkage panel (USLP 1.0), containing 1536 SNPs, was the first developed (Hyten et al. 2010); however, larger arrays, such as a SoySNP50 K Illumina array (Song et al. 2013), 180K AXIOM ® SoyaSNP Affymetrix array (Lee et al. 2015b), and an NJAU 355K SoySNP Affymetrix array (Wang et al. 2016), were subsequently developed from sequence analyses of several cultivated soybeans (*Glycine max* L. Merr.) and wild soybean (*G. soja* Siebold et Zucc.) genotypes. A set of 6000 SNPs for a medium-scale Infinium array was selected from the SoySNP50K array to maximally represent haplotype blocks, assess genetic diversity within cultivated soybean and *G. soja*, and facilitate genotyping in the soybean research community (Song et al. 2014). Recently, a soybean tropical collection containing 169 cultivars was genotyped using a high-throughput BARCSoySNP6K BeadChip assay, which provides a high-resolution map of genome-wide markers and can facilitate analysis of complex traits in soybean (Contreras-Soto et al. 2017).

In this study, we conducted a genome study of soybean with 96 diverse accessions genotyped with BARCSoySNP6K BeadChips to identify molecular markers associated with QTL regions for oil traits in soybean. Candidate genes within significant association loci that were potentially involved in the regulation of oil traits were also predicted. In addition, we identified the best alleles for significant oil traits, which can be used by soybean breeders in crossing programs.

Materials and methods

Plant material and field experiments

The association panel for the genome analysis consisted of a diverse collection of 96 soybean accessions (including 62 plant introductions (PIs) and 34 soybean cultivars) originating from different countries of the world (Table S1 and Fig. 1). Accessions were selected to represent a range of germplasm with respect to soybean oil content. Seeds were obtained from the germplasm collections of Embrapa-Soybean (Brazil) and the Department of Genetics, ESALQ, University of Sao Paulo (Brazil).

All accessions were planted and cultivated between November and March of the 2009–2010 and 2010–2011 agricultural years in the experimental area of the Department of Genetics, Piracicaba, Sao Paulo, Brazil. Each plot contained 20 plants, which were planted in rows 1.5 m in length and spaced 0.8 m from the nearest plots. In both experimental years, a Federer augmented design was used, with the genotypes organized in two experimental sets with common checks.

Total oil was extracted and analyzed using a Butt apparatus and hexane as the solvent. Measurements of the five fatty acids were conducted by gas chromatography (chromatograph model 3900, Varian, Palo Alto, CA). For each accession in both years, the average values of seed oil and fatty acid contents from three replicates were used for the association analysis (Priolli et al. 2015).

Genotyping

Seeds of each accession were planted in seedling plates in standard soil mix. Plants were grown in the greenhouse (24–25 °C, approximately 33% humidity). Total genomic DNA was isolated from lyophilized leaf tissue bulked from five plants per accession using the DNeasy Plant Kit (Qiagen). DNA concentration was quantified with a spectrophotometer (NanoDrop Technologies Inc., Centerville, DE, USA) and normalized at 50 ng/µl for marker genotyping.

SNP genotyping was performed at Centro de Genômica Funcional ESALQ/USP in Piracicaba, Sao Paulo, Brazil, using BARCSoySNP6K Illumina Infinium BeadChips (Illumina, Inc., San Diego, CA, USA). The assay consisted of a series of standard protocols, such as incubation, DNA amplification, hybridization of samples to the bead assay, extension, and imaging of the bead assay (Song et al. 2013). The SNP alleles were called using the Illumina Genome Studio Genotyping Module (Illumina, Inc. San Diego, CA). Data were first filtered excluding redundant SNPs, nonpolymorphic SNPs and SNPs with more than 10% missing data to calculate population structure and principal component analysis resulting in 5914 SNP loci. In addition, for the association

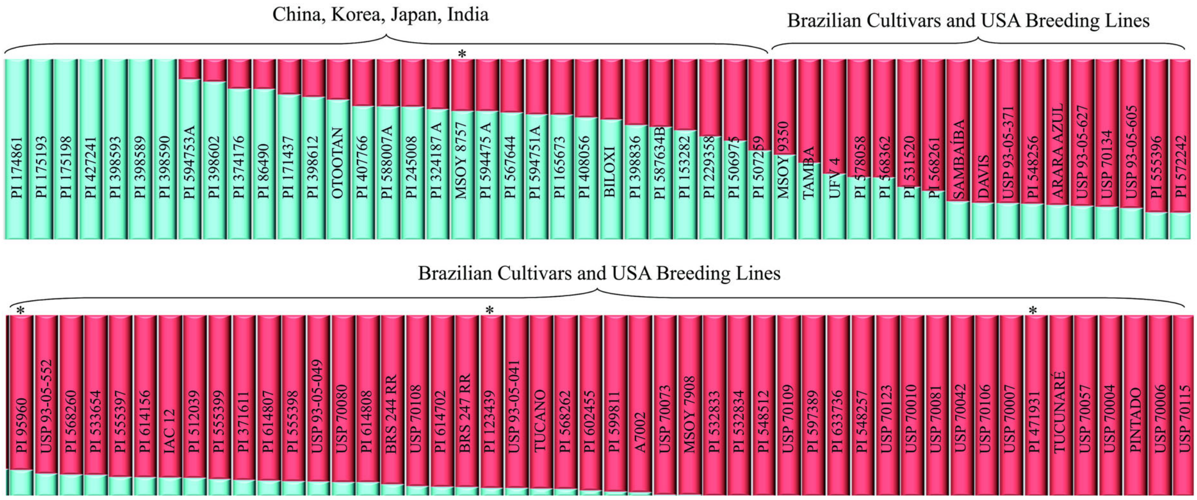

Fig. 1 Two germplasm clusters, red and blue, based on Bayesian analysis of the 96 soybean accessions analyzed by using 5220 SNP loci. Details of the identification of accessions and their geographical origins are indicated. '*' denotes an accession present in a cluster that does not correspond to its origins

analysis, we also excluded SNP loci with minor allele frequencies of less than 1% and SNP loci with more than 25% missing data, retaining 5520 SNP loci. As recommended by Hwang et al. (2014), all heterozygous loci were treated as missing data.

Statistical analysis

The 2-year data were averaged for each trait. Kolmogorov–Smirnov two-sample statistical tests (K–S test) (Snedecor and Cochran 1989) were applied to test for heterogeneity in the distribution of each trait in a year.

The trials with genotypes organized in experimental sets with regular treatments and common checks were analyzed according to Zimmermann (2014). The statistical model was based on an augmented randomized complete block design and analyzed using the equation Yij = m + ti + bj + eij, where Yij is the observation of the ith treatment in the jth block, with j = 1, 2, …, b; I = 1, 2, …, p, p + 1, p + 2, …, p + t, where p is the number of progeny or regular treatments, t is the number of checks, and p + t = v is the total number of treatments; m is the general mean; ti is the effect of the ith treatment, with I = 1, 2, …, p, p +1, p +2, …, p + t; bj is the effect of the jth block, with j = 1, 2, …, b; and eij is the normally distributed random effect. The analyses of individual and joint variances were carried out using the restricted maximum likelihood (REML) method, considering all parameters of the model as random. Data were estimated by combining the 2 years using the LME4 R package (R Development Core Team 2015). Broad-sense heritability (BSH) was calculated with the formula BSH = σ^2G/(σ^2G + $\sigma^2\varepsilon$/n), where σ^2G is the genotype variance, $\sigma^2\varepsilon$ is the error component variance, and n represents years (Nyquist 1991).

The genetic structure based on the 5914 SNPs was investigated using a Bayesian model-based Markov chain Monte Carlo (MCMC) clustering method implemented using the program STRUCTURE v. 2.3.3 (Pritchard et al. 2000; Hubisz et al. 2009). The following parameters were applied to the analysis: diploid locus, admixture model and correlated allele frequencies. Following a burn-in period of 50,000, five independent runs were performed for each K value (from 1 to 10), with 500,000 iterations, as previously optimized (Priolli et al. 2015). The true value of K(ΔK) was chosen according to the method of Evanno (Evanno et al. 2005) using STRUCTURE HARVESTER 0.6.7 (Earl and vonHoldt 2012). Graphs of the STRUCTURE results were produced using CLUMPP (Jakobsson and Rosenberg 2007). We used a Q matrix from the structure to assign individuals to different Ks (referred to here as 'clusters' for simplicity) using a critical level of > 50% for each. Principal components analysis (PCA) was conducted in R using the APE (Paradis et al. 2004) and GGPLOT2 packages (Ginestet 2011). Genetic diversity was estimated using the package diveRsity version 1.9.90 (Keenan et al. 2013) in the R software (R Development Core Team 2015). Variations in allelic frequencies were quantified using F_{ST}. The statistical significance of departures from zero was tested using bootstrapping over the loci in the R package diveRsity.

The linkage disequilibrium (LD) block structure was examined using 5220 loci in TASSEL 5.0 software (Bradbury et al. 2007) by estimating the squared frequency correlation (r^2) of alleles in each chromosome. Nonlinear regression curves were used to estimate the LD decay with distance, and the LD decay rate was determined as the physical distance between markers at which the average r^2 dropped to half its maximum value.

The BSH, total sample size, number of SNPs and average two-locus LD (r^2) between SNP markers were estimated to calculate the statistical power of each association analysis by using the GWAPower package (Feng et al. 2011).

A compressed mixed linear model (Zhang et al. 2010) incorporating the trait data, population structure (Q matrix) and pairwise kinship (K matrix) was used to identify marker-trait associations using the TASSEL program. The K matrix was automatically obtained by the centered-IBS method using TASSEL. We also generated quantile–quantile (QQ) plots of the observed versus expected *P* values at each SNP. Markers were identified as significantly associated with traits based on a significance threshold of $P < 1.916 \times 10^{-4}$, where *P* value < 1/n (n = number of markers). Manhattan plots of -log10 (*P*) values for each SNP vs. chromosomal position were generated from the TASSEL results. Genes with known functional descriptions related to SNP peaks were selected as candidate genes using the Wm82 Genome Browser of SoyBase (https://soybase.org/).

Some specific locus alleles were significantly associated with certain oil traits, and the contributions of these alleles to the phenotypic values were assessed. To graphically evaluate the associations of the polymorphisms, a binary logistic regression model was built using the GGPLOT2 package.

Results

Phenotypic data

Oil traits in soybean are complex traits controlled by both genetic and environmental factors that require multiple phenotypic scoring. As shown in Supp. Table S1, the mean value for oil content in 2010 over all accessions was 18.91% of seed dry mass, and the soybean oil concentrations showed means of 10.70, 3.27, 24.10, 53.01 and 6.40% for palmitic, stearic, oleic, linoleic and linolenic fatty acids, respectively. In 2011, the corresponding means were 18.83, 10.48, 3.16, 24.98, 52.62 and 6.31%, respectively. Normal distribution testing according to the K–S two-sample test ($P < 0.05$) showed that the frequency distribution for the oil traits in each year did not depart from normality. The ANOVA (Supp. Table S2) showed significant effects of genotype across different environments in total oil, palmitic, stearic, oleic, linoleic and linolenic acid contents. However, the high BSH estimates (Table 1) indicated that the phenotypic values in these 2 years were relatively stable for the different accessions, suggesting that there were major genetic components conditioning the oil traits in this population.

To assess the oil breeding potential of the panel, PIs and cultivars were analyzed separately (Table 2). The means of oil content, palmitic acid and linoleic acid were higher for the cultivars than for PIs. However, the range of variation among the PIs was two- to threefold higher than that of the cultivars for all oil traits, and the extreme values belonged to the PI group. A higher oleic acid content (> 50% concentration in soybean oil) was observed in accessions PI 531520, PI 568261 and PI 568260. For palmitic acid content, the lowest values (~ 5% soybean oil) were observed in PI 599811, PI 602455 and PI 568260. PI 531520 showed the lowest value of linolenic acid and a high value of oleic acid. PI 471931 had the highest oil content. Notably, all these oil components are important in soybean breeding programs aimed at oil quality.

Genotyping, population structure and linkage disequilibrium

A total of 5220 SNP loci distributed in the soybean genome were selected based on BARCSoySNP6K genotyping. These SNPs covered a region of 947 Mb

Table 1 Maximum and minimum oil trait values (% dry mass in seeds and % concentration in soybean oil) observed in 96 soybean accessions

	Oil content	Palmitic acid	Stearic acid	Oleic acid	Linoleic acid	Linolenic acid
Max	23.30	15.50	4.55	56.18	61.10	11.56
Min	12.85	3.24	2.16	13.32	28.16	2.41
Mean	18.87	10.59	3.21	24.54	52.81	6.36
STD	2.49	1.67	0.43	8.05	6.26	1.46
G	*	*	*	*	*	*
BSH	0.89	0.91	0.64	0.87	0.88	0.77

Means, variances and heritabilities evaluated for 2 years under Brazilian field conditions are also shown

G genotype across different environments, *STD* standard deviation, *BSH* broad-sense heritability

*Significant at $P < 0.001$

Table 2 Seed oil (% dry mass) and fatty acid contents (% in soybean oil) for the soybean cultivars and plant introductions

	Cultivar				Plant introduction			
	Mean ± SE	CV (%)	Range	Range variation	Mean ± SE	CV (%)	Range	Range variation
Oil content	20.60 ± 0.86	4	18.86–23.11	4.25	17.93 ± 2.58	14	12.85–23.30	10.45
Palmitic acid	10.85 ± 0.67	6	9.09–12.02	2.93	10.45 ± 2.00	19	3.24–15.50	12.27
Stearic acid	3.18 ± 0.40	13	2.49–4.30	1.81	3.23 ± 0.45	14	2.16–4.55	2.39
Oleic acid	23.21 ± 4.55	20	15.78–37.35	21.57	25.27 ± 9.35	37	13.32–56.18	42.86
Linoleic acid	54.25 ± 3.56	7	42.57–61.08	18.51	52.03 ± 7.22	14	28.16–61.10	32.94
Linolenic acid	6.11 ± 0.87	14	4.31–7.84	3.53	6.50 ± 1.67	26	2.41–11.59	9.18

Mean values for 2 years under Brazilian field conditions are shown. CV (%) = (std/mean)*100

in the soybean genome, which represents 86% of the 1100-Mb soybean genome. SNP markers were identified on each chromosome, with the number of markers ranging from 211 (chromosome 12) to 336 (chromosome 13) and averaging 261 (Supp. Table S3). These values indicated that the Illumina Infinium platform identified SNPs that were well distributed throughout the soybean genome.

Using SNP loci, we performed a Bayesian clustering analysis (STRUCTURE) to determine the population structure of our panel. According to Evanno's method, the most likely K value (number of clusters) was K = 2 (Fig. 1), with 65 and 31 individuals predicted in each cluster. Notably, clusters 1 (red) and 2 (blue) corresponded to accessions from America (Brazil and the USA) and Asia (China, Korea, Japan, and India). To quantify the population structure of the panel, we performed principal component analysis (PCA) (Fig. 2). The dispersion plots of the first and second principal components explained 9.47% and 6.08% of the variance between the accessions, with a clear discrimination of clusters according to the Bayesian division. The two-dimensional PCA plot suggested a broader genetic base in cluster 2 (Asian materials) than cluster 1 (American accessions). The pedigrees of the soybean accessions (Supp. Table S4) confirmed the narrower genetic base of cluster 1, which consists mainly of cultivars and soybean breeding material.

Despite the presence of only two clusters, which might have suggested less genetic diversity among the genotypes of the soybean panel, the measure of differentiation, F_{ST}, was estimated at 0.1135, indicating that, despite their low number, the two clusters were significantly different (Supp. Table S5), and this contrast could be useful for identifying loci in

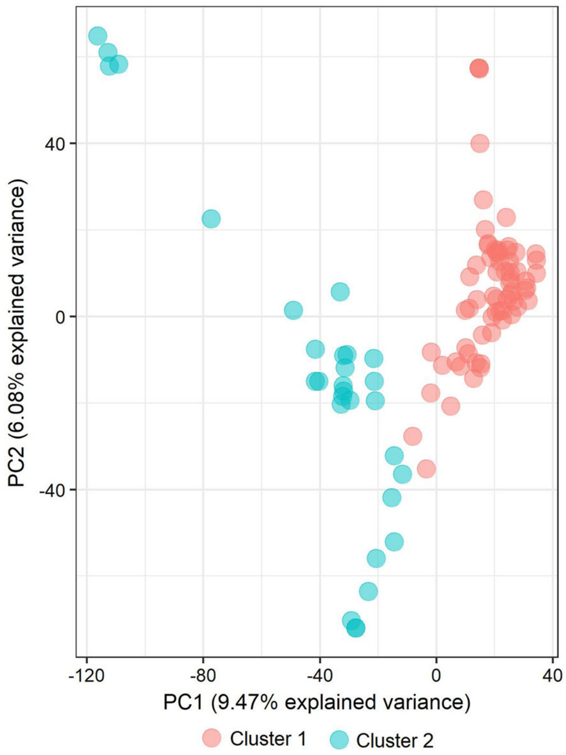

Fig. 2 Dispersion plot of the first (9.47%) and second (6.08%) principal components based on analysis of soybean accessions using 5220 SNP loci. The red and blue points represent accessions from America (1) and Asia (2), respectively, according to the clusters identified using STRUCTURE

association analysis. Cluster 1 showed the highest number of individuals and number of alleles, but allelic richness among the clusters was similar and not significant, confirming that both presented the same genetic variability.

The distribution of the correlation coefficients (r^2) between SNPs located at different physical distances on each chromosome can be observed in Supp. Figure 1. Slow LD decay was observed with increased distance (Kb) in all 20 chromosomes, with the presence of large blocks in LD in each chromosome. The haplotype blocks spanned between 15,000 and 30,000 Kb, except that of chromosome 19, which spanned 6000 Kb. The LD block structures of all the chromosomes (Fig. 3) showed that the r^2 value declined as the physical distance between the loci increased. The decay of LD with physical distance between SNPs occurred at approximately 300 Kb (r^2 = 0.16), suggesting structure of the soybean genome within this distance.

Genome analysis

The GWAPower simulation indicates the sample size required to reach the maximum power for an analysis. Considering the parameters specific to the present study, such as SNP number (5220 loci), LD of 0.16 and heritability of 0.8267 (average for all oil traits), the minimum adequate sample size is 96 individuals. This result indicates that our analysis based on these soybean accessions was adequate to obtain maximum resolution.

Using a linear mixed model (MLM) with corrections for multiple tests, the totals of 1, 16 and 2 SNPs for oil content, palmitic acid and oleic acid, respectively, exceeded the threshold of significance ($-\log_{10} P \geq 3.72$) (Fig. 4). The 48.10 Mb position on chromosome 19 showed the highest level of significance (P value = 7.61×10^{-7}), comprising one SNP associated with palmitic acid content. Chromosomes 8 and 12 showed the most SNPs, three in each, which were associated with palmitic acid content. Chromosomes 10 and 18 showed one SNP each associated with oleic acid content, and chromosome 9 showed one SNP associated with oil content. No overlap was found between the loci associated with these traits; however, two regions, one on chromosome 8 and the other on chromosome 15, showed SNPs associated with palmitic acid content that were less than 0.5 Mb (500 kb) apart. The distribution of the QQ plots of total oil, palmitic acid and oleic acid content (Supp. Figure 2) showed values in a normal curve, adequate for the compressed MLM model to reduce false positives in the significant traits.

Based on the association analysis and the genes annotated in SoyBase (www.soybase.org), we identified causal genes for loci significantly associated with each trait (Table 3). Although many of the SNP loci were in intergenic regions, 26% were in coding regions (CDS), introns or the 5′ UTRs of genes with functional annotation. These genes included genes involved in fatty acid metabolism and regulation, such as genes encoding methyltransferase, translation-initiation factor, glycosyltransferase, kinase protein and storage proteins.

To identify alleles associated with the three significant traits, the most significant SNP loci of each trait were selected, and the contribution of each allele to the trait value was recorded. The results of the binary logistic regression indicated that the frequency of the

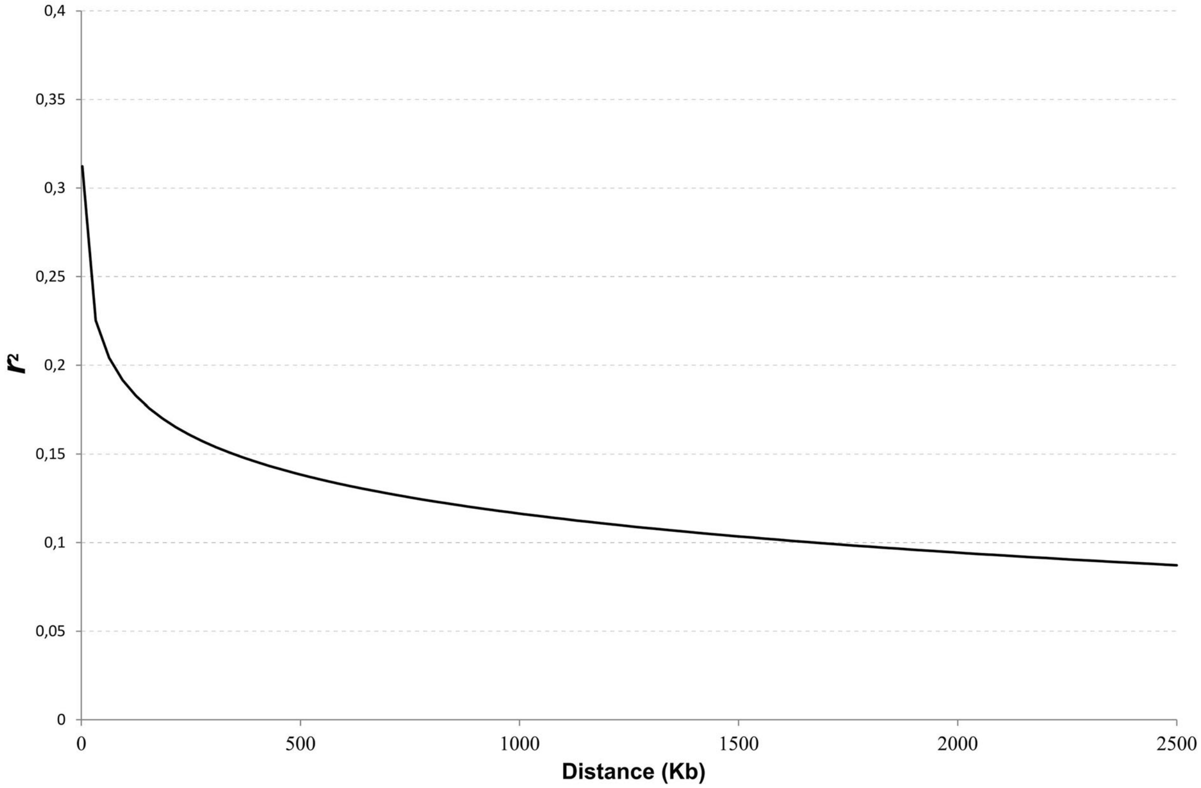

Fig. 3 Pairwise LD values (r^2) plotted against genetic distance estimated among 5220 SNP loci and 96 soybean accessions

C allele in ss715635790 (Fig. 5) decreased as palmitic acid content increased, whereas for the s715629367 and ss715603267 loci, the frequency of the C allele increased as oleic acid content and oil content increased. Although they are located in regions that do not contain any previously discovered QTLs or genes affecting these traits, these alleles belong to SNP loci on different chromosomes and may prove valuable for future breeding-by-design of soybean lines to enhance oil content and/or soybean oil quality.

Discussion

Our findings show SNP loci associated with oil traits in soybean and that our soybean panel can be useful to identify such polymorphisms. Consistent with our results, previous studies have identified SNP loci associated with the five main fatty acids in soybean using universal SNP chips and a similar experimental setup (Li et al. 2015; Leamy et al. 2017). The current work also complements a growing body of work demonstrating the power of genome-wide association studies to identify molecular markers associated with oil content in soybean (Hwang et al. 2014; Vaughn et al. 2014; Cao et al. 2017) and provides data specific to Brazilian field conditions.

Breeding for oil traits is focused on the quantity and quality of soybean oil, including the contents of the five main fatty acids. Because the genetics of these traits are well known, and desired sources of germplasm are available, oil traits can be manipulated in a breeding program. The high BSH values that we document here suggest that a 96-soybean panel can be useful for oil trait breeding programs. Previous studies have reported heritability estimates for oil traits that are moderate to high, although these traits are quantitative traits controlled by multiple genes (Fehr et al. 1991; Panthee et al. 2006; Hyten et al. 2004).

The range of variation for oil traits was greater among the PIs than among the Brazilian cultivars in this study, indicating that the PI population can be used to find genes controlling these traits. The expansion of genetic diversity by incorporating alleles

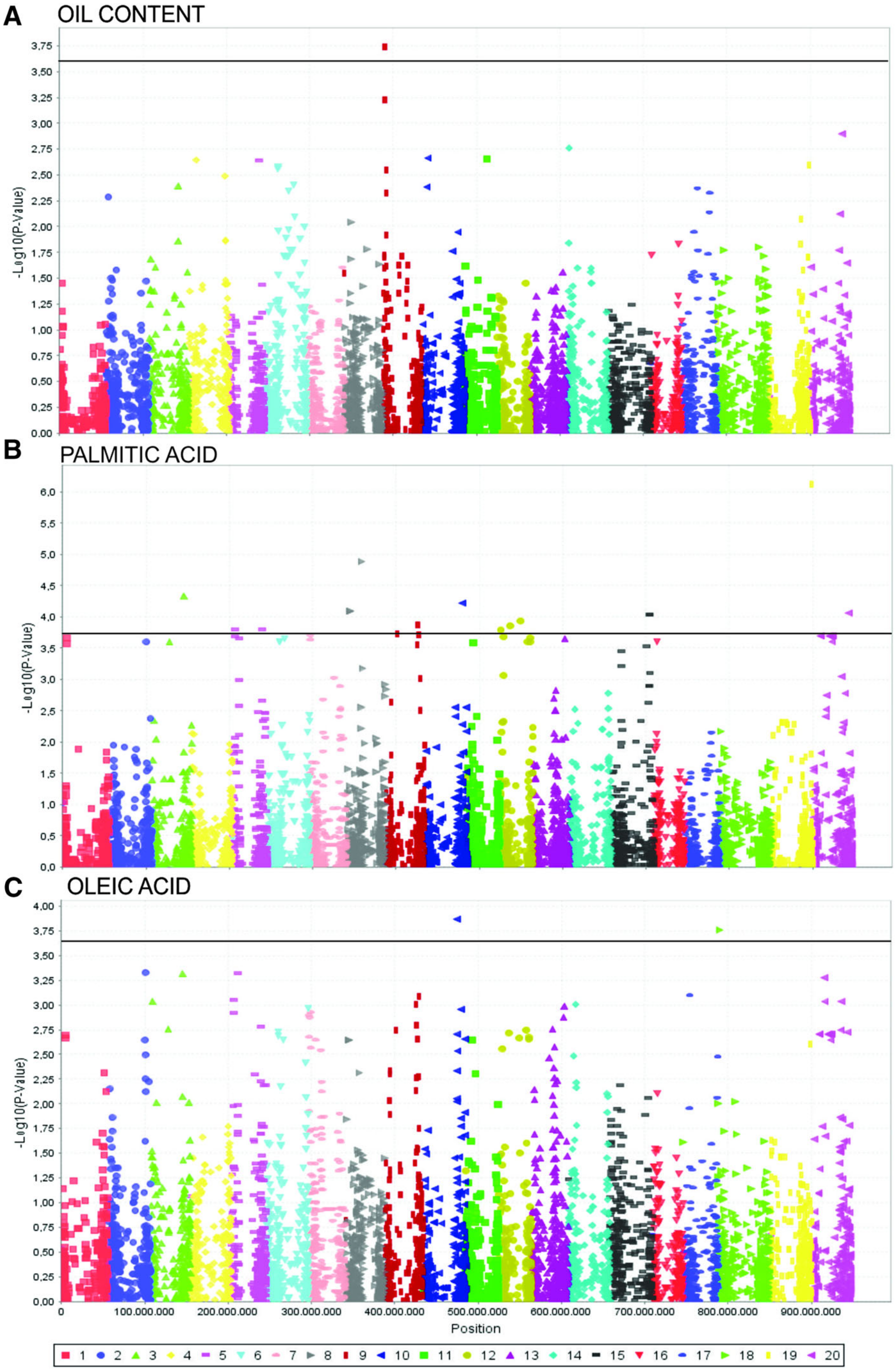

Fig. 4 Manhattan plots of genome-wide association study for oil traits in soybean. Negative log10-transformed *P* values from a genome-wide scan by using mixed linear models (MLM) for oil content (**a**), palmitic acid (**b**) and oleic acid (**c**) are plotted against positions on each of the 20 chromosomes. The significant trait-associated SNPs (Bonferroni adjusted) are distinguished by the threshold line

from PIs has been proposed in several studies from countries with a narrow genetic base for soybean, such as Brazil and the USA (Hiromoto and Vello 1986; Gizlice et al. 1994; Sneller 2003; Wysmierski and Vello 2013). However, PIs have presented low agronomic value in relation to seed yield, and genetic diversity and agronomic value are independent traits (Sneller 1994). To avoid the low agronomic potential of some PIs, researchers have advised selection of those soybean lines that present the best agronomic characteristics prior to using them in breeding

Table 3 SNPs significantly associated with oil traits and predicted candidate genes

Trait	SNP name[a]	Chromosome	Location[b]	*P* value	R^2 marker[c]	SNP biallele	Position	Gene and functional annotation[d]
Oil content	ss715603267	9	2037326	1.82E−04	0.1796	C/T	Intergenic	
Palmitic acid	ss715585707	3	36236670	4.59E−05	0.2681	A/G	Intergenic	
	ss715590297	5	4395428	1.59E−04	0.2294	T/G	Intergenic	
	ss715591234	5	35961486	1.58E−04	0.2310	A/G	Intron	Glyma05g169100; the cupin superfamily, seed storage proteins
	ss715599971	8	17747858	1.31E−05	0.3083	C/T	5′ UTR	Glyma08g23340; serine/threonine protein kinase
	ss715601594	8	3785831	8.23E−05	0.2467	A/G	CDS	Glyma08g048600; methyltransferase activity
	ss715601602	8	3822597	8.23E−05	0.2467	C/T	Intergenic	
	ss715603045	9	14230692	1.87E−04	0.2220	G/T	Intergenic	
	ss715603976	9	40507917	1.33E−04	0.2340	C/A	Intergenic	
	ss715607504	10	45629506	6.01E−05	0.2555	T/C	CDS	Glyma10g225800; translation-initiation factor
	ss715611451	12	12706240	1.39E−04	0.2307	A/G	Intergenic	
	ss715611484	12	1341187	1.64E−04	0.2278	A/G	Intergenic	
	ss715611893	12	21177984	1.16E−04	0.2357	G/A	Intergenic	
	ss715622072	15	46801411	9.26E−05	0.2522	T/C	Intergenic	
	ss715622144	15	47326917	9.24E−05	0.2481	T/C	Intergenic	
	ss715635790	19	48195830	7.61E−07	0.3935	A/C	Intergenic	
	ss715638459	20	43815687	8.65E−05	0.2444	T/C	Intergenic	
Oleic acid	ss715606967	10	40581832	1.37E−04	0.2379	A/C	CDS	Glyma10g172200; granule bound starch synthase Ia: Glycosyltransferase
	ss715629367	18	1789189	1.74E−04	0.2260	C/T	Intergenic	

[a] SoySNP50K nomenclature (https://soybase.org)

[b] Location in base pairs for the peak SNP is provided according to *Glycine max* Wm82.a2v1

[c] Proportion of variance explained for marker-trait associations (R^2) according to Tassel software

[d] *Glycine max* genome assembly version Glyma.Wm82.a2 (Gmax2.0) (https://soybase.org)

programs (Vello et al. 1984; Sneller 1994; Wysmierski and Vello 2013).

Genome analysis using BeadChip platforms has allowed the evaluation of the genetic structure of soybean germplasm based on a large number of markers (Hyten et al. 2010; Song et al. 2013; Lee et al. 2015b; Wang et al. 2016). The two main genetic groups identified by the STRUCTURE analysis corresponded to the Asian and American gene pools, as identified in a previous study with 142 microsatellite markers (Priolli et al. 2015). The PCA suggested that the genetic base was higher in cluster 2 (Asian accessions) than cluster 1 (American accessions). These findings are consistent with previous studies using molecular markers that showed substructure based on geographical origin (Ude et al. 2003; Li et al. 2008) as well as studies on the genetic bases of both germplasms (Hiromoto and Vello 1986; Gizlice et al. 1994; Sneller 1994; Wysmierski and Vello 2013).

The extent of LD is an important factor determining the efficiency of association analysis. The decay of LD with physical distance between SNPs occurred at 300 kb (r^2 = 0.16), which is comparable to the results of previous studies (220–270 Kb) that used larger and

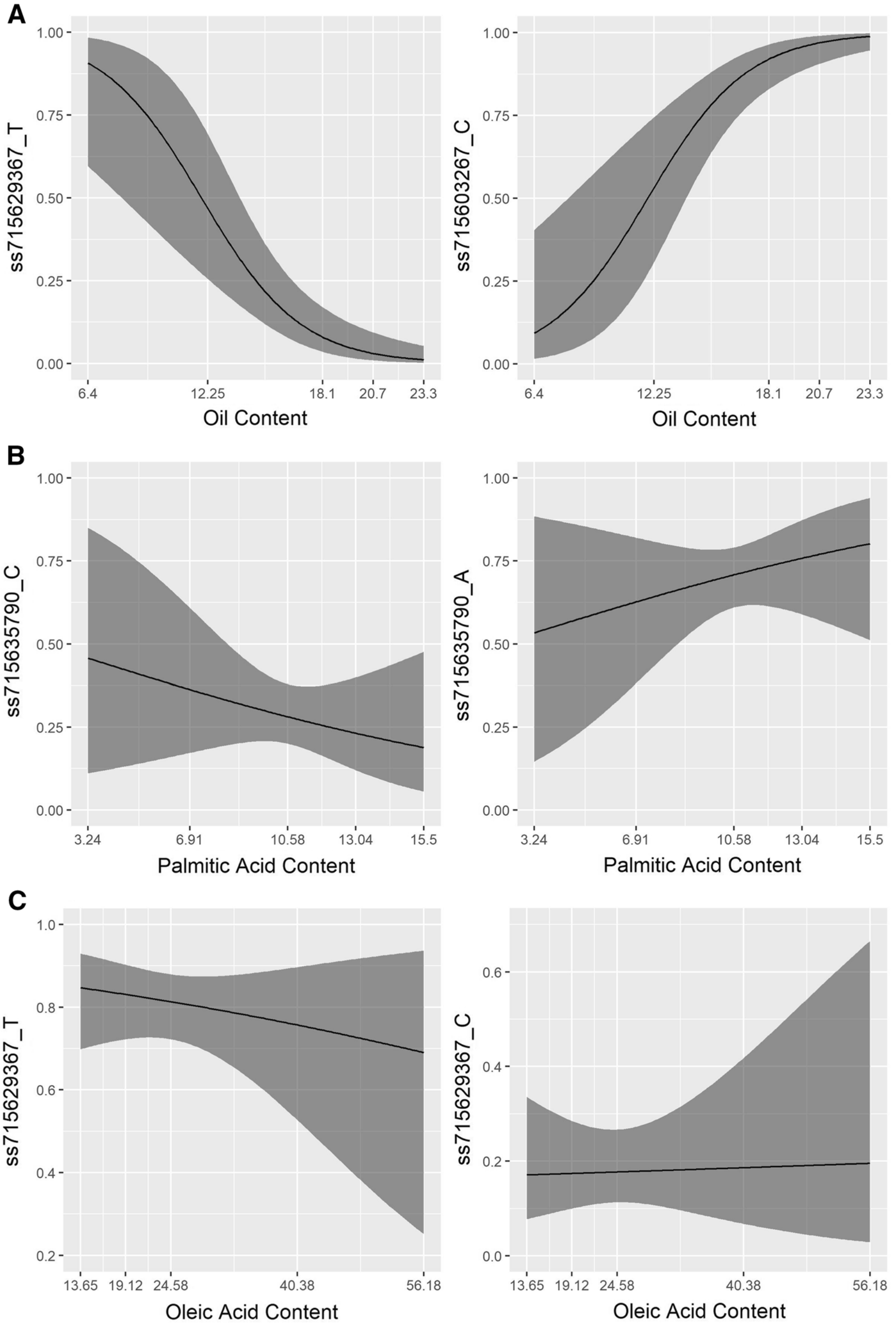

Fig. 5 Fitted logistic regression describing the associations between oil traits and three SNP polymorphisms in soybean panel **a** ss715603267: T-to-C allele in soybean seed oil content; **b** ss715635790: C-to-A allele in palmitic acid content; and **c** ss715629367: T-to-C allele in oleic acid content. Gray shadows show 95% confidence intervals

more genetically diverse populations (Vuong et al. 2015; Zhang et al. 2017). The more genetically diverse the germplasm, the more rapid the expected decay, which provides more opportunity for selection. Although the observed extent of LD was problematic because it resulted in the inclusion of tens to hundreds of candidate genes within an LD block, the results indicated that the use of these accessions had no substantial disadvantage compared to the use of the other sets of soybean germplasm, reaching reasonable resolution. In a study with Brazilian soybean cultivars, the length of the blocks was very similar among chromosomes, with most blocks being 51–500 kb (Contreras-Soto et al. 2017).

We discovered a total of 19 SNPs on ten different chromosomes that were associated with oil traits in our soybean panel. The corrected statistical test for the P values (of $P < 1.916 \times 10^{-4}$, Bonferroni correction) minimized the probability that the null hypothesis was falsely rejected by concentrating on a balance between false and true positives. Because it is stringent, this correction might miss some important associations, as confirmed by the absence of SNP loci associated with stearic, linoleic and linolenic acid content (data not shown).

Six of the sixteen loci significantly associated with palmitic acid were near or in the same linkage groups as previously identified SNPs. For instance, a genome-wide association study conducted with soybean accessions found SNPs located in genes Glyma05g07630 and Glyma12g01380a on chromosomes 5 and 12, respectively (Li et al. 2015). Our study identified SNPs ss715590297 and ss71561484 within 3.0 Kb of these genes. The SNP loci ss715591234, ss715603045, ss715603976, and ss715611451 were also located at a distance less than 4 Mb from genes according to the same study. QTLs in 40.6 and 44.5 Mb positions of chromosomes 9 and 19, respectively, were associated with palmitic acid in a study of linkage mapping (Smallwood et al. 2017), where the SNPs ss715603976 and ss715635790 were identified in our study. Similarly, 4.9 Mb position was found to be associated with oil content in a previous genome-wide association study (Hwang et al. 2014) and is where the ss715603267 locus was found in our study. The region at 9.9 Mb on chromosome 9 was reported twice as associated with oil content on SoyBase based on linkage mapping. Usually, SNPs that are reported in multiple studies using different sources of oil germplasm are good candidate genes for the validation of associations detected via association analysis.

The analysis of the SNP annotations revealed an extensive network of terms associated with several physiological metabolisms that may be associated with the metabolism of oils, such methylation enzymes, translation-initiation factors, glycosyltransferase, kinase protein and storage proteins; however, such terms were associated with only 26% of the genes or candidate genes identified here. In a genome study where an annotated gene approach in a model plant was adopted to design the genotyping array, the majority (93%) of the 1205 SNPs were located in the coding regions (CDS), untranslated regions (UTRs) and introns of 1074 annotated genes (Li et al. 2015). In another study, the soybean reference genome was used to search for all genes associated with seed composition traits of the wild soybean genome, and a total of 29 SNPs were found, of which 8 (27.6%) were located in candidate genes (Leamy et al. 2017). Both strategies, the SNPs developed using the model plant and the utilization of more diverse germplasm, were successful in the identification of candidate genes.

Beyond major gene effects, many QTLs with minor effects on oil traits have been discovered in soybean (Hyten et al. 2004; Lee et al. 2015a; Smallwood et al. 2017). These results can explain one of the probable causes for the presence of SNPs in introns and intergenic regions influencing target traits in our study. Another factor may have been the extension of LD, which in our soybean panel persisted for long distances, suggesting that the use of few markers would have resulted in the detection of additional QTLs. BARCSoySNP6K does not cover all genes in the soybean genome (Song et al. 2014), and one highly significant marker may be either the causative gene itself or in close linkage to the causative gene. Another factor can be the location of the SNP in the soybean genome. According to the authors, the BeadChip was developed using several quality criteria, including the genome region (euchromatic vs. heterochromatic). Although five-sixths of the SNPs came from euchromatic regions, the heterochromatic regions, which have lower numbers of genes, are also present (Song et al. 2014).

It is possible to manipulate the proportions of some fatty acids over a wide range by traditional plant breeding techniques (Graef et al. 2009; Clemente and Cahoon 2009). Increasing oleic acid levels and decreasing linoleic and linolenic acid levels make soybean oil healthier for human consumption. Soybean accessions with reduced palmitic acid are desirable because saturated palmitic acid associated with a diverse lipoprotein profile gives rise to negative health effects in humans (Mensink and Katan 1990). Moreover, to optimize the fuel characteristics of soybean oil for use in biodiesel, it has been suggested that oils that are high in oleic acid and low in palmitic acid should be developed (Graef et al. 2009). Because genome-wide allelic and haplotype data are available for relevant breeding lines and haplotype-trait associations have been established, it may be possible for soybean breeders to undertake breeding-by-design approaches. For example, considering our findings, soybean breeders interested in optimizing the quality characteristics of soybean oil can focus on the three C alleles of the loci ss715635790, ss715629367 and ss715603267, because they can yield soybean seeds with lower palmitic acid content as well as higher oleic acid and total oil contents.

In conclusion, the present study revealed the phenotypic variability of our association panel, indicating the potential of these materials to obtain new combinations of favorable alleles to oil traits, in addition to promoting the amplification of the genetic base for breeding programs. Our analysis also confirmed previous findings and the utility of BARC-SoySNP6K BeadChips for genome analysis and its direct applicability for soybean improvement. In total, 16, 2 and 1 SNP loci were significantly associated with palmitic, oleic and oil content, and their candidate genes were predicted. We suggest that by using favorable alleles, soybean breeders can rapidly improve oil traits in soybean.

Acknowledgements This work was supported by the Sao Paulo Research Foundation (FAPESP, project FAPESP/BIOEN 2016/01823-9). M.M. Bajay thanks Coordination for the Improvement of Higher Education Personnel (CAPES); J.B. P and N.A. Vello thank the National Council for Scientific and Technological Development (CNPq) for scholarships.

References

Aransiola EF, Ojumu TV, Oyekola OO, Madzimbamuto TF, Ikhu-Omoregbe DIO (2014) A review of current technology for biodiesel production: state of the art. Biomass Bioenergy 61:276–297

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635

Cao YC, Li SG, Wang ZL, Chang FG, Kong JJ, Gai JY, Zhao TJ (2017) Identification of major quantitative trait loci for seed oil content in soybeans by combining linkage and genome-wide association mapping. Front Plant Sci 8:1222

Clemente TE, Cahoon EB (2009) Soybean oil: genetic approaches for modification of functionality and total content. Plant Physiol 151:1030–1040

Contreras-Soto RI, de Oliveira MB, Costenaro-da-Silva D, Scapim CA, Schuster I (2017) Population structure, genetic relatedness and linkage disequilibrium blocks in cultivars of tropical soybean (*Glycine max*). Euphytica 213:173

Deshmukh R, Sonah H, Patil G, Chen W, Prince S, Mutava R, Vuong T, Valliyodan B, Nguyen HT (2014) Integrating omic approaches for abiotic stress tolerance in soybean. Front Plant Sci 5:244

Diers BW, Shoemaker RC (1992) Restriction-fragment-length-polymorphism analysis of soybean fatty-acid content. J Am Oil Chem Soc 69:1242–1244

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Fehr WR, Welke GA, Hammond EG, Duvick DN, Cianzio SR (1991) Inheritance of elevated palmitic acid content in soybean seed oil. Crop Sci 31:1522–1524

Feng S, Wang SC, Chen CC, Lan L (2011) GWAPOWER: a statistical power calculation software for genome-wide association studies with quantitative traits. BMC Genet 12:12

Ginestet C (2011) GGPLOT2: elegant graphics for data analysis. J R Stat Soc Ser A Stat Soc 174:245–246

Gizlice Z, Carter TE, Burton JW (1994) Genetic base for North-American public soybean cultivars released between 1947 and 1988. Crop Sci 34:1143–1151

Graef G, LaVallee BJ, Tenopir P, Tat M, Schweiger B, Kinney AJ, Van Gerpen JH, Clemente TE (2009) A high-oleic-acid and low-palmitic-acid soybean: agronomic performance and evaluation as a feedstock for biodiesel. Plant Biotechnol J 7:411–421

Hiromoto DM, Vello NA (1986) The genetic base of Brazilian soybean (*Glycine-max* (L) Merrill) cultivars. Braz J Genet 9:295–306

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9:1322–1332

Hwang EY, Song QJ, Jia GF, Specht JE, Hyten DL, Costa J, Cregan PB (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genom 15:1

Hyten DL, Pantalone VR, Saxton AM, Schmidt ME, Sams CE (2004) Molecular mapping and identification of soybean fatty acid modifier quantitative trait loci. J Am Oil Chem Soc 81:1115–1118

Hyten DL, Choi IY, Song QJ, Specht JE, Carter TE, Shoemaker RC, Hwang EY, Matukumalli LK, Cregan PB (2010) A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci 50:960–968

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806

Keenan K, McGinnity P, Cross TF, Crozier WW, Prodöh PA (2013) diveRsity: an R package for the estimation and exploration of population genetics parameters and their associated errors. Methods Ecol Evol 4:782–788

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. Cell 155:27–38

Leamy LJ, Zhang HY, Li CB, Chen CY, Song BH (2017) A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). BMC Genom 18:18

Lee JD, Bilycu KD, Shannon JG (2015a) Genetics and breeding for modified fatty acid profile in soybean seed oil. J Crop Sci Biotech 10:201–210

Lee YG, Jeong N, Kim JH, Lee K, Kim KH, Pirani A, Ha BK, Kang ST, Park BS, Moon JK, Kim N, Jeong SC (2015b) Development, validation and genetic analysis of a large soybean SNP genotyping array. Plant J 81:625–636

Li YH, Guan RX, Liu ZX, Ma YS, Wang LX, Li LH, Lin FY, Luan WJ, Chen PY, Yan Z, Guan Y, Zhu L, Ning XC, Smulders MJM, Li W, Piao RH, Cui YH, Yu ZM, Guan M, Chang RZ, Hou AF, Shi AN, Zhang B, Zhu SL, Qiu LJ (2008) Genetic structure and diversity of cultivated soybean (*Glycine max* (L.) Merr.) landraces in China. Theor Appl Genet 117:857–871

Li HW, Zhao TJ, Wang YF, Yu DY, Chen SY, Zhou RB, Gai JY (2011) Genetic structure composed of additive QTL, epistatic QTL pairs and collective unmapped minor QTL conferring oil content and fatty acid components of soybeans. Euphytica 182:117–132

Li YH, Reif JC, Ma YS, Hong HL, Liu ZX, Chang RZ, Qiu LJ (2015) Targeted association mapping demonstrating the complex molecular genetics of fatty acid formation in soybean. BMC Genom 16:841

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10:565–577

Mensink R, Katan M (1990) Effect of dietary trans fatty acids on high-density and low-density lipoprotein cholesterol levels in healthy subjects. N Engl J Med 323:439–445

Monteros MJ, Burton JW, Boerma HR (2008) Molecular mapping and confirmation of QTLs associated with oleic acid content in N00-3350 soybean. Crop Sci 48:2223–2234

Nyquist WE (1991) Estimation of heritability and prediction of selection response in plant-populations. CRC Crit Rev Plant Sci 10:235–322

Panthee DR, Pantalone VR, Saxton AM (2006) Modifier QTL for fatty acid composition in soybean oil. Euphytica 152:67–73

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Priolli RHG, Campos JB, Stabellini NS, Pinheiro JB, Vello NA (2015) Association mapping of oil content and fatty acid components in soybean. Euphytica 203:83–96

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

R Development Core Team (2015) R: a language and environment for statistical computing [Internet]. R Foundation for Statistical Computing, Vienna

Smallwood CJ, Gillman JD, Saxton AM, Bhandari HS, Wadl PA, Fallen BD, Hyten DL, Song Q, Pantalone VR (2017) Identifying and exploring significant genomic regions associated with soybean yield, seed fatty acids, protein and oil. J Crop Sci Biotech 20(4):243–253

Snedecor GW, Cochran WG (1989) Statistical methods, 8th edn. Iowa State University Press, Ames

Sneller CH (1994) Pedigree analysis of elite soybean lines. Crop Sci 34:1515–1522

Sneller CH (2003) Impact of transgenic genotypes and subdivision on diversity within elite North American soybean germplasm. Crop Sci 43:409–414

Song QJ, Hyten DL, Jia GF, Quigley CV, Fickus EW, Nelson RL, Cregan PB (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS ONE 8:e54985

Song Q, Jia G, Quigley C, Fickus E, Hyten D, Nelson R, Cregan P (2014) Soybean BARCSoySNP6K Beadchip—a tool for soybean genetics research. In: Plant animal genome XXII, Jan 10–15, 2014, San Diego. Abstract No. P306. https://pag.confex.com/pag/xxii/webprogram/Paper10932.html. Accessed 28 June 2018

Spencer MM, Pantalone VR, Meyer EJ, Landau-Ellis D, Hyten DL (2003) Mapping the Fas locus controlling stearic acid content in soybean. Theor Appl Genet 106:615–619

Ude GN, Kenworthy WJ, Costa JM, Cregan PB, Alvernaz J (2003) Genetic diversity of soybean cultivars from China, Japan, North America, and North American ancestral lines determined by amplified fragment length polymorphism. Crop Sci 43:1858–1867

Vaughn JN, Nelson RL, Song QJ, Cregan PB, Li ZL (2014) The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3 (Bethesda) 4:2283–2294

Vello NA, Fehr WR, Bahrenfus JB (1984) Genetic-variability and agronomic performance of soybean populations developed from plant introductions. Crop Sci 24:511–514

Vuong TD, Sonah H, Meinhardt CG, Deshmukh R, Kadam S, Nelson RL, Shannon JG, Nguyen HT (2015) Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. BMC Genom 16:593

Wang J, Chu SS, Zhang HR, Zhu Y, Cheng H, Yu DY (2016) Development and application of a novel genome-wide SNP array reveals domestication history in soybean. Sci Rep 6:20728

Wilson R, Burton JW, Pantalone VR, Dewey RE (2002) New gene combinations governing saturated and unsaturated FA composition in soybean. In: Kuo TM, Gardner HW (eds) Lipid biotechnology. Marcel Dekker Inc, New York, pp 95–113

Wysmierski PT, Vello NA (2013) The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. Genet Mol Biol 36:547–555

Zhang ZW, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu JM, Arnett DK, Ordovas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355-U118

Zhang HY, Song QJ, Griffin JD, Song BH (2017) Genetic architecture of wild soybean (*Glycine soja*) response to soybean cyst nematode (*Heterodera glycines*). Mol Genet Genom 292:1257–1265

Zimmermann FJP (2014) Estatística aplicada à pesquisa agrícola, 2nd edn. Embrapa, Brasília