Breeding Science 67: 435–449 (2017)
doi:10.1270/jsbbs.17024

Research Paper

Genome-wide association mapping for flowering and maturity in tropical soybean: implications for breeding strategies

Rodrigo Iván Contreras-Soto*[1,5,6], Freddy Mora[2], Fabiane Lazzari[3], Marco Antônio Rott de Oliveira[4], Carlos Alberto Scapim[1] and Ivan Schuster[3]

[1] *Departamento de Agronomia, Universidade Estadual de Maringá*, Av. Colombo, 5790, Maringá PR, 87020-900, Brazil
[2] *Institute of Biological Sciences, University of Talca*, Talca, 3460000, Chile
[3] *Dow Agrosciences*, Rod. Anhanguera S/N Km 330, Cravinhos SP, 14140-000, Brazil
[4] *COODETEC*, BR 467 km 98, 85813-450, Cascavel, PR, Brazil
[5] *Instituto de Ciencias Agronómicas, Universidad de O'Higgins*, Av. Libertador Bernardo O'Higgins 611, Rancagua, 2820000, Chile
[6] *Centro de Estudios Avanzados en Fruticultura*, Camino a Las Parcelas 882 Km 105, Ruta 5 Sur, Rengo, 2940000, Chile

Knowledge of the genetic architecture of flowering and maturity is needed to develop effective breeding strategies in tropical soybean. The aim of this study was to identify haplotypes across multiple environments that contribute to flowering time and maturity, with the purpose of selecting desired alleles, but maintaining a minimal impact on yield-related traits. For this purpose, a genome-wide association study (GWAS) was undertaken to identify genomic regions that control days to flowering (DTF) and maturity (DTM) using a soybean association mapping panel genotyped for single nucleotide polymorphism (SNP) markers. Complementarily, yield-related traits were also assessed to discuss the implications for breeding strategies. To detect either stable or specific associations, the soybean cultivars (N = 141) were field-evaluated across eight tropical environments of Brazil. Seventy-two and forty associations were significant at the genome-wide level relating respectively to DTM and DTF, in two or more environments. Haplotype-based GWAS identified three haplotypes (Gm12_Hap12; Gm19_Hap42 and Gm20_Hap32) significantly co-associated with DTF, DTM and yield-related traits in single and multiple environments. These results indicate that these genomic regions may contain genes that have pleiotropic effects on time to flowering, maturity and yield-related traits, which are tightly linked with multiple other genes with high rates of linkage disequilibrium.

Key Words: quantitative trait loci, pleiotropy, tropical soybean, linkage disequilibrium.

Introduction

Flowering, maturity and plant height are key complex traits determining soybean productivity and adaptability (Cober and Morrison 2010, Zhang *et al.* 2015a). Most of these traits have been studied through correlation with yield to improve the understanding of their relationship to yield components (Fox *et al.* 2015, Li *et al.* 2008a, Mansur *et al.* 1996). Moreover, to improve relevant agronomic traits in breeding programs, where large populations are evaluated every year, genotyping with a small number of markers would be more feasible (Schuster 2011). Consequently, it is desirable to identify molecular markers in genetically superior progenies or exotic plant introduction with favorable alleles, which should be successfully introgressed using marker assisted selection (MAS) (Fox *et al.* 2015).

Yield-quantitative trait loci (QTL) are often detected within the context of specific soybean breeding populations and environments, since some conditions in any given environment, geographic region or year can change the grain yield (Guzman *et al.* 2007, Orf *et al.* 1999). According to Palomeque *et al.* (2010), studies have identified QTLs associated with traits of interest that appear to be independent of the environment but dependent on the genetic background in which they found. The difficulty of identifying yield-QTL effective for MAS across a wide range of genetic and/or environmental contexts might be addressable by using preliminary yield trials to model target haplotypes within each context and then immediately selecting inbred lines that target genotypes in real time (Sebastian *et al.* 2010). Sebastian *et al.* (2010) demonstrated that using MAS with haplotypes

Communicated by Sachiko Isobe
Received March 14, 2017. Accepted May 23, 2017.
First Published Online in J-STAGE on November 16, 2017.
*Corresponding author (e-mail: contrerasudec@gmail.com)

to improve grain yield is possible if focused within a specific genetic and environmental context. In addition, the context-specific approach has already been adopted as a major component of MAS strategies known commercially as Accelerated Yield Technology (AYT) at Pioneer Hi-Bred International.

Genome-wide association studies (GWAS) using individual Single Nucleotide Polymorphism (SNPs) and haplotype information have been used to improve agronomical traits in soybean (Contreras-Soto *et al.* 2017, Hao *et al.* 2012, Zhang *et al.* 2015a). A haplotype block is a genomic region in which two or more polymorphic loci (i.e., SNP) in close proximity tend to be inherited together with high probability (Abdel-Shafy *et al.* 2014). These blocks are believed to be caused by recombination hotspots with extremely rare recombination within stretches of DNA, where the enclosed SNPs consequently segregate together from one generation to the next, acting as combined multi-site alleles (Greenspan and Geiger 2004). The combination of SNP alleles in a haplotype block on one chromosome covers the observed variation and can have higher linkage disequilibrium (LD) with the allele of a QTL than individual SNP alleles that are used to construct the haplotype (Abdel-Shafy *et al.* 2014). Furthermore, haplotype association is likely to be more powerful in the presence of LD (Garner and Slatkin 2003). Lorenz *et al.* (2010) used simulated phenotype data to show that the use of SNP-based haplotypes can increase power over the use of single-SNP markers in GWAS. Using haplotypes for QTL mapping could compensate for the bi-allelic limitation of SNPs, and substantially improve the efficiency of QTL mapping (Yang *et al.* 2011). According to Song *et al.* (2015), highly selfing species, such as soybean, are in many ways uniquely suitable for haplotype block mapping. Therefore, the aim of this study was to identify haplotypes across multiple environments that contribute to time to flowering and maturity in tropical soybean, with the view to improve the selection of desired alleles for these traits, but with minimal impact on yield.

Material and Methods

Plant material and field evaluation

The association panel of this study consisted of 141 cultivars of tropical soybean (**Supplemental Table 1**), which were field evaluated in five locations that represent eight environments of Brazil: Cascavel (24°52′54.9″S 53°32′30.4″W) in the growing seasons 2012/2013, 2013/2014 and 2014/2015 (Cas12/13, Cas13/14 and Cas14/15, respectively); Palotina (24°21′06.5″S 53°45′24.9″W) in the growing season 2014/2015 (Pal14/15); Primavera do Leste (15°34′37.6″S 54°20′41.8″W) in the growing season 2012/2013 (Pri12/13), Rio Verde (17°45′49.0″S 51°01′49.3″W) in the growing season 2013/2014 and 2014/2015 (Rio13/14 and Rio14/15), and Sorriso (12°32′43.6″S 55°42′41.8″W) in the growing season 2014/2015 (Sorr14/15). These locations were chosen on the basis of their diversity of latitude and altitude. Field trials were arranged in a complete block design with two replications. Fertilizer and field management practices recommended for optimum soybean production were used according to Embrapa (2011).

Phenotypic data analysis

Seed yield (SY), 100-seed weight (SW), plant height (PH), number of days to flowering (DTF) and maturity (DTM) were measured in the 141 soybean cultivars across the eight environments. Flowering dates were recorded when 50% of plants in a plot had open flowers. DTF was measured by counting days from emergence to flowering, when approximately 50% of plants per plot had at least one open flower (R1), and DTM was measured by counting the days from planting to the date when plants had 95% of their pods dry (R8 on the scale of Fehr and Caviness 1977). Field data were analyzed on the basis of the following mixed linear model:

$$y_{ijk} = \mu + g_i + l_j + (gl)_{ij} + b_{k(j)} + e_{ijk}$$

where μ is the total mean, g_i is the genetic effect of the i^{th} genotype, l_j is the effect of the j^{th} environment, $(gl)_{ij}$ is the interaction effect between the i^{th} genotype and the j^{th} environment (G × E), $b_{k(j)}$ is the random block effect within the j^{th} environment, and e_{ijk} is a random error following $N(0, \sigma_e^2)$. Adjusted entry means (AEM) were calculated for each of the 141 entries (i^{th} genotype: g_i) with the LSMEANS option of MIXED procedure, and these were used as a dependent variable in the posterior association analysis. AEM (denoted as M_i) was

$$M_i = \hat{\mu} + \hat{g}_i$$

where $\hat{\mu}$ and $\hat{g}_i$ are the generalized least-squares estimates of μ and g_i, respectively. To estimate AEM for all cultivars at each of the eight environments, g was regarded as fixed and b as random, as proposed by Stich *et al.* (2008). The Restricted Likelihood Ratio Test (RLRT) was calculated to confirm the heterogeneity of residual variance (across environments) using the GLIMMIX procedure in SAS, according to the following:

$$\mathrm{RLRT} = 2 \cdot \log\left[\frac{\mathrm{L(M_{HV})}}{\mathrm{L(M_{CV})}}\right]$$

where M_{HV} and M_{CV} are the models with heterogeneous and common (homogenous) variances, respectively. The asymptotic distribution of the RLRT statistic is Chi-square with p degrees of freedom ($\mathrm{RLRT} \sim \chi_p^2$), where p is the difference in the number of parameters included in the M_{HV} and M_{CV} models (in this case $P = 7$). Consequently, error variances were assumed to be heterogeneous among locations, and these were computed using the COVTEST homogeneity option, with RANDOM _residual_ statement and GROUP option in the GLIMMIX procedure (Mora *et al.* 2016). Analysis of *Deviance* (ANODEV) was conducted to evaluate the significance of the effects of the five traits across environments by using the MIXED procedure in SAS (Nelder

and Wedderburn 1972). The PROC CORR procedure was used to analyze Pearson correlations among variables by environment since G × E interactions were significant. Broad-sense heritability (h^2) for the five traits at each environment was estimated as the proportion of genetic variance (σ_g^2) over the total variance ($\sigma_g^2 + \sigma_e^2$), according to the formula:

$$h^2 = \frac{\sigma_g^2}{\left(\sigma_g^2 + \sigma_e^2\right)}$$

Association panel, SNP genotyping and population structure

Cultivars were genotyped for 6,000 single nucleotide polymorphisms (SNPs) using the Illumina BARCSoySNP6K BeadChip, corresponding to a subset of SNPs from the SoySNP50K BeadChip (Song *et al.* 2013). Genotyping was conducted by Deoxi Biotechnology Ltda. ® in Aracatuba, Sao Paulo, Brazil. A total of 3,780 SNP markers, including polymorphic and non-redundant SNPs, SNP markers with greater than 10% minor allele frequency (MAF) and missing data values lower than 25% were used for subsequent analysis, with heterozygous markers treated as missing data. Haplotype blocks were constructed using the Solid Spine method implemented in the software Haploview (Barrett *et al.* 2005), and have been previously reported by Contreras-Soto *et al.* (2017) (**Supplemental Table 2**). This method considers that the first and last markers in a block are in strong LD with all intermediate markers, thereby providing more robust block boundaries. A cutoff of 1% was used, meaning that if addition of a SNP to a block resulted in a recombinant allele at a frequency exceeding 1%, then that SNP was not included in the block. Then, these LD blocks were used to conduct the haplotype-based GWAS.

A Bayesian model-based method was used to infer population structure using 3,780 SNPs, implemented in the program InStruct (Gao *et al.* 2007). Posterior probabilities were estimated using five independent runs of the Markov Chain Monte Carlo (MCMC) sampling algorithm for the numbers of genetically differentiated groups (k) varying from 2 to 10, without prior population information. The MCMC chains were run for a burn-in of 5,000, followed by 50,000 iterations. The convergence of the log likelihood was determined by the value of the Gelman-Rubin statistic. The best estimate of k was determined according to the lowest value of the average log(Likelihood) and Deviance Information Criterion (DIC) values among the simulated groups (Gao *et al.* 2007), as defined by Spiegelhalter *et al.* (2002).

$$\mathrm{DIC} = \overline{\mathrm{D}} + \mathrm{pD}$$

where $\overline{\mathrm{D}}$ is a Bayesian measure of model fit that is defined as the posterior expectation of the deviance ($\overline{\mathrm{D}} = E_{\theta/y}[-2 \cdot \ln f(y/\theta)]$); pD is the effective number of parameters, which measures the complexity of the model.

SNP-based GWAS

AEM of each cultivar were used to perform SNP-based and haplotype-based GWAS for SY, PH, SW, DTF and DTM. To consider the effects of population structure and genetic relatedness among the cultivars, the following unified mixed-model (Cappa *et al.* 2013, Yu *et al.* 2006) of association was employed (in matrix form):

$$\mathbf{y} = \mathbf{S\alpha} + \mathbf{Qv} + \mathbf{Zu} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is a vector of adjusted phenotypic observations; $\boldsymbol{\alpha}$ is a vector of SNP effects (fixed); $\mathbf{v}$ is a vector of population structure effects (fixed); $\mathbf{u}$ is a vector of polygene background effects (random); and $\boldsymbol{\varepsilon}$ is a vector of residual effects. $\mathbf{S}$, $\mathbf{Q}$ and $\mathbf{Z}$ are incidence matrices for $\mathbf{a}$, $\mathbf{v}$, and $\mathbf{u}$, respectively. According to Yu *et al.* (2006), the variances of $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are $\mathrm{Var}(\mathbf{u}) = 2\mathbf{K}\sigma_g^2$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{R}\sigma_e^2$, respectively. $\mathbf{K}$ and $\mathbf{R}$ are the kinship and residual variance matrices, respectively. This is a structured association model (Q model), which considers the genetic structure of the association panel included in the association mixed model. The kinship coefficient matrix (K) that explains the most likely identity by state of each allele between cultivars was estimated using the program TASSEL (Bradbury *et al.* 2007, Endelman and Jannink 2012). Mixed linear models with Q and K by themselves and MLM considering Q + K models were also run in TASSEL (Bradbury *et al.* 2007, Yu *et al.* 2006). The Bayesian information criterion (BIC) (Schwarz 1978) was used for model selection, which is defined as:

$$\mathrm{BIC} = -2 \cdot \log L + p \cdot \log(n)$$

where L is the restricted maximum likelihood for a determined model, p the number of parameters to be estimated in the model, and n the sample size. BIC values were computed using the TASSEL program following Yu *et al.* (2006).

Haplotype-based GWAS

Haplotype-based GWAS was performed on the basis of LD information. Haplotype-based association mapping was performed by using adjusted phenotypes (y) as the dependent traits and the information of haplotype blocks in the model, as follows:

$$y = 1_n + \mathrm{H}_i + u + e_i$$

Where 1_n is a vector of n ones, with n representing the number of soybean cultivars, H_i is the incidence matrix of haplotype genotypes for the individuals at the i-th haplotype locus; The element of $\mathrm{H}_i(\mathrm{H}_{ij})$ is equal to the number of the i-th copies of haplotypes-blocks carried by the j-th cultivar. For this analysis u represents the polygenic gene effect or kinship matrix ($\mathbf{K}$) with variance $\mathrm{Var}(\mathbf{u}) = 2\mathbf{K}\sigma_g^2$ and the residual effects e_i with variance $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{R}\sigma_e^2$. A limit of detection (LOD) value higher than 3 was used as the threshold P-value for haplotype-trait associations according to Hwang *et al.* (2014). Then, only the significant haplotypes were used to estimate the phenotypic variance explained by haplotypes. The percentage of variation explained by the haplotype-based method was calculated using a simple regression performed in TASSEL as follows:

$$R^2_{LR} = 1 - \exp\left[\left(\frac{-2}{n}\right) * (\log LM - \log L0)\right]$$

Where LR is the Likelihood Ratio; *n* represents the number of observations (i.e., number of soybean cultivars); logLM and logL0 are the likelihood functions of the reduced and the intercept-only models, respectively (Sun *et al.* 2010). The Chi-square test was performed to check phenotypic differences among haplotype blocks using the CONTRAST option of the GENMOD procedure in SAS (SAS Institute, Inc., Cary, NC).

Additionally, the genomic regions or SNPs in haplotypes blocks identified in this study were compared to the genomic locations of QTLs previously reported for the traits under study. Genes, QTLs and markers annotated in Glyma1.01 and NCBI RefSeq gene models in SoyBase (www.soybase.org) were used as references.

Results

Phenotypic analysis, heritability and correlation between traits

Analysis of deviance indicated that the effects of genotype (G), environment (E) and their interaction (G × E) were statistically significant ($\chi^2 > 0.01$) for all traits under study (**Supplemental Table 3**). Highly significant differences were observed among traits and environments (**Supplemental Figs. 1–5**). On average, PH ranged from 38.27 cm (Rio13/14) to 103.45 cm (Cas12/13). SY and SW data ranged from 670.23 kg ha^{-1} (Rio13/14) to 3319.00 kg ha^{-1} (Cas14/15) and 11.96 g (Rio13/14) to 15.50 g (Rio14/15), respectively. As expected, DTF and DTM varied widely, ranging from 30 (Pri12/13) to 47 (Cas13/14) days, and 88 (Pri12/13) to 133 (Cas14/15) days, respectively (**Table 1**). The high phenotypic variability was confirmed by analysis of deviance, which revealed that all traits were severely influenced by environmental factors, showing significant G × E interaction (**Supplemental Table 3**). Over the eight environments, SY was moderately heritable with a value of 56%, whereas SW, DTM, PH and DTF showed high heritabilities: 81.7%, 91.7%, 93.4% and 94.6%, respectively.

Analysis of phenotypic correlation was conducted by environment since residual heterogeneity was observed among the environments and the G × E interaction was significant for all traits. In most of the environments, significant and positive phenotypic correlations were observed between SY and SW, with correlation coefficients ranging from 0.15 (Pri12/13; *P*-value < 10^{-2}) to 0.58 (Cas14/15; *P*-value < 10^{-4}), and with no correlation between SY and SW in Pal14/15 and Sorr14/15. SY and SW showed different patterns of phenotypic correlation with DTF and DTM across environments. The same was observed among SY and SW with PH. In most of the environments, PH and SW showed negative correlations, although non-significant at the 0.05 level. However, PH, DTF and DTM were low to highly positively correlated traits, and statistically different of zero (*P*-value < 10^{-4}), with correlation coefficients ranging from 0.13 for PH and DTF at Rio14/15 (*P*-value < 10^{-2}) to 0.84 for DTM and DTF at Rio13/14 (*P*-value < 10^{-4}) (**Supplemental Table 4**).

Genome-wide association across environments and traits

According to the deviance information criterion (from the posterior Bayesian clustering analysis), the most probable number of subpopulations was nine (**Supplemental Fig. 6**). The results based on Bayesian information criterion (BIC) consistently showed a better fit for the Q + K model over either Q or K alone (**Supplemental Table 5**). In total, 33, 29, 57, 72 and 40 linkage disequilibrium blocks were significantly associated with SY, SW, PH, DTM and DTF, respectively (**Tables 2–6**, **Supplemental Tables 6–10**). The haplotypes blocks explained considerable phenotypic variation: 17.6% to 96.8%, 13.6% to 33.2%, 45.2% to 99.4%, 12.7% to 59.9% and 12.9% to 42.7% for SY, SW, PH, DTM and DTF, respectively (**Tables 2–6**).

For SY, thirty-three haplotype blocks were effectively associated across environments. These haplotypes were identified on chromosomes 5, 9, 10, 11, 12, 15 and 19, and showed uncharacterized gene annotation or were located in intergenic regions (**Table 2**). The haplotype region Gm12_Hap12 encompasses a genomic region of 420 kb and contain the satt568 and satt442 markers, which are related to the seed protein 28-2 and 28-3 QTLs, respectively (Liang *et al.* 2010, Yang *et al.* 2011) (**Fig. 1**). Interestingly, QTLs related to reproductive stage and pod maturity were located near this haplotype. The SSR marker satt192, related to seed glycitein 9-7, was found within Gm12_Hap12. The SNPs located at this chromosomal location were confirmed as an exclusive haplotype region because they were associated with seed yield at Cas14/15, which correspond to Cascavel in the growing season 2014/15.

In Cascavel (Cas14/15), the haplotypes Gm12_Hap12a and Gm12_Hap12b were significantly different than haplotype Gm12_Hap12c. On average, Gm12_Hap12a and Gm12_Hap12b produced 3509.0 kg ha^{-1} and 3354.1 kg ha^{-1}, while haplotype Gm12_Hap12c yielded 2323.0 kg ha^{-1}, 34% and 31% lower than the haplotypes Gm12_Hap12a and Gm12_Hap12b, respectively. These haplotypes were well distributed in our association mapping panel (TAAT 32%, TAAC 45% and CGGT 23%) (**Table 2**).

For SW, twenty-nine haplotype blocks were significantly associated across environments and chromosomes (**Table 3**). Particularly, the Gm13_Hap41 was associated with one QTL related to SW, seed weight 40-1 (Rossi *et al.* 2013), and two QTLs for Pod maturity 20-1 and Lodging 27-6 (Li *et al.* 2008b, Rossi *et al.* 2013) (**Fig. 2**).

Most of the SNPs effectively associated with PH across environments were located on chromosome 19, including haplotype regions Gm19_Hap42 and Gm19_Hap43. These haplotypes were consistent across all environments. Gm19_Hap42 is a region containing the Determinate stem 1 gene (Dt1 or GmTFL1) (Cober *et al.* 2000), found 18.6 kb

Table 1. Descriptive statistics of phenotypic variation, heritability (h^2) across environments and variance components (G and G × E) of seed yield (SY), seed weight (SW), plant height (PH), days to maturity (DTM) and flowering (DTF) of 141 cultivars of soybean evaluated in eight environments

Trait	Environment	Mean	SD	Min	Max	G	G × E	h^2 (%)
SY (kg ha^{-1})	Cas12/13	2457.59	820.92	806.00	6563.00	75068	351055	56.7
	Pri12/13	1910.82	767.17	233.00	4372.00			
	Cas13/14	1863.71	623.59	125.00	5127.00			
	Rio13/14	670.23	305.47	128.00	1780.00			
	Cas14/15	3319.00	1297.25	176.00	7149.00			
	Pal14/15	1442.93	667.79	299.00	3669.00			
	Rio14/15	1559.18	814.61	136.00	4284.00			
	Sorr14/15	1775.69	800.04	152.00	4916.00			
	Mean							
SW (100seed gr)	Cas12/13	12.08	2.31	7.90	25.50	1.50	2.05	81.7
	Pri12/13	12.59	1.99	9.00	25.80			
	Cas13/14	13.49	2.26	7.90	25.00			
	Rio13/14	11.96	1.69	8.20	23.90			
	Cas14/15	12.33	3.15	6.30	19.40			
	Pal14/15	12.30	1.92	7.60	18.40			
	Rio14/15	15.50	1.93	10.30	21.00			
	Sorr14/15	14.78	1.96	10.10	21.80			
	Mean							
PH (cm)	Cas12/13	103.45	19.89	55.00	220.00	209.30	101.83	93.4
	Pri12/13	48.36	11.98	20.00	90.00			
	Cas13/14	97.59	19.54	45.00	205.00			
	Rio13/14	38.27	11.59	20.00	75.00			
	Cas14/15	90.34	24.45	30.00	180.00			
	Pal14/15	74.25	23.04	30.00	130.00			
	Rio14/15	46.17	13.72	20.00	95.00			
	Sorr14/15	55.52	18.50	23.00	100.00			
	Mean							
DTF (days)	Cas12/13	46.16	10.38	28.00	80.00	44.63	18.42	94.6
	Pri12/13	30.29	5.90	24.00	52.00			
	Cas13/14	47.75	9.41	29.00	82.00			
	Rio13/14	40.49	7.31	28.00	77.00			
	Cas14/15	46.58	10.85	26.00	76.00			
	Pal14/15	46.76	6.89	32.00	70.00			
	Rio14/15	37.39	7.26	24.00	54.00			
	Sorr14/15	31.42	6.09	25.00	46.00			
	Mean							
	Cas12/13	126.33	15.39	104.00	256.00	82.00	53.57	91.7
	Pri12/13	88.83	9.84	40.00	172.00			
	Cas13/14	124.89	15.66	97.00	248.00			
DTM (days)	Rio13/14	99.02	13.71	82.00	182.00			
	Cas14/15	133.89	10.59	106.00	164.00			
	Pal14/15	119.59	6.48	106.00	138.00			
	Rio14/15	104.98	9.38	80.00	123.00			
	Sorr14/15	98.27	5.56	75.00	123.00			
	Mean							

G × E = Genotype × Environment interaction.
G = Genotype.

upstream of the peak SNP ss715635425, which has been previously associated with PH and days to maturity in soybean (Contreras-Soto *et al.* 2017, Zhang *et al.* 2015b). In addition, other yield QTLs have previously been identified in this region, including seed yield 11-6, and plant height 13-8 and 4-2 (Lee *et al.* 1996, Specht *et al.* 2001) (**Table 4**, **Fig. 3**). Therefore, this QTL region should be considered as a relevant QTL responsible for PH.

For PH, interesting or discriminant haplotypes were located in our association mapping panel, i.e., the haplotype Gm19_Hap43c (GCG), which was associated in most of the environments and showed significant differences with the haplotype responsible for tallest plants (Gm19_Hap43d) in Cas13/14 and Cas14/15. On average, soybean plants with this haplotype showed heights of 93.5 and 84.4 cm of height in Cas13/14 and Cas14/15, respectively, and represented 72% of the total panel (**Table 4**). However, in Pal14/15, the haplotype Gm19_Hap43a (ATA) was significantly different than the haplotype responsible for smaller plants (Gm19_Hap43c), and consequently produced higher seed yield plants with significant differences among the others haplotypes (Gm19_Hap43a =1848.2 kg ha^{-1}, yielding 28% more

Table 2. Haplotype block associated with seed yield in 141 cultivars of tropical soybean

Env	Position (bp)			SN	Hap_ID	HapA	HF	SY[a]	R^2 (%)	Nearby genes or QTLs
	Chr	Start	End							
Cas13/14	9	38523430	38906660	3	Gm9_Hap22a	CCC	29	2126.2a	21.3	DNA-binding protein RHL1-like
					Gm9_Hap22b	CTC	3	1861.8ab		
					Gm9_Hap22c	TTC	61	1761.7b		
					Gm9_Hap22d	TTT	20	1717.9b		
Cas14/15	12	5622210	6052289	4	Gm12_Hap12a	TAAC	55	3509.0a	41.4	uncharacterized LOC102667945
					Gm12_Hap12b	TAAT	37	3354.1a		
					Gm12_Hap12c	CGGT	28	2323.0b		
Pal14/15	19	44965128	45370594	6	Gm19_Hap42a	AATxAA	34	1815.1a	96.8	Beta-fructofuranosidase insoluble isoenzyme 1-like
					Gm19_Hap42b	GCCGGG	88	1219.3a		
					Gm19_Hap42c	ACCGGG	2	374.2b		
					Gm19_Hap42d	AATGAA	–	–		
Pal14/15	10	3962673	4360182	6	Gm10_Hap8a	TATxTA	16	1999.6a	17.6	uncharacterized LOC100499780
					Gm10_Hap8b	CCGCTA	8	1522.3b		
					Gm10_Hap8c	CCGCCG	30	1516.2bc		
					Gm10_Hap8d	TCTxTA	34	1227.9bcd		
					Gm10_Hap8e	CCGCCA	22	1162.2bcd		
					Gm10_Hap8f	TCTCTA	–	–		
					Gm10_Hap8g	TATCTA	–	–		
					Gm10_Hap8h	TCGCTA	3	–		
Pal14/15	19	45478438	45643073	3	Gm19_Hap43a	ATA	31	1848.2a	50.6	Intergenic
					Gm19_Hap43b	ACG	2	1113.0b		
					Gm19_Hap43c	GCG	89	1112.9b		
					Gm19_Hap43d	GTA	2	–		
Pal14/15	11	4462645	4806173	5	Gm11_Hap11a	CCxAA	31	1699.9a	45.6	Probable 125 kDa kinesin-related protein-like
					Gm11_Hap11b	TATCA	6	1548.8ab		
					Gm11_Hap11c	CCTAC	21	1144.4bc		
					Gm11_Hap11d	TATAA	10	1060.8bc		
Sorr14/15	5	5621714	5794460	3	Gm5_Hap7a	CAC	18	2027.8a	18.7	uncharacterized LOC100818074
					Gm5_Hap7b	CGT	21	1956.5a		
					Gm5_Hap7c	TAT	78	1743.5a		
Sorr14/15	15	5621714	5794460	3	Gm15_Hap11a	TCC	7	2024.1a	30.1	uncharacterized LOC100785341
					Gm15_Hap11b	CCC	92	1898.9ab		
					Gm15_Hap11c	TTA	27	1510.8b		

Env: Environment; Chr: Chromosome; SN: Number of SNPs by haplotype; Hap_ID: Haplotype ID; HapA: Allelic haplotypes; HF: Haplotype frequency; SY: mean for seed yield (kg*ha^{-1}) of haplotypes at each environment.
[a] = Different letter means statistical differences.

than the mean of Pal14/15) (**Tables 2**, **4**).

The haplotype Gm19_Hap42a should differentiate indeterminate growth type in soybean cultivars, whereas Gm19_Hap42b should differentiate determinate soybean cultivars. Gm19_Hap42b showed significant differences with the haplotype responsible for the tallest plants (Gm19_Hap42a) at environments Cas12/13, Cas13/14, Cas14/15, Pal14/15, Rio14/15 and Sorr14/15 (**Table 4**). Interestingly, in Pal14/15 for SY, this haplotype was not significantly different from the plants that yielded more (**Table 2**).

For DTM, seventy-two haplotypes were associated across six environments. Of these, forty-two were located on intergenic regions and did not contain putative genes related to DTM. Specifically, some yield loci have previously been associated at the haplotype genomic region Gm20_Hap32: seed yield 12-3 and 15-15, plant height 14-1 and 26-15, and seed weight 36-5 (Han *et al.* 2012, Kabelka *et al.* 2004, Sun *et al.* 2006, Yuan *et al.* 2002) (**Table 5**, **Fig. 4**).

For DTF, forty haplotypes were associated across six environments. Most of these were located on intergenic regions of different chromosomes and showed no relationship with genes or markers. The haplotype Gm12_Hap12 was associated with DTF in the environments Cas12/13 and Cas13/14, and, interestingly, the same haplotype was associated with SY (**Tables 2**, **6**). Particularly, for DTF and SY, the haplotypes Gm12_Hap12a (TAAC) and Gm12_Hap12b (TAAT) showed significant differences with Gm12_Hap12c (54 and 55 days, respectively). In fact, these haplotypes showed the lowest days to flowering (precocity) (46 and 47 days, and 44 and 46 days in Cas12/13 and Cas13/14, respectively) and the highest yielding plants when compared with Gm12_Hap12c (**Table 6**).

Discussion

Phenotypic variation and correlation between traits

The heritability values observed in our panel indicate that much of the phenotypic variation was genetic. Heritability

Table 3. Haplotype block associated with 100-seed weight in 141 cultivars of tropical soybean

Env	Position (pb)			SN	Hap_ID	HapA	HF	SW[a]	R^2 (%)	Nearby genes or QTLs
	Chr	Start	End							
Pri12/13	11	4875880	4971452	2	Gm11_Hap12a	CT	19	14.0a	27.6	Probable Xaa-Pro aminopeptidase P-like
					Gm11_Hap12b	TC	61	12.5b		
					Gm11_Hap12c	CC	14	11.9b		
					Gm11_Hap12d	TT	22	11.4b		
Pri12/13	11	5074720	5248257	2	Gm11_Hap13a	AA	65	11.9a	13.6	Syntaxin-112 like
					Gm11_Hap13b	GA	18	12.1a		
Cas13/14	11	5074720	5248257	2	Gm11_Hap12a	GA	18	13.4a	15.7	Syntaxin-112 like
					Gm11_Hap12b	AA	65	12.9a		
Pal14/15	13	32225680	32347696	3	Gm13_Hap41a	GAC	51	13.0a	33.2	Glyma13g205300 Glyma13g207600 Glyma13g207900
					Gm13_Hap41b	GGT	31	12.5b		
					Gm13_Hap41c	AAT	28	11.1b		
Pal14/15	13	31956416	32154461	4	Gm13_Hap42a	GTAG	7	14.0a	22.3	Glyma13g209500
					Gm13_Hap42b	GCGG	23	12.7a		
					Gm13_Hap42c	GTGG	44	12.7a		
					Gm13_Hap42d	GTAA	6	12.6a		
					Gm13_Hap42e	ATAA	25	10.9ab		
					Gm13_Hap42f	ACAG	1	10.3ab		
Rio14/15	9	42458021	42790738	4	Gm9_Hap27a	TTTA	18	16.1a	20.2	Auxin-responsive protein IAA8-like
					Gm9_Hap27b	GCTA	76	15.8a		
					Gm9_Hap27c	GCCG	31	14.3b		
Sorr14/15	2	8544380	8819494	4	Gm2_Hap22a	AATG	4	16.3a	17.9	Auxilin-like protein 1-like
					Gm2_Hap22b	ACCA	24	15.9ab		
					Gm2_Hap22c	AATA	15	14.7bc		
					Gm2_Hap22d	GCTG	74	14.5bc		
Sorr14/15	11	5303401	5800217	4	Gm11_Hap14a	CATC	21	16.4a	23.1	Intergenic
					Gm11_Hap14b	TCTC	45	14.5b		
					Gm11_Hap14c	TATT	24	14.5b		
					Gm11_Hap14d	TCCC	12	14.2b		
					Gm11_Hap14e	TCTT	7	14.1b		

Env: Environment; Chr: Chromosome; SN: Number of SNPs by haplotype; Hap_ID: Haplotype ID; HapA: Allelic haplotypes; HF: Haplotype frequency; SW: mean for 100-seed weight (g/100seed) of haplotypes at each environment.
[a] = Different letter means statistical differences.

for SY (56%) was moderately high but smaller compared to Kim *et al.* 2012 (66%) and similar to Fox *et al.* 2015 (59%). On the other hand, the heritabilities for SW, PH, DTM and DTF were high and similar to those estimated by Hao *et al.* (2012) for SW and Zhang *et al.* (2015a) for PH, DTM and DTF. SY had a positive significant correlation with SW at six of the eight environments. Previous reports have also shown a significant positive correlation for SY and SW in soybean (Hao *et al.* 2012, Recker *et al.* 2014). For SY and PH, more positive than negative phenotypic correlations were observed. In addition, as suggested by Zhang *et al.* (2015a), the results based on multiple environments indicate that PH is a key factor for yield. On the other hand, most of the environments showed negative phenotypic correlations between SW and PH. However, Recker *et al.* (2014) showed a significant positive phenotypic correlation between these traits. At the moment, it is difficult to identify the potential relationship between these traits. Our results confirmed the inconsistent pattern of observed phenotypic correlation between seed yield and other important agronomic traits in soybean (Kim *et al.* 2012). The correlation among flowering-related traits with PH revealed the high phenotypic correlations between PH, DTM and DTF across multiple environments, suggesting close relationships among these traits.

Haplotype by environment interaction

The present study showed that some haplotype associations were location and year specific; however, stable haplotypes across environments was also found. According to Palomeque *et al.* (2010), QTLs for a specific trait are not always stable across environments and/or genetic backgrounds. The lack of validation in a different genetic background across environments could imply that these QTLs were not stable or that epistatic effects could be influencing the results. Another possibility is the presence of QTL by environment interactions, which represents a major challenge in genetic determinants of complex traits.

On the other hand, for plant height, strong and consistent genomic regions within haplotypes across environments were identified (i.e., Gm19_Hap42; Gm19_Hap43). For example, in Cascavel environments, the same haplotype region (Gm19_Hap42) was associated with plant height in the 2012/13, 2013/14 and 2014/15 growing seasons (Cas12/13,

Table 4. Haplotype block associated with plant height in 141 cultivars of tropical soybean

Env	Position (bp)			NS	Hap_ID	HapA	HF	PH [a]	R^2 (%)	Nearby genes or QTLs
	Chr	Start	End							
Cas12/13	19	45478438	45643073	3	Gm19_Hap43d	GTA	2	121.3a	52.0	Intergenic
					Gm19_Hap43a	ATA	31	115.4a		
					Gm19_Hap43c	GCG	89	98.9a		
					Gm19_Hap43b	ACG	2	90.0a		
Cas12/13	19	44965128	45370594	6	Gm19_Hap42a	AATxAA	31	114.7a	99.1	Sd yld 11-6 Pl ht 4-2 Pl ht 13-8 Dt1 gene (GmTFL1) Sat_286*
					Gm19_Hap42b	GCCGGG	88	99.2b		
					Gm19_Hap42c	ACCGGG	2	83.8b		
Cas12/13	13	36964799	37050736	3	Gm13_Hap53a	GGA	8	120.6a	23.9	uncharacterized LOC102670348
					Gm13_Hap53b	GGG	45	102.4ab		
					Gm13_Hap53c	AAA	53	100.9b		
					Gm13_Hap53d	GAA	4	91.3bc		
Pri12/13	19	45478438	45643073	3	Gm19_Hap43d	GTA	2	55.0a	63.4	Intergenic
					Gm19_Hap43a	ATA	31	53.9a		
					Gm19_Hap43c	GCG	89	46.7a		
					Gm19_Hap43b	ACG	2	43.8a		
Pri12/13	14	8027761	8527621	6	Gm14_Hap21a	CGGGTA	3	63.8a	46.1	Intergenic
					Gm14_Hap21b	CGGGGA	28	54.3a		
					Gm14_Hap21c	TTTAGA	16	49.5ab		
					Gm14_Hap21d	CGTATA	7	48.2ab		
					Gm14_Hap21e	TTTATA	39	47.9b		
					Gm14_Hap21f	TTTAGG	13	44.6b		
					Gm14_Hap21g	CGTAGA	1	–		
Cas13/14	19	45478438	45643073	3	Gm19_Hap43a	ATA	31	109.3a	51.6	Intergenic
					Gm19_Hap43d	GTA	2	106.3a		
					Gm19_Hap43c	GCG	89	93.5b		
					Gm19_Hap43b	ACG	2	83.8b		
Cas13/14	19	44965128	45370594	6	Gm19_Hap42a	AATxAA	31	108.8a	99.1	–*
					Gm19_Hap42b	GCCGGG	88	93.7b		
					Gm19_Hap42c	ACCGGG	2	75.0b		
Cas14/15	19	44965128	45370594	6	Gm19_Hap42a	AATxAA	31	103.8a	99.1	–*
					Gm19_Hap42b	GCCGGG	88	84.6b		
					Gm19_Hap42c	ACCGGG	2	63.8b		
Cas14/15	19	45478438	45643073	3	Gm19_Hap43d	GTA	2	111.3a	55.5	Intergenic
					Gm19_Hap43a	ATA	31	105.3a		
					Gm19_Hap43c	GCG	89	84.4b		
					Gm19_Hap43b	ACG	2	82.5b		
Pal14/15	19	44965128	45370594	6	Gm19_Hap42a	AATxAA	31	94.8a	99.4	–*
					Gm19_Hap42b	GCCGGG	88	63.2b		
					Gm19_Hap42c	ACCGGG	2	32.6b		
Pal14/15	19	45478438	45643073	3	Gm19_Hap43a	ATA	31	94.2a	79.0	Intergenic
					Gm19_Hap43d	ACG	2	69.0ab		
					Gm19_Hap43c	GCG	89	63.2b		
					Gm19_Hap43d	GTA	2	–		
Rio14/15	19	45478438	45643073	3	Gm19_Hap43d	GTA	2	66.3a	64.7	Intergenic
					Gm19_Hap43a	ATA	31	55.8b		
					Gm19_Hap43c	GCG	89	41.4b		
					Gm19_Hap43b	ACG	2	33.8bc		
Rio14/15	19	44965128	45370594	6	Gm19_Hap42a	AATxAA	31	54.7a	98.2	–*
					Gm19_Hap42c	ACCGGG	2	43.8ab		
					Gm19_Hap42b	GCCGGG	88	41.2b		
Sorr14/15	19	45478438	45643073	3	Gm19_Hap43d	GTA	2	68.9a	45.2	Intergenic
					Gm19_Hap43a	ATA	31	68.1a		
					Gm19_Hap43c	GCG	89	50.5ab		
					Gm19_Hap43b	ACG	2	44.5ab		
Sorr14/15	19	44965128	45370594	6	Gm19_Hap42a	AATxAA	31	67.1a	69.1	–*
					Gm19_Hap42c	ACCGGG	2	57.5ab		
					Gm19_Hap42b	GCCGGG	88	50.4b		

Env: Environment; Chr: Chromosome; SN: Number of SNPs by haplotype; Hap_ID: Haplotype ID; HapA: Allelic haplotypes; HF: Haplotype frequency; PH: mean for plant height (cm) of haplotypes at each environment.
[a] = Different letter means statistical differences.

Table 5. Haplotype block associated with days to maturity in 141 cultivars of tropical soybean

Env	Position (bp)			NS	Hap_ID	HapA	HF	DTM[a]	R^2 (%)	Nearby genes or QTLs
	Chr	Start	End							
Cas12/13	20	45458003	45857761	6	Gm20_Hap32a	GGGGGC	1	150.5a	59.9	LOC100789709 splicing factor U2AF-associated protein 2-like
					Gm20_Hap32b	GGGGAA	1	140.5b		
					Gm20_Hap32c	GAGGGA	8	132.9b		
					Gm20_Hap32d	GGGGGA	22	126.8b		
					Gm20_Hap32e	GAGGGC	67	126.2b		
					Gm20_Hap32f	AAAAAA	19	116.3bc		
Pri12/13	5	5621714	5794460	3	Gm5_Hap7a	CGT	21	98.7a	17.4	Intergenic
					Gm5_Hap7b	TAT	78	87.4b		
					Gm5_Hap7c	CAC	18	87.1b		
Pri12/13	9	6027763	6079042	2	Gm9_Hap13a	CA	14	94.7a	15.5	Intergenic
					Gm9_Hap13b	TA	53	87.6b		
					Gm9_Hap13c	CC	50	87.0b		
Pri12/13	9	43818290	44104810	3	Gm9_Hap30a	CCA	10	89.3a	18.2	Transcription initiation factor TFIID subunit 1-like
					Gm9_Hap30b	CTA	59	87.5a		
					Gm9_Hap30c	CCG	16	87.4a		
					Gm9_Hap30d	TTA	33	87.2a		
Pri12/13	1	49910518	50206347	5	Gm1_Hap17a	TxATA	15	95.5a	53.4	Intergenic
					Gm1_Hap17b	GGGGC	8	88.9ab		
					Gm1_Hap17c	TGGGC	83	87.4ab		
Pri12/13	4	45298627	45435298	2	Gm4_Hap25a	CC	47	89.2a	21.3	Intergenic
					Gm4_Hap25b	CT	16	89.1a		
					Gm4_Hap25c	TT	47	84.8b		
Pri12/13	11	7368580	7405714	2	Gm11_Hap18a	TA	37	89.3a	14.1	Intergenic
					Gm11_Hap18b	Cx	94	87.2a		
Pri12/13	5	2440984	2911445	6	Gm5_Hap2a	GGAAAA	3	91.3a	19.8	LOC100813996 transportin-3-like
					Gm5_Hap2b	GGGGAC	41	89.6a		
					Gm5_Hap2c	GGGAGC	11	87.8a		
					Gm5_Hap2d	TAAAAA	57	87.3a		
					Gm5_Hap2e	GAGAGC	1	–		
Pri12/13	2	831795	1033638	4	Gm2_Hap3a	TAAT	17	95.8a	19.2	ATG8i protein
					Gm2_Hap3b	CGAC	13	89.2a		
					Gm2_Hap3c	TGAT	31	88.2a		
					Gm2_Hap3d	CGCT	19	87.9a		
					Gm2_Hap3e	CGCC	19	86.9a		
					Gm2_Hap3f	CGAT	7	85.4a		
Pri12/13	9	43003730	43315338	3	Gm9_Hap28a	TTA	14	91.9a	18.2	uncharacterized LOC100793859
					Gm9_Hap28b	CCG	90	87.8ab		
					Gm9_Hap28c	CTA	3	87.6ab		
					Gm9_Hap28d	TTG	6	87.6ab		
					Gm9_Hap28e	TCG	11	84.6b		
Cas13/14	7	16625092	16979586	5	Gm7_Hap33a	ATGTT	22	127.7a	20.1	Intergenic
					Gm7_Hap33b	GCACT	45	123.7a		
					Gm7_Hap33c	GCACC	54	122.5a		
Cas13/14	9	41903227	42370093	7	Gm9_Hap26a	GxTTCTA	55	126.2a	31.1	Intergenic
					Gm9_Hap26b	AATTTTA	29	124.9ab		
					Gm9_Hap26c	GAxGCCC	11	120.3ab		
					Gm9_Hap26d	GAxGCTC	15	113.8b		
Cas14/15	2	40565506	40813466	4	Gm2_Hap48a	CAAT	14	141.1a	21.8	uncharacterized LOC100819417
					Gm2_Hap48b	CGGC	88	136.1b		
					Gm2_Hap48c	CGAT	5	127.9bc		
					Gm2_Hap48d	AAAT	13	121.8bc		
Pal14/15	2	13674975	14161558	3	Gm2_Hap33a	AAC	5	123.8a	21.8	Intergenic
					Gm2_Hap33b	ACC	10	120.6a		
					Gm2_Hap33c	ACA	12	119.9a		
					Gm2_Hap33d	GCA	55	119.9a		
					Gm2_Hap33e	GAC	1	119.0a		
					Gm2_Hap33f	AAA	22	116.3a		
Rio14/15	16	30267608	30519426	5	Gm16_Hap26a	GGGCG	111	106.6a	34.1	Intergenic
					Gm16_Hap26b	AATAA	18	97.7b		
Rio14/15	9	32388671	32695242	2	Gm9_Hap19a	AC	32	111.9a	12.7	Intergenic
					Gm9_Hap19b	GC	47	107.5b		
					Gm9_Hap19c	AT	35	99.1c		
Rio14/15	19	7322454	7358532	2	Gm19_Hap10a	GG	61	109.8a	23.9	Intergenic
					Gm19_Hap10b	AG	11	108.0a		
					Gm19_Hap10c	AA	48	98.9b		
Rio14/15	19	8115198	8436529	3	Gm19_Hap11a	GCT	10	109.8a	27.5	Intergenic
					Gm19_Hap11b	GCC	61	109.7a		
					Gm19_Hap11c	TTC	18	102.0b		
					Gm19_Hap11d	TTT	25	97.3b		
Rio14/15	4	47740685	48222393	2	Gm4_Hap31a	CA	2	109.5a	13.8	Intergenic
					Gm4_Hap31b	CG	109	107.6a		
					Gm4_Hap31c	TA	16	98.4b		

Env: Environment; Chr: Chromosome; SN: Number of SNPs by haplotype; Hap_ID: Haplotype ID; HapA: Allelic haplotypes; HF: Haplotype frequency; DTM: mean for days to maturity (days) of haplotypes at each environment.
[a] = Different letter means statistical differences.

Table 6. Haplotype block associated with days to flowering in 141 cultivars of tropical soybean

Env	Position (bp)			NS	Hap_ID	HapA	HF	DTF[a]	R^2 (%)	Nearby genes or QTLs
	Chr	Start	End							
Cas12/13	12	5622210	6052289	4	Gm12_Hap12c Gm12_Hap12a Gm12_Hap12b	CGGT TAAC TAAT	28 55 37	53.9a 45.6b 43.7b	34.6	uncharacterized LOC102667945*
Cas13/14	12	5622210	6052289	4	Gm12_Hap12c Gm12_Hap12a Gm12_Hap12b	CGGT TAAC TAAT	28 55 37	54.9a 46.7b 45.8b	41.9	–*
Cas13/14	17	8794927	9008173	4	Gm17_Hap10a Gm17_Hap10b	GCCG AATA	67 48	51.6a 42.1b	38.7	Intergenic
Cas13/14	15	49446994	49521249	2	Gm15_Hap45a Gm15_Hap45b Gm15_Hap45c Gm15_Hap45d	CC AC CT AT	61 5 2 61	52.8a 49.0ab 45.5ab 43.8b	18.5	LOC100804065 cysteine synthase-like
Rio13/14	12	14306367	14775930	5	Gm12_Hap21a Gm12_Hap21b	TTCAT CCTGG	40 79	43.2a 39.5b	42.7	Intergenic
Rio13/14	9	6155810	6470091	4	Gm9_Hap14a Gm9_Hap14b Gm9_Hap14c Gm9_Hap14d Gm9_Hap14e	GGCA GACA AATG AACA AATA	19 39 43 7 7	46.5a 40.0a 38.9a 38.0a 35.6b	26.5	Intergenic
Cas14/15	6	50711282	50936449	5	Gm6_Hap52a Gm6_Hap52b Gm6_Hap52c Gm6_Hap52d Gm6_Hap52e Gm6_Hap52f	TTGCG CTGCG TGGCG CGGCG TTGTA TTATA	8 20 10 39 6 26	56.6a 49.8ab 48.6ab 48.6ab 47.4ab 40.1b	26.1	Intergenic
Cas14/15	12	38680709	38970900	2	Gm12_Hap35a Gm12_Hap35b Gm12_Hap35c Gm12_Hap35d	AG AT CG CT	5 65 8 43	62.2a 48.4b 45.9bc 43.9c	12.9	LOC102660802 micronuclear linker histone polyprotein-like
Cas14/15	20	41883051	42297577	4	Gm20_Hap27a Gm20_Hap27b Gm20_Hap27c	GCGG ACGG ATTA	11 19 91	52.7a 50.2a 45.9a	32.9	Intergenic
Rio14/15	2	41787747	42088045	4	Gm2_Hap51a Gm2_Hap51b Gm2_Hap51c Gm2_Hap51d	GGCG GATA TGTA TATA	11 44 6 58	42.6a 41.6a 34.0b 33.9b	18.1	APO protein 3, mitochondrial-like**
Sorr14/15	2	41787747	42088045	4	Gm2_Hap51a Gm2_Hap51b Gm2_Hap51d Gm2_Hap51c	GGCG GATA TATA TGTA	11 44 58 6	36.2a 34.4a 28.8b 28.6b	19.8	–**

Env: Environment; Chr: Chromosome; SN: Number of SNPs by haplotype; Hap_ID: Haplotype ID; HapA: Allelic haplotypes; HF: Haplotype frequency; DTF: mean for days to flowering (days) of haplotypes at each environment.
[a] = Different letter means statistical differences.

Cas13/14 and Cas14/15), and explained most phenotypic variation (99.14%). Specifically, the haplotypes Gm19_Hap42a (AATxAA) and Gm19_Hap42b (GCCGGG) may help in marker-assisted selection of indeterminate and determinate growth habit soybean cultivars, respectively. QTLs controlling plant height are spread over all 20 chromosomes (Soybase 2016); however, this QTL region could be considered a relevant QTL responsible for PH (Contreras-Soto *et al.* 2017). In fact, Zhang *et al.* (2015b) previously reported this region as associated with PH and DTM in soybean. In soybean, stem growth habit is regulated by an epistatic interaction between two genes, Dt1 and Dt2 (Bernard 1972). The present study reported the haplotype association with the Dt1 gene (**Table 4**, **Fig. 3**), which maintains the indeterminate growth habit (dt1dt1 plants are fully determinate); however was not identified the Dt2 gene, which in the presence of Dt1, produces semideterminate plants. Additionally, our study reported a seed yield QTL in this region. As plant height is one of the major factors determining yield potential in soybean, Gm19_Hap42 (with its large effect on plant height) may also affect soybean yield substantially, as previously reported by Zhang *et al.* (2015b). In addition, Kato *et al.* (2015) suggest that the indeterminate growth habit is an advantageous characteristic in breeding for high yield of early maturing soybean varieties. The present study identified the Dt1 gene associated to plant height (indeterminate

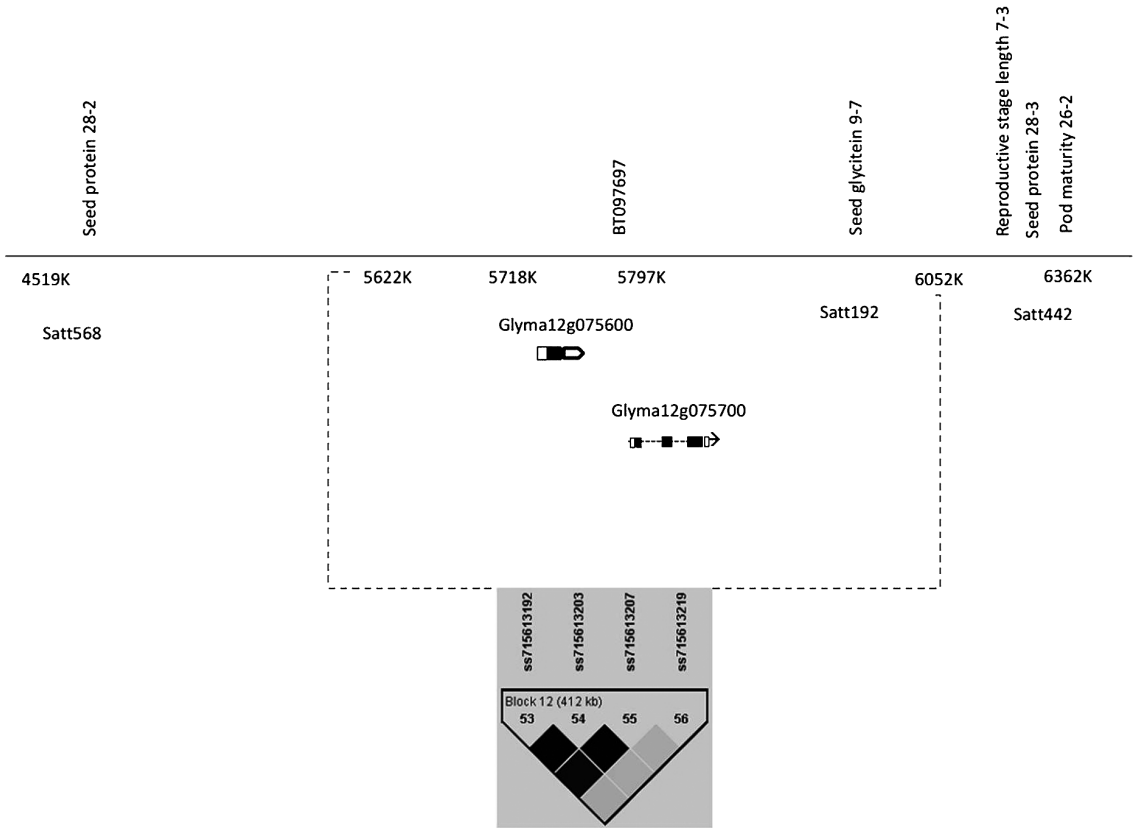

Fig. 1. Candidate region for major-effect loci: ss715613192 ss715613203, ss715613207 and ss715613219 on Gm12_Hap12 associated with SY, DTF and pod maturity in soybean. In the top panel, the QTLs and the proposed genomic region Glyma12g075600 annotated as a double-stranded RNA-binding protein 2-like, which encodes a Ribonuclease III protein (BT097697). Glyma12g075700 is another gene close to this haplotype region that encodes a senescence regulator protein in soybean. SSR markers related to seed protein and glycitein, pod maturity and reproductive stage. The bottom panel depicts a haplotype region of 412 kb associated with SY (intensity of black color indicates the r^2, and higher intensity means higher r^2).

plants) and co-associated to yield QTL, suggesting that this genomic region would also be responsible for high yield in soybean; however fine-mapping and cross-validation of the genes localized near of this haplotype should be performed.

Co-associated haplotype genomic regions among yield and flowering traits

For several traits, some molecular markers located at candidate genomic regions were co-localized on the same haplotype block. The co-association of a single gene or two linked genes to multiple traits that are phenotypically related has been previously reported (Sun *et al.* 2013). On Chromosome 19 (haplotype Gm19_Hap42), four QTL regions for plant height, seed yield, SCN (soybean cyst nematode) and terminal flower harbored three genes related to TERMINAL FLOWER 1 (TFL1), Basic leucine zipper (bZIP) transcription factor family protein and a beta-fructofuranosidase insoluble isoenzyme 1-like. TFL1 is an ortholog of the Antirrhinum CENTRORADIALIS (CEN) and acts as a floral repressor by preventing the expression of LFY and AP1 (Bradley *et al.* 1996, Liu *et al.* 2010). This gene corresponds to the *Dt1* locus, which controls soybean growth habit (Tian *et al.* 2010) and has been designated *GmTFL1* (Glyma19g37890). *GmTFL1* transcripts have been shown to accumulate in shoot apical meristems during early vegetative growth in both determinate and indeterminate growth habit soybeans; however, *GmTFL1* transcripts are abruptly lost after flowering in determinate lines while remaining in indeterminate ones (Liu *et al.* 2010). Consequently, this generates the difference of main stem nodes and flowering periods between indeterminate and determinate plants. Additionally, on the same haplotype region, the SSR Sat_286 has been identified and has exhibited a high accuracy in discrimination tests for growth habit in soybean (Vicente *et al.* 2016).

The LOC100789709 gene on chromosome 20 (Gm20_Hap32), described as a splicing factor U2AF-associated protein, was related to DTM in soybean. This gene is a homolog of atU2AF in *Arabidopsis thaliana*. Wang and Brendel (2006) demonstrated that altered expression levels of atU2AF^{35}a or atU2AF^{35}b causes pleiotropic phenotypes in flowering time, leaf morphology, flower, and silique shape in *A. thaliana*; specifically, pleiotropic phenotypes have been observed in mutants and transgenic lines. Homozygous

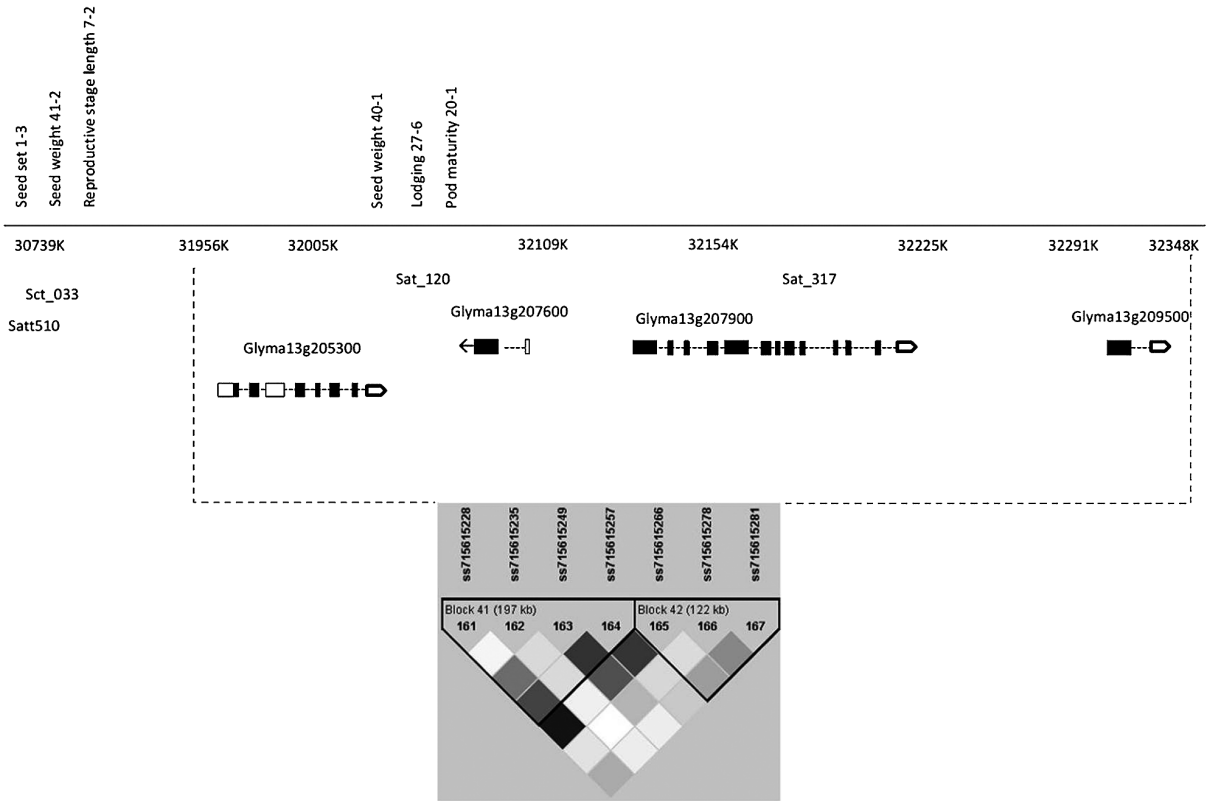

Fig. 2. Candidate region for major-effect loci: ss715615228, ss715615235, ss715615249 and ss715615257 located on haplotype Gm13_Hap41 and loci ss715615266, ss715615278 and ss715615281 located on haplotype Gm13_Hap42. Gm13_Hap41 was associated with SW, lodging and pod maturity in soybean. In the top panel, the QTLs and the proposed genomic regions are Glyma13g205300, Glyma13g207600 and Glyma13g207900, which encode an unknown protein, a nuclear transcription factor Y (subunit Gamma), and a dihydroxy-acid dehydratase, respectively. Additionally, Gm13_Hap42 were associated with SW and was identified as annotated gene Glyma13g209500, which encodes a 60S ribosomal protein. The bottom panel depicts haplotype regions of 197 kb and 122 kb associated with the aforementioned traits (intensity of black color indicates the r^2, and higher intensity means higher r^2).

atU2AF^{35}a T-DNA insertion plants and atU2AF^{35}b transgenic plants showed late flowering under both long and short day conditions. In fact, the altered expression of this gene may also affect days to flowering and maturity in soybean, confirming the haplotype association with this latter trait. Additionally, in this candidate region, some loci controlling grain yield have previously been associated: seed yield 12-3 and 15-15, plant height 14-1 and 26-15, and seed weight 36-5 (Han *et al.* 2012, Kabelka *et al.* 2004, Sun *et al.* 2006, Yuan *et al.* 2002). These results suggest that the morphological correlations between yield components and time to flowering and maturity traits are related on a genetic basis, suggesting gene pleiotropy and high rates of linkage disequilibrium (Chen and Lubberstedt 2010).

On chromosome 12 the haplotype Gm12_Hap12 was significantly associated with SY, DTF and DTM traits in all environments under study. This result may suggest that this region contains a single gene that has pleiotropic effects and is tightly linked with multiple genes. Recker *et al.* (2014) evaluated multiple environments to show that SY and DTM are positively correlated, while SY was not significantly correlated with DTF. In the present study, variable correlation results were obtained at individual environments, e.g., for SY and DTM: r = –0.65 (at Cas14/15-Cascavel) to r = 0.39 and 0.26 (at Pri12/13-Primavera do Leste and Sorr14/15-Sorriso, respectively); For SY and DTF: r = –0.44 (at Cas12/13-Cascavel) to r = 0.46 (at Rio14/15-Rio Verde). As such, these results should be interpreted at the environment level considering that these traits exhibit QTL-by-environment interactions. In Cascavel, the haplotype Gm12_Hap12 should be used to improve yield and precocity in the current soybean program. Specifically, the haplotypes Gm12_Hap12a and Gm12_Hap12b showed significant differences from Gm12_Hap12c for DTF and SY. In fact, these haplotypes showed the lowest days to flowering (precocity) (46 and 47 days, and 44 and 46 days, respectively) and the highest yield plants when compared with Gm12_Hap12c. Furthermore, the fine mapping of such regions could help to discern the specific genetic elements controlling these traits. For instance, in this genomic region, two annotated (candidate) genes were identified (Glyma12g075600 and Glyma12g075700), which, in fact, should be validated.

Finally, the results of this study suggest that the BARCSoySNP6K BeadChip and haplotype-based genome-

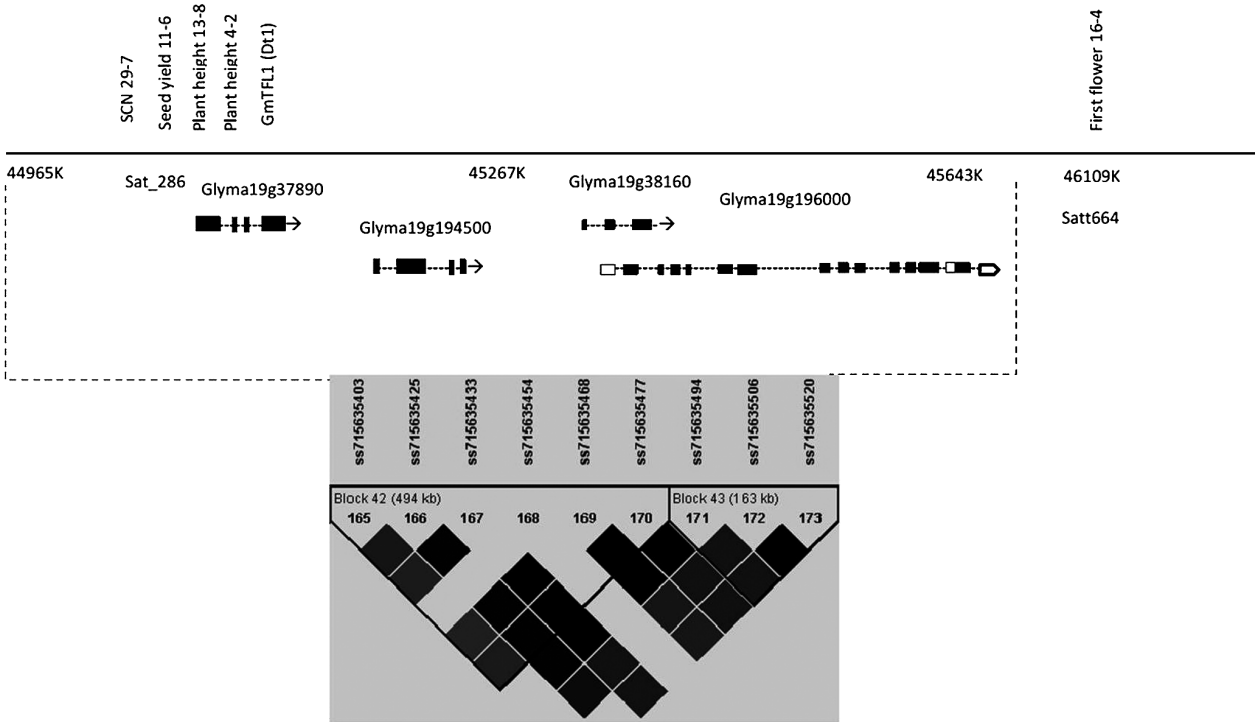

Fig. 3. Candidate region for major-effect loci: ss715635403, ss715635425, ss715635433, ss715635454 and ss715635468 located on Gm19_Hap42 and loci ss715635494, ss715635506 and ss715635520 located on Gm19_Hap43. Gm19_Hap42 was associated with PH, SY and SCN in soybean. In the top panel, the QTLs and the putative genomic region are Glyma19g37890 (Dt1 or GmFLT1), which determines stem growth habit in soybean, Glyma19g194500, which encodes an abscisic acid-insensitive protein, Glyma19g38160, which encodes a beta-fructofuranosidase isoenzyme and Glyma19g196000, which encodes a spindly related enzyme. The bottom panel depicts haplotype regions of 494 kb (Gm19_Hap42) and 163 kb (Gm19_Hap43) associated with the aforementioned traits (intensity of black color indicates the r^2, and higher intensity means higher r^2).

wide association are valuable sources of information for discovering genomic regions that control quantitative traits in soybean. This research identified useful associated markers that have not been previously reported and that were detected in multiple environments. This will facilitate assessing and validating causal genetic variation of complex quantitative traits and may eventually be used to accelerate the optimization of molecular breeding. However, as with any molecular markers, we emphasize that the identified haplotypes should be validated before large-scale use.

Acknowledgments

RICS thanks to Coordenação de aperfeiçoamento de pessoal de nível superior (CAPES) of Brazil and Programa de Pós-graduação em Genética e Melhoramento from the State University of Maringa for the research resources and scholarship.

Literature Cited

Abdel-Shafy, H., R.H. Bortfeldt, J. Tetens and G.A. Brockmann (2014) Single nucleotide polymorphism and haplotype effects associated with somatic cell score in German Holstein cattle. Genet. Sel. Evol. 46: 35.

Barrett, J.C., B. Fry, J. Maller and M.J. Daly (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.

Bernard, R.L. (1972) Two genes affecting stem termination in soybeans. Crop Sci. 12: 235–239.

Bradbury, P., Z. Zhang, D. Kroon, T. Casstevens, Y. Ramdoss and E. Buckler (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635.

Bradley, D., R. Carpenter, L. Copsey, C. Vincent, S. Rothstein and E. Coen (1996) Control of inflorescence architecture in *Antirrhinum*. Nature 379: 791–797.

Cappa, E.P., Y.A. El-Kassaby, M.N. Garcia, C. Acuña, N.M. Borralho, D. Grattapaglia and S.N. Marcucci Poltri (2013) Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in *Eucalyptus globulus*. PLoS ONE 8: e81267.

Chen, Y. and T. Lubberstedt (2010) Molecular basis of trait correlations. Trends Plant Sci. 15: 454–461.

Cober, E.R., J. Madill and H.D. Voldeng (2000) Early tall determinate soybean genotype E1E1e3e3e4e4dt1dt1 sets high bottom pods. Can. J. Plant Sci. 80: 527–531.

Cober, E.R. and M.J. Morrison (2010) Regulation of seed yield and agronomic characters by photoperiod sensitivity and growth habit genes in soybean. Theor. Appl. Genet. 120: 1005–1012.

Contreras-Soto, R.I., F. Mora, M.A.R. de Oliveira, W. Higashi, C.A. Scapim and I. Schuster (2017) A Genome-Wide Association Study for Agronomic Traits in Soybean Using SNP Markers and SNP-

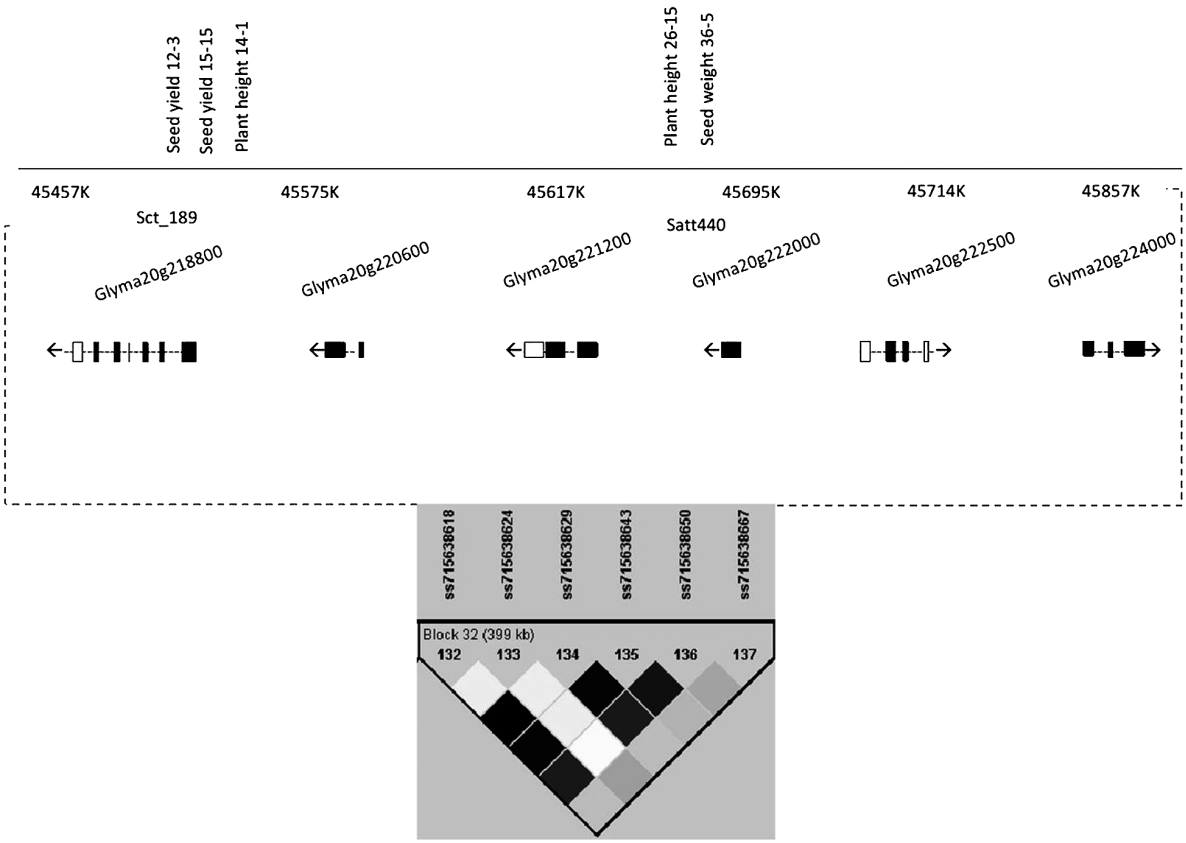

Fig. 4. Candidate region for major-effect loci: ss715638618, ss715638624, ss715638629, ss715638643, ss715638650 and ss715638667 are located on Gm20_Hap32 and associated with DTM. Additionally, QTLs for SY, SW and PH were identified. In the top panel, the QTLs and the proposed genomic regions include Glyma20g218800, Glyma20g220600, Glyma20g221200, Glyma20g222500, Glyma20g222000 and Glyma20g224000 which encode a splicing factor U2AF-associated protein, beta catenin-related armadillo repeat-containing, GDSL Esterase/Lipase, serine/threonine-protein phosphatase PP1 isozyme 2-related, At-hook motif nuclear-localized protein 19-related and Trihelix transcription factor GTL2, respectively. The bottom panel depicts a haplotype region of 399 kb associated with the aforementioned traits (intensity of black color indicates the r^2, and higher intensity means higher r^2).

Based Haplotype Analysis. PLoS ONE 12: e0171105.
Embrapa. (2011) Tecnologias de produção de soja – região central do Brasil 2012 e 2013. - Londrina: Embrapa Soja, p. 261 (Sistemas de Produção / Embrapa Soja, n.15).
Endelman, J.B. and J.-L. Jannink (2012) Shrinkage estimation of the realized relationship matrix. G3 (Bethesda) 2: 1405–1413.
Fehr, W.R. and C.E. Caviness (1977) Stages of soybean development. Spec. Rep. 80. Iowa State Univ, Ames.
Fox, C.M., T.R. Cary, R.L. Nelson and D.W. Diers (2015) Confirmation of a Seed Yield QTL in Soybean. Crop Sci. 55: 992–998.
Gao, H., S. Williamson and C.D. Bustamante (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176: 1635–1651.
Garner, C. and M. Slatkin (2003) On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci. Genet. Epidemiol. 24: 57–67.
Greenspan, G. and D. Geiger (2004) Model-based inference of haplotype block variation. J. Comput. Biol. 11: 493–504.
Guzman, P.S., B.W. Diers, D.J. Neece, S.K. St. Martin, A.R. LeRoy, C.R. Grau, T.J. Hughes and R.L. Nelson (2007) QTL associated with yield in three backcross-derived populations of soybean. Crop Sci. 47: 111–122.
Han, Y., D. Li, D. Zhu, H. Li, X. Li, W. Teng and W. Li (2012) QTL analysis of soybean seed weight across multi-genetic backgrounds and environments. Theor. Appl. Genet. 125: 671–683.
Hao, D., H. Cheng, Z. Yin, S. Cui, D. Zhang, H. Wang and D. Yu (2012) Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. Theor. Appl. Genet. 124: 447–458.
Hwang, E.Y., Q. Song, G. Jia, J.E. Specht, D.L. Hyten, J. Costa and P.B. Cregan (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genomics 15: 1.
Kabelka, E.A., B.W. Diers, W.R. Fehr, A.R. LeRoy, I.C. Baianu, T. You, D.J. Neece and R.L. Nelson (2004) Putative alleles for increased yield from soybean plant introductions. Crop Sci. 44: 784–791.
Kato, S., K. Fujii, S. Yumoto, M. Ishimoto, T. Shiraiwa, T. Sayama, A. Kikuchi and T. Nishio (2015) Seed yield and its components of indeterminate and determinate lines in recombinant inbred lines of soybean. Breed. Sci. 65: 154–160.
Kim, K.S., B.W. Diers, D.L. Hyten, M.A.R. Mian, J.G. Shannon and R.L. Nelson (2012) Identification of positive yield QTL alleles from exotic soybean germplasm in two backcross populations. Theor. Appl. Genet. 125: 1353–1369.
Lee, S.H., M.A. Bailey, M.A.R. Mian, T.E. Carter, D.A. Ashley, R.S.

Hussey, W.A. Parrott and H.R. Boerma (1996) Molecular markers associated with soybean plant height, lodging, and maturity across locations. Crop Sci. 36: 728–735.

Li, D., T.W. Pfeiffer and P.L. Cornelius (2008a) Soybean QTL for yield and yield components associated with *Glycine soja* alleles. Crop Sci. 48: 571–581.

Li, W., D.H. Zheng, K. Van and S.-H. Lee (2008b) QTL Mapping for major agronomic traits across two years in soybean (*Glycine max* L. Merr.). J. Crop Sci. Biotechnol. 11: 171–190.

Liang, H., Y. Yu, S. Wang, Y. Lian, T. Wang, Y. Wei, P. Gong, X. Liu, X. Fang and M. Zhang (2010) QTL Mapping of Isoflavone, Oil and Protein Contents in Soybean (*Glycine max* L. Merr.). Agric. Sci. China 9: 1108–1116.

Liu, B., S. Watanabe, T. Uchiyama, F. Kong, A. Kanazawa, Z. Xia, A. Nagamatsu, M. Arai, T. Yamada, K. Kitamura *et al.* (2010) The soybean stem growth habit gene *Dt1* is an ortholog of Arabidopsis *TERMINAL FLOWER1*. Plant Physiol. 153: 198–210.

Lorenz, A.J., M.T. Hamblin and J.-L. Jannink (2010) Performance of Single Nucleotide Polymorphisms versus Haplotypes for Genome-Wide Association Analysis in Barley. PLoS ONE 5: e14079.

Mansur, L.M., J.H. Orf, K. Chase, T. Jarvik, P.B. Cregan and K.G. Lark (1996) Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci. 36: 1327–1336.

Mora, F., Y.A. Quitral, I. Matus, J. Russell, R. Waugh and A. del Pozo (2016) SNP-based QTL mapping of 15 complex traits in barley under rain-fed and well-watered conditions by a mixed modeling approach. Front. Plant Sci. 7: 909.

Nelder, J.A. and R.W.M. Wedderburn (1972) Generalized Linear Models. J. Roy. Stat. Soc. 135: 370–384.

Orf, J.H., K. Chase, T. Jarvik, L.M. Mansur, P.B. Cregan, F.R. Adler and K.G. Lark (1999) Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. Crop Sci. 39: 1642–1651.

Palomeque, L., L.J. Liu, W.B. Li, B.R. Hedges, E.R. Cober, M.P. Smid, L. Lukens and I. Rajcan (2010) Validation of mega-environment universal and specific QTL associated with seed yield and agronomic traits in soybeans. Theor. Appl. Genet. 120: 997–1003.

Recker, J.R., J.W. Burton, A. Cardinal and L. Miranda (2014) Genetic and Phenotypic Correlations of Quantitative Traits in Two Long-Term, Randomly Mated Soybean Populations. Crop Sci. 54: 939–943.

Rossi, M.E., J.H. Orf, L.-J. Liu, Z. Dong and I. Rajcan (2013) Genetic basis of soybean adaptation to North American vs. Asian mega-environments in two independent populations from Canadian × Chinese crosses. Theor. Appl. Genet. 126: 1809–1823.

Schuster, I. (2011) Marker-assisted selection for quantitative traits. Crop Breed. Appl. Biotechnol. S1: 50–55.

Schwarz, G. (1978) Estimating the dimension of a model. Ann. Stat. 6: 461–464.

Sebastian, S.A., L.G. Streit, P.A. Stephens, J.A. Thompson, B.R. Hedges, M.A. Fabrizius, J.F. Soper, D.H. Schmidt, R.L. Kallem, M.A. Hinds *et al.* (2010) Context-specific marker-assisted selection for improved grain yield in elite soybean populations. Crop Sci. 50: 1196–1206.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson and P.B. Cregan (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS ONE 8: e54985.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson and P.B. Cregan (2015) Fingerprinting soybean germplasm and its utility in genomic research. G3 (Bethesda) 5: 1999–2006.

SoyBase (2016) USDA-ARS Soybean Genetics and Genomics Database. USDA, Washington, DC. www.soybase.org/search/qtllist.php

Specht, J.E., K. Chase, M. Macrander, G.L. Graef, J. Chung, J.P. Markwell, M. Germann, J.H. Orf and K.G. Lark (2001) Soybean response to water: a QTL analysis of drought tolerance. Crop Sci. 41: 493–509.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin and A. van der Linde (2002) Bayesian measures of model complexity and fit. J. R. Stat. Soc. Series B Stat. Methodol. 64: 583–639.

Stich, B., J. Mohring, H.P. Piepho, M. Heckenberger, E.S. Buckler and A.E. Melchinger (2008) Comparison of mixed-model approaches for association mapping. Genetics 178: 1745–1754.

Sun, D., W. Li, Z. Zhang, Q. Chen, H. Ning, L. Qiu and G. Sun (2006) Quantitative trait loci analysis for the developmental behavior of soybean (*Glycine max* L. Merr). Theor. Appl. Genet. 112: 665–673.

Sun, G., C. Zhu, M.H. Kramer, S.S. Yang, W. Song, H.P. Piepho and J. Yu (2010) Variation explained in mixed-model association mapping. Heredity (Edinb) 105: 333–340.

Sun, S., M.Y. Kim, K. Van, Y.W. Lee, B. Li and S.H. Lee (2013) QTLs for resistance to Phomopsis seed decay are associated with days to maturity in soybean (*Glycine max*). Theor. Appl. Genet. 126: 2029–2038.

Tian, Z., X. Wang, R. Lee, Y. Li, J.E. Specht, R.L. Nelson, P.E. McClean, L. Qiu and J. Ma (2010) Artificial selection for determinate growth habit in soybean. Proc. Natl. Acad. Sci. USA 107: 8563–8568.

Vicente, D., I. Schuster, F. Lazzari, J.P.D. Paranzini, M.A.R. de Oliveira and C.E.C. Prete (2016) Mapping and validation of molecular markers of genes *Dt1* and *Dt2* to determine the type of stem growth in soybean. Acta Sci. Agron. 38: 61–68.

Wang, B.B. and V. Brendel (2006) Molecular characterization and phylogeny of U2AF35 homologs in plants. Plant Physiol. 140: 624–636.

Yang, K., J. Moon, N. Jeong, H. Chun, S. Kang, K. Back and S. Jeong (2011) Novel major quantitative trait loci regulating the content of isoflavone in soybean seeds. Genes Genomics 33: 685–692.

Yu, J., G. Pressoir, W. Briggs, B.I. Vroh, M. Yamasaki, J. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

Yuan, J., V.N. Njiti, K. Meksem, M.J. Iqbal, K. Triwitayakorn, M.A. Kassem, G.T. Davis, M.E. Schmidt and D.A. Lightfoot (2002) Quantitative trait loci in two soybean recombinant inbred line populations segregating for yield and disease resistance. Crop Sci. 42: 271–277.

Zhang, H., D. Hao, H.M. Sitoe, Z. Yin, Z. Hu, G. Zhang and D. Yu (2015a) Genetic dissection of the relationship between plant architecture and yield component traits in soybean (*Glycine max*) by association analysis across multiple environments. Plant Breed. 134: 564–572.

Zhang, J., Q. Song, P.B. Cregan, R.L. Nelson, X. Wang, J. Wu and G.-L. Jiang (2015b) Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. BMC Genomics 16: 217.