**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Desequilíbrio de ligação, análise de associação genômica ampla e sinais de
seleção em soja**

Maisa Curtolo

Tese apresentada para obtenção do título de Doutora em
Ciências. Área de concentração: Genética e
Melhoramento de Plantas

**Piracicaba
2018**

Maisa Curtolo
Bacharel em Biotecnologia

**Desequilíbrio de ligação, análise de associação genômica ampla e sinais de seleção
em soja**

Orientador:
Prof. Dr. **JOSÉ BALDIN PINHEIRO**

Tese apresentada para obtenção do título de Doutora em
Ciências. Área de concentração: Genética e
Melhoramento de Plantas

Piracicaba
2018

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Curtolo, Maisa

Desequilíbrio de ligação, análise de associação genômica ampla e sinais de seleção em soja / Maisa Curtolo. - - Piracicaba, 2018.
103 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. *Glycine max* 2. Desequilíbrio de ligação 3. Diversidade genética 4. Estrutura populacional 5. GWAS 6. Sinais de seleção I. Título

À minha família.

Dedico.

AGRADECIMENTOS

À Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ/USP) e ao programa de pós-graduação em Genética e Melhoramento de Plantas do Departamento de Genética pela oportunidade de realizar o curso de Doutorado;

A todos os professores do Departamento de Genética da ESALQ - USP, pelos ensinamentos e conhecimentos compartilhados;

A todos os funcionários do Departamento de Genética, em especial ao Domingos de Sálvio Amaral, Márcio Araújo Silva, Berdan e Léia pela competência e compromisso com o trabalho;

Ao Prof. Dr. Luís Eduardo Aranha Camargo pela oportunidade da realização do estágio de graduação e pelos ensinamentos que tive durante os anos de convivência;

Ao Prof. Dr. José Baldin Pinheiro pela orientação, conversas, ensinamentos, pelo exemplo profissional e amizade;

À Prof^a. Dr^a. Maria Imaculada Zucchi pelo apoio, amizade e por sempre dizer:” calma, var dar certo”;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão da bolsa de estudo;

Ao Prof. Dr. David Hyten por me receber no período de doutorado sanduíche na Universidade de Nebraska em Lincoln;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) a concessão da bolsa do PDSE para a realização do doutorado sanduíche;

À minha família, que sempre me incentiva a buscar novos desafios;

Ao meu companheiro Rafael, que mesmo a distância, esteve sempre me apoiando em todos os momentos. Obrigada por todo o amor e companheirismo; Agradeço também pelas suas contribuições que foram fundamentais para a finalização dessa tese;

Aos colegas do Laboratório de Diversidade Genética e Melhoramento de Plantas pela convivência, cafés e conhecimentos compartilhados. Principalmente a Vanessa, a Diane, Alessandro, Maiara, Emanuel e ao João, meus grandes amigos que nunca mediram esforços para me ajudar;

Aos colegas do Laboratório de Genética Molecular, Marina, Rafael, Ferchu, Jerônimo, Thais, Débora, Rafinha e Silvania, pelos ensinamentos, conversas e amizade;

Aos meus amigos Gabriel, Tatiane, Sueme, Santhiago, Adriana, Júlia, Mariana e Thais por tornar meus dias mais felizes;

Enfim, a todos aqueles que de forma direta ou indiretamente tenham contribuído para a realização deste trabalho.

Muito obrigada!

SUMÁRIO

RESUMO.....	7
ABSTRACT	8
1. INTRODUÇÃO	9
REFERÊNCIAS	10
2. EFEITOS DA ESTRUTURA POPULACIONAL E PARENTESCO NA MENSURAÇÃO DO DESEQUILÍBRIO DE LIGAÇÃO EM SOJA	13
RESUMO.....	13
ABSTRACT	13
2.1. INTRODUÇÃO	14
2.2. MATERIAL E MÉTODOS	17
2.2.1. <i>Material vegetal e genotipagem</i>	17
2.2.2. <i>Filtragem e imputação de dados perdidos</i>	17
2.2.3. <i>Estrutura populacional e matriz de parentesco</i>	18
2.2.4. <i>Desequilíbrio de ligação</i>	18
2.3. RESULTADOS E DISCUSSÃO	19
2.3.1. <i>Genotipagem</i>	19
2.3.2. <i>Estrutura da população e matriz de parentesco</i>	20
2.3.3. <i>Desequilíbrio de ligação</i>	22
2.3.4. <i>Desequilíbrio de ligação inter-cromossômico</i>	27
2.4. CONCLUSÃO	28
REFERÊNCIAS	28
FIGURAS	34
TABELAS	40
3. ARQUITETURA GENÉTICA DE MÚLTIPLOS CARACTERES DE SOJA VIA ANÁLISE DE ASSOCIAÇÃO GENÔMICA AMPLA.....	43
RESUMO.....	43
ABSTRACT	43
3.1. INTRODUÇÃO	44
3.2. MATERIAL E MÉTODOS	45
3.2.1. <i>Material vegetal</i>	45
3.2.2. <i>Genotipagem e controle de qualidade dos marcadores</i>	46
3.2.3. <i>Fenotipagem e análises estatísticas</i>	47
3.2.4. <i>Análise de associação genômica-ampla (GWAS)</i>	47
3.3. RESULTADOS E DISCUSSÃO	48
3.3.1. <i>Análise fenotípica</i>	48
3.3.2. <i>Estrutura da população e matriz de parentesco</i>	50
3.3.3. <i>Mapeamento associativo</i>	51
3.4. Conclusão	57
REFERÊNCIAS	57
FIGURAS	63
TABELAS	64
4. IDENTIFICAÇÃO DE REGIÕES DO GENOMA DIFERENTEMENTE AFETADAS PELA SELEÇÃO EM POPULAÇÕES DE SOJA VISANDO O AUMENTO DA BASE GENÉTICA DA CULTURA EM PROGRAMAS DE MELHORAMENTO DO BRASIL	69
RESUMO.....	69
ABSTRACT	69

4.1.	INTRODUÇÃO	70
4.2.	MATERIAL E MÉTODOS.....	72
4.2.1.	<i>Material vegetal e genotipagem</i>	72
4.2.2.	<i>Diversidade genética</i>	72
4.2.3.	<i>Screening de indicativos de sinais de seleção</i>	73
4.2.4.	<i>Estrutura de blocos de haplótipos</i>	73
4.3.	RESULTADOS E DISCUSSÃO	73
4.3.1.	<i>Diversidade genética</i>	73
4.3.2.	<i>Diferenciação entre as populações</i>	75
4.3.3.	<i>Redução da diversidade nucleotídica</i>	78
4.3.4.	<i>Estrutura de blocos de haplótipos</i>	79
4.3.5.	<i>Regiões candidatas</i>	81
4.4.	<i>Conclusão</i>	85
	REFERÊNCIAS	86
	FIGURAS.....	93
	TABELAS	96

RESUMO

Desequilíbrio de ligação, análise de associação genômica ampla e sinais de seleção em soja

Com a disponibilidade de tecnologias de genotipagem robustas, fornecendo milhares de marcadores moleculares a um baixo custo por amostra, tornou-se mais acessível escanear o genoma todo. Logo, as abordagens de mapeamento associativo, baseado no cálculo do desequilíbrio de ligação, e a busca de regiões do genoma que apresentam sinais de seleção foram favorecidas. Essas ferramentas apresentam um grande potencial a ser explorado em programas de melhoramento de soja, auxiliando na obtenção de informações valiosas em relação à arquitetura genética dos caracteres de interesse agrônomico e à dinâmica dos processos de seleção. Portanto, os objetivos deste estudo foram: (i) verificar os efeitos da relação genética entre genótipos no desequilíbrio de ligação; (ii) obter informações em relação a arquitetura genética para dez caracteres por meio da abordagem de mapeamento associativo; e (iii) identificar regiões sob seleção diferencial em uma população representada por cultivares brasileiras comparadas a de uma população de genótipos exóticos de diferentes origens geográficas. Para isto, 95 genótipos de soja foram genotipados utilizando a plataforma da *Affymetrix (180 K Axiom® Soybean Genotyping Array)*. O efeito da estrutura populacional sobre DL foi investigado utilizando como correções as matrizes resultantes do software STRUCTURE, DAPC, PCA e a matriz de parentesco. Para o mapeamento associativo, os genótipos foram fenotipados para dez caracteres. Foram realizadas várias abordagens para detecção de sinais de seleção: estimação da diferenciação populacional nas regiões do genoma por meio do índice de fixação (F_{ST}) e *cross-population composite likelihood ratio test* (XP-CLR); identificação de regiões do genoma com redução de diversidade nucleotídica (μ); e a presença de blocos de haplótipos. Ao utilizar medidas de correções para estimação do DL no genoma da soja observamos a influência da estrutura da população e do parentesco nos padrões de DL. Pelo mapeamento associativo, foram identificados 181 marcadores associados para dez caracteres avaliados em soja. Além disso, foi verificada a complexa interação entre regiões envolvidas no controle de caracteres quantitativos (QTL) e ambientes. Regiões diferencialmente selecionadas foram identificadas entre a população de genótipos brasileiros e de materiais de origens diversas, demonstrando que essas passaram por processos de seleção divergente.

Palavras-chave: *Glycine max*; Diversidade genética; QTLs; Estrutura populacional; GWAS; Seleção artificial

ABSTRACT

Linkage disequilibrium, genome-wide association analysis and signals of selection on soybean

The availability of robust genotyping technologies providing thousands of markers with a low-cost per sample, whole-genome scans are becoming more accessible. Hence, associative mapping approaches, based on the calculation of linkage disequilibrium (LD), and the detection of signals of selection were favored. These tools have a great potential to be explored in soybean breeding programs, helping to obtain valuable information about the genetic architecture of the characters of agronomic interest and of the dynamics of selection processes. Therefore, the objectives of this study were: (i) to verify the effects of the genetic relationship among genotypes on soybean linkage disequilibrium; (ii) to obtain information about the genetic architecture in ten soybean traits with an association mapping approach; and (iii) to identify regions under differential selection between a population represented by Brazilian cultivars compared to a population of exotic genotypes from different geographic origins. For this, 95 soybean genotypes were genotyped using an Affymetrix (180 K Axiom® Soybean Genotyping Array) platform. The effect of the population structure on LD was investigated using as corrections the matrices resulting from the software STRUCTURE, DAPC, PCA and a kinship matrix. We phenotyped ten soybean traits in order to carry the association mapping. Several approaches were carried aiming to detect selection signals: estimation of population genetic differentiation in the genomic regions with fixation index (F_{ST}) and *cross-population composite likelihood ratio test* (XP-CLR); identification of genomic regions with reduced nucleotide diversity (μ); and the presence of haplotype blocks. Using the LD corrections, it was showed the influence of population structure and kinship on LD patterns in soybean. By the association mapping, we identified 181 markers associated with the ten traits evaluated in soybean. In addition, complex interactions between QTL and environments were verified. Differentially selected regions were identified between the Brazilian genotypes population and the population of materials from diverse origins, demonstrating that they underwent divergent selection processes.

Keywords: *Glycine max*; Genetic diversity; QTLs; Population structure; GWAS; Artificial selection

1. INTRODUÇÃO

A soja (*Glycine max* L.) é uma das culturas mais importantes do mundo, contribuindo com mais da metade da produção mundial de óleo e proteína [1]. Devido a sua importância, esforços no que tange ao desenvolvimento de cultivares de soja com potenciais produtivos maiores e, principalmente, melhor adaptados às condições ambientais adversas, torna-se imprescindível para o desenvolvimento desta cultura. Assim, o entendimento do controle genético dos caracteres agronômicos da soja tem tido grande destaque.

Concomitantemente aos avanços das tecnologias de genotipagem de marcadores SNPs (*Single Nucleotide Polymorphisms*) em grande escala e sequenciamento completo do genoma, tornou-se possível escanear o genoma todo. Com a disponibilidade dessas ferramentas, a busca de regiões envolvidas no controle de caracteres fenotípicos é favorecida, devido à cobertura de regiões do genoma antes não acessadas. Assim, a abordagem de mapeamento associativo é considerada uma ferramenta com grande potencial a ser explorada para o entendimento da arquitetura genética dos caracteres selecionados em programas de melhoramento. Assim como o mapeamento de ligação, o mapeamento associativo tem como objetivo a identificação de regiões do genoma envolvidas no controle de caracteres quantitativos [2]. Contudo, esse explora o histórico dos eventos de recombinação ocorridos em uma população, não necessitando o desenvolvimento de populações de mapeamento, como bi-parentais, retro-cruzamentos e duplo-haplóides, configurando-se como uma poderosa ferramenta alternativa à abordagem convencional [3].

O mapeamento associativo baseia-se no cálculo do desequilíbrio de ligação (DL) entre locos marcadores [4]. Entretanto, estudos dessa natureza, bem como para obter informações a respeito do DL de uma população, requerem conhecimento prévio acerca da diversidade e estruturação da população a ser utilizada, a fim de minimizar ocorrências de falso-positivos [5,6]. A diferenciação populacional e o parentesco elevado entre os indivíduos podem resultar na subdivisão de populações, e consequentemente gerarem um DL superestimado [7]. A superestimação do DL afeta a mensuração da quantidade e densidade de marcadores necessários para estudos de mapeamento associativo, podendo comprometer a resolução do mapeamento [8]. Adicionalmente, pode fornecer informações não acuradas a respeito da taxa de decaimento do DL de uma espécie [7].

Além dos métodos de mapeamento, a identificação de sinais de seleção no genoma também se constitui como ferramenta potencial na identificação de regiões de importância agronômica [9]. O reconhecimento dessas regiões pode contribuir na identificação de segmentos responsáveis por grandes efeitos fenotípicos, que por sua vez podem ser evidenciados pela co-localização de QTLs identificados em estudos anteriores via abordagens de mapeamento [10]. Os estudos dos sinais de seleção deixados no genoma, não só é interessante por auxiliar na identificação das regiões envolvidas no controle de caracteres quantitativos, mas também por fornecer informações a respeito da dinâmica das regiões das envolvidas nos processos de seleção [13]. Assim, esses estudos são de grande importância para a definição das estratégias dos programas de melhoramento e, portanto, para o desenvolvimento de novos cultivares [10].

A abordagem de detecção de sinais de seleção apresenta a vantagem de não necessitar de grandes populações de mapeamento e da coleta de dados fenotípicos [14]. Além do mais, essa permite utilizar a comparação de populações divergentes, com históricos de seleção distintos, revelando as regiões diferencialmente mais afetadas nas populações. Além disso, permite acessar informações específicas de uma população, revelando as alterações no genoma decorrentes de processos de domesticação e do melhoramento genético [10].

Desta forma, os objetivos foram: (i) verificar os efeitos da relação genética entre genótipos no desequilíbrio de ligação; (ii) obter informações em relação a arquitetura genética para dez caracteres via abordagem de mapeamento associativo; e (iii) identificar regiões sob seleção diferencial em uma população representada por cultivares brasileiras comparadas a de uma população de genótipos exóticos de diferentes origens geográficas (*Plant Introductions*, PI).

REFERÊNCIAS

1. Yan L, Hofmann N, Li S, Ferreira ME, Song B, Jiang G, et al. Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC Genomics*. *BMC Genomics*; 2017;18: 1–11. doi:10.1186/s12864-017-3922-0
2. Al-Maskri AY, Sajjad M, Khan SH. Association mapping: A step forward to discovering new alleles for crop improvement. *Int J Agric Biol*. 2012;14: 153–160. doi:10.13140/2.1.1925.9524
3. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*. *Plant Methods*; 2013;9: 29. doi:10.1186/1746-4811-9-29

4. Gupta PK, Rustgi S, Kulwal PL. Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol.* 2005;57: 461–85. doi:10.1007/s11103-005-0257-z
5. Mackay I, Powell W. Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci.* 2007;12: 57–63. doi:10.1016/j.tplants.2006.12.001
6. Mezouk S, Dubreuil P, Bosio M, Décousset L, Charcosset A, Praud S, et al. Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. *Theor Appl Genet.* 2011;122: 1149–1160. doi:10.1007/s00122-010-1519-y
7. Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb).* Nature Publishing Group; 2012;108: 285–291. doi:10.1038/hdy.2011.73
8. Mangin B, Sandron F, Henry K, Devaux B, Willems G, Devaux P, et al. Breeding patterns and cultivated beets origins by genetic diversity and linkage disequilibrium analyses. *Theor Appl Genet.* Springer Berlin Heidelberg; 2015;128: 2255–2271. doi:10.1007/s00122-015-2582-1
9. Wen Z, Boyse JF, Song Q, Cregan PB, Wang D. Genomic consequences of selection and genome-wide association mapping in soybean. *BMC Genomics.* BMC Genomics; 2015;16: 671. doi:10.1186/s12864-015-1872-y
10. Tomicic M, Djordjevic V, Obrecht D, Miladinovic J, Brbaklic L, Mikic A, et al. Tracking footprints of selection associated with soybean adaptation to Central-East Europe environments. *Euphytica.* Springer Netherlands; 2015;203: 701–713. doi:10.1007/s10681-014-1313-z
11. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. *Genome Res.* 2010;20: 393–402. doi:10.1101/gr.100545.109
12. Randhawa IAS, Khatkar MS, Thomson PC, Raadsma HW. Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genet.* 2014;15: 1–19. doi:10.1186/1471-2156-15-34
13. Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (*Mus musculus*). *PLoS Genet.* 2012;8. doi:10.1371/journal.pgen.1002891
14. Nielsen R, Williamson S, Kim Y, Nielsen R, Williamson S, Kim Y, et al. Genomic scans for selective sweeps using SNP data. *Genome Res.* 2005; 15:1566–1575. doi:10.1101/gr.4252305

2. EFEITOS DA ESTRUTURA POPULACIONAL E PARENTESCO NA MENSURAÇÃO DO DESEQUILÍBRIO DE LIGAÇÃO EM SOJA

RESUMO

A soja é a cultura agrícola com o maior aumento da produtividade nas últimas décadas. Esse aumento gradativo na produtividade da cultura deve-se principalmente as melhorias introduzidas no campo, associadas ao progresso de melhoramento genético. Recentemente, a análise de associação genômica ampla tornou o processo de identificação de genes mais acessível, sendo considerada uma das ferramentas promissora para auxiliar no entendimento da arquitetura genética dos caracteres de importância para a soja. Assim, o desequilíbrio de ligação e os estudos de estrutura populacional são extremamente importantes para obtenção de resultados confiáveis nas análises de associação genômica ampla, devido à influência destes no número de falsos positivos na análise. No presente estudo, a estrutura da população, o parentesco e o desequilíbrio de ligação (DL) foram avaliados em um conjunto de 95 genótipos de soja provenientes de diversas localidades do mundo, genotipados com a plataforma *180 k Axiom Soybean Genotyping Array*. O efeito da estrutura populacional sobre DL foi investigado por oito abordagens: r^2 (estimativa de DL sem qualquer correção), r^2s [usando a estrutura populacional de DAPC (r^2s_{dapc}), software STRUCTURE (r^2s_{struc}) e abordagem PCA (r^2s_{pca})], r^2v (usando a matriz de parentesco como uma correção) e r^2vs (considerando o parentesco e também cada das abordagens de estrutura populacional como correções). Para visualizar DL em todo o genoma, *heatmaps* foram criados e foi estimada a taxa de decaimento para cada abordagem. Análise do DL considerando r^2 revelou baixo a médio DL (0,175-0,45), considerando a distância de até 10 Mb. No entanto, quando as correções foram consideradas, verificamos que os pares de marcadores com baixo DL foram eliminados. Além disso, o DL estimado sem correção o decaimento em 0,1 permaneceu acima de 10 Mb, enquanto que o DL considerando a matriz de parentesco e a abordagem PCA como correções (r^2sv_{pca}), proporcionou um decaimento inferior a 0,1 em 5Mb. Adicionalmente, os padrões de LD em cromossomos diferentes revelaram a percepção de dependência funcional entre regiões. Os resultados mostram a importância de considerar a estrutura da população e o parentesco entre os genótipos em estudos de DL, bem como em nas análises de associação genômica ampla, devido à eliminação do DL falsos. No entanto, quando as espécies com baixos níveis de DL estão presentes, como espécies de autógamas, ao utilizar medidas de correções sem um estudo prévio das consequências pode-se reduzir as chances de sucesso na detecção de associações, devido à redução da extensão dos blocos de DL. Esses resultados destacam a influência da estrutura da população e do parentesco nos padrões de DL na soja. Além disso, são apresentados novos *insights* sobre a estrutura do genoma e dos fenômenos individuais de cada cromossomo.

Palavras-chave: Inter cromossomos; Decaimento; SNPs; Falsos positivos; GWAS

ABSTRACT

Soybean is the crop with the highest increases in yield in the last decades. The improvements in crop productivity are mainly provided due to technological progress associated with breeding processes. Recently, association-mapping approaches made the process of gene identification more accessible, being considered one of the most promising tools for genetic breeding. Thus, linkage disequilibrium and population structure studies are extremely important to obtain reliable results in the association mapping, due to their influence on the number of false positives. In the present study, population structure, genetic relationship and linkage disequilibrium (LD) were evaluated for a set of 95 worldwide soybean accessions, using the Axiom Soybean Genotyping Array. The effect of population structure on LD was investigated by

four approaches: r^2 (estimate of LD without any correction), r^2s [using the population structure from DAPC (r^2s_{dapc}), STRUCTURE software (r^2s_{struc}), and PCA approach (r^2s_{pca})], r^2v (using the kinship matrix as a correction), and r^2vs (considering kinship and each population structure approaches as corrections). To visualize LD throughout the genome, heatmaps and the decay were visualized for each approach. LD analysis considering r^2 for the whole genome revealed low to medium LD (0.175-0.45). However, when the corrections were used, we verified that pairs of markers with low LD were eliminated. Additionally, the LD measured with no correction remained above 10 Mb, while r^2 considering the kinship matrix and PCA approaches as corrections, provided LD inferior to 0.1 in 5Mb. The results show the importance of considering the population structure and genetic relationship among genotypes in LD studies, as well as in GWAS, by the elimination of false LD. However, when species with low LD levels are present, as inbred species, the correction measures used here reduced the chances of success for the detection of associations, due to the reduction of LD blocks, resulting from structure and relationship between the individuals. These results highlight the influence of population structure and kinship on LD patterns in soybean. Additionally, novel insights on genome structure and individual chromosome phenomena uncovered are presented. LD patterns in different chromosomes' reveal the perception of functional dependency.

Keywords: Inter-chromosomal; Decay, SNPs; False-positives; GWAS

2.1. Introdução

O desequilíbrio de ligação (DL) é definido como qualquer associação não aleatória ou preferencial de alelos de diferentes locos [1], resultando em desvios das frequências alélicas em relação às frequências esperadas sob independência [2]. A análise do DL ao longo do genoma possibilita a obtenção de informações sobre eventos passados, revelando respostas importantes sobre a história evolutiva de uma população, evidenciando o histórico de melhoramento e os padrões de subdivisão geográfica. Assim, a análise do DL em segmentos do genoma refletirá a história da seleção natural e artificial, conversão de genes, mutações e outros fatores que causam a evolução gênica [1,3]. O DL é importante para o sucesso de estudos genéticos, como o estudo de associação genômica ampla (Genome-Wide Association Studies- GWAS), sendo crucial para determinar a densidade de marcadores necessários nestas análises e definindo a resolução de mapeamento [4].

O DL é afetado por diversos fatores, sendo um indicador sensível dos processos genéticos que modificam a estrutura dos cromossomos e, portanto, do genoma [1]. Além da taxa de recombinação e mutação que afetam diretamente no DL, o tamanho populacional, sistema reprodutivo, deriva genética, seleção e estrutura da população podem ser citados como os principais mecanismos que interferem na magnitude e no decaimento do DL de uma população [5]. Outros fatores tais como, a migração e idade da população também podem modificar os padrões do DL [1,6].

Atualmente, a estrutura da população e o parentesco têm recebido mais atenção no estudo de DL e GWAS, por citado por ser a principal causa de ocorrência de falsos positivos nas análises [5]. Uma população qualquer pode apresentar indivíduos mais ou menos aparentados, compartilhando mais ou menos alelos. Em resultado a esta diferenciação, as populações podem apresentar sub-estruturação, que pode ser resultante da influência de vários fatores, como poucos eventos de recombinações entre os indivíduos da população [1]. Caso diferentes locos estejam fixados em diferentes sub-populações, ao considerar como uma única população podendo detectar um falso DL. Sendo que, tais alelos, podem ou não estar ligados fisicamente. Assim, a mistura de diferentes sub-populações pode gerar DL errôneo, sendo indispensável que a estrutura de população e também o parentesco sejam considerados nos estudos de DL e em GWAS [4].

O DL entre pares de locos em um genoma pode ser mensurado por várias estatísticas [7]. A escolha da estatística mais apropriada depende do objetivo do estudo em questão. As estatísticas mais utilizadas para o cálculo do DL entre pares de locos são D' e r^2 [8]. Ambos são inflacionados pelo reduzido tamanho amostral e pela frequência de alelos raros. Entretanto, o D' apesar de ser útil para a mensuração do DL não considera o histórico de recombinação da população, sendo esta uma informação de extrema importância quando se trata de estudos GWAS[4,8]. Portanto, considerando o objetivo de estudos futuros de GWAS, e melhor entendimento da história evolutiva do genoma, a estatística mais apropriada é a r^2 de Hill e Robertson [9]. Além disso, a estimativa r^2 pode ser interpretada como uma correlação entre os pares de marcadores [5], podendo assim, variar de 0 (ausência do desequilíbrio de ligação) a 1 (presença do desequilíbrio de ligação). Contudo, devido à diferenciação populacional que, em suma, resultam na subdivisão entre populações e parentesco elevado entre os indivíduos, o DL pode ser superestimado [10]. A superestimação do DL pode afetar na mensuração da quantidade e densidade de marcadores necessários para estudos de GWAS, e comprometer a resolução do mapeamento [11]. Em populações autógamas, como a soja, a endogamia influencia ainda mais drasticamente do que nas alógamas na estrutura da população. As medidas de DL consideram que os cruzamentos são aleatórios e não preferenciais, como ocorre na autogamia. Desta forma, a chance de erros na mensuração do DL é maior em autógamas. Mangin et al. [10] sugerem a utilização de medidas de correções para o cálculo do DL. Assim, esses autores propuseram a utilização de três novas medidas de DL que levam em conta, juntamente com o cálculo do r^2 , a estrutura da população, o parentesco e ambos em conjunto. Mangin et al. [11] e Cericola et al. [12] verificaram que com a utilização destas correções o DL decaiu de acordo com o esperado para as espécies em

estudo, sendo que quando não utilizaram as correções um elevado DL foi encontrado, mesmo em regiões distantes fisicamente.

A estrutura de uma população, ou de uma amostra de indivíduos, pode ser conhecida por diversas maneiras. O conhecimento da origem geográfica dos indivíduos pode ser um artifício para a separação em sub-populações. Entretanto, nem sempre temos acesso a esta informação, além de que outros processos podem influenciar mais fortemente a estruturação do que a procedência geográfica. Os métodos mais utilizados atualmente para inferir a estrutura da população, independentemente de ter ou não informações *a priori* sobre os indivíduos, são: a Análise de Componentes Principais (ACP) [13], a inferência *Bayesiana* com aplicação do programa STRUCTURE [14] e recentemente, a Análise Discriminante de Componentes Principais (DAPC) também tem sido utilizada com sucesso [15]. Basicamente, as três abordagens fornecem a estrutura da população, no entanto em situações complexas, na qual as populações ancestrais não são bem diferenciadas ambas as análises devem ser investigadas para buscar a melhor interpretação e confiabilidade nos resultados apresentados [16].

A soja cultivada é uma planta autógama pertencente à classe das dicotiledôneas, membro da família Fabaceae, gênero *Glycine*, subgênero soja e espécie *Glycine max* (L.) Merrill [17]. A espécie possui $2n=40$ cromossomos, tetraplóide antigo, resultante do processo de diploidização na qual atualmente apresenta comportamento citológico diplóide [18]. A cultura é uma das principais fontes de proteína e óleo vegetal, representando a planta com o maior teor de proteína por hectare em uma lavoura [19,20]. A soja representa uma das principais culturas de exploração agrícola, sendo um dos propulsores da economia mundial. Para que a cultura continue a receber destaque, incrementos na produtividade e demais caracteres de importância econômica são necessários. A utilização de novas abordagens no melhoramento é considerada essencial para acelerar os ganhos genéticos de um programa de melhoramento.

Desta forma, o presente estudo teve como objetivo explorar a utilização de marcadores moleculares SNPs (*Single Nucleotide Polymorphisms*) para o estudo do DL considerando algumas medidas de correções como, estrutura da população e parentesco dos indivíduos, a fim de elucidar a importância destas correções para a redução dos falsos-positivos em estudo de DL e demonstrar sua importância para os demais estudos que são baseados no DL.

2.2. Material e métodos

2.2.1. Material vegetal e genotipagem

Para este estudo foi utilizado um painel de acessos de soja composto por 95 genótipos, dentre estes; 80 são PIs (*Plant Introductions*) de soja selecionadas em estudo prévio realizado por Mulato et al. [21], provenientes de 38 diferentes localidades do mundo; e 15 são genótipos melhorados de origem brasileira, também já reportados em estudo por Marconato et al. [22].

Para a genotipagem, folhas dos 95 genótipos foram coletadas, de uma única planta, cultivada em casa de vegetação, após o aparecimento do primeiro par de folhas verdadeiras. As folhas foram maceradas com auxílio de um macerador automático e após isto se procedeu a extração de DNA pelo protocolo descrito por Doyle e Doyle [23]. A qualidade do DNA foi avaliada por eletroforese em gel de agarose a 1% (p/v) corados com SYBR[®] Safe (Invitrogen[™]). A quantificação foi feita no aparelho QuantiFluor[®], e as amostras foram diluídas para a concentração de 100 ng μL^{-1} e acondicionadas em placas para genotipagem com marcadores SNPs (*Single Nucleotide Polymorphisms*). A genotipagem foi realizada na plataforma da Affymetrix (*Axiom[®] Soybean Genotyping Array*) contendo 180.961 marcadores SNPs para soja cultivada e selvagem.

2.2.2. Filtragem e imputação de dados perdidos

A partir dos marcadores obtidos pela genotipagem com o *Axiom[®] Soybean Genotyping Array*, foi realizada uma filtragem utilizando os critérios estabelecidos pelos desenvolvedores do arranjo para garantir a qualidade e reprodutibilidade dos marcadores. Assim, de acordo com os critérios recomendados, os seguintes parâmetros foram utilizados: $\text{DQC} \geq 0,82$ (*Dish quality control*, medida da resolução da distribuição dos valores de contraste), $\text{QC call rate (Quality control call rate)} \geq 92$, $\text{Average call rate for passing} \geq 97$, $\text{Minor allele cutoff} \geq 2$ e marcadores *Poly High Resolution*. Além disso, uma segunda filtragem foi realizada, na qual restaram marcadores com menos de 10% de dados perdidos e SNPs com $\text{MAF} < 0,05$ (*Minor Allele Frequency*) foram eliminados.

A imputação de genótipos foi realizada a fim de não perder uma quantidade de dados significativa [24]. Desta forma, os dados perdidos resultantes das filtrações anteriores foram codificados e passaram pela imputação com o programa BEAGLE 4.1. [25].

2.2.3. Estrutura populacional e matriz de parentesco

A estrutura populacional do painel de soja foi inferida através de três abordagens. O programa STRUCTURE [26] foi utilizado para estimar a probabilidade de cada genótipo ter sido originado de cada sub-população a partir das frequências alélicas do painel [26]. Para cálculo do número de sub-populações mais provável (K) foram utilizados os seguintes parâmetros: 10 simulações independentes com K variando entre 2 e 10, 500.000 *burn-ins*, 1.000.000 passos de Monte Carlo Cadeias de Markov (MCMC), modelo de não-mistura (*no admixture*) e frequências alélicas correlacionadas. A seleção do número K mais provável foi realizada por meio dos valores de ΔK [27]. A segunda abordagem foi a *Discriminant Analysis of Principal Components* (DAPC) [15], realizada com o pacote *adegenet* [28] na plataforma R [29]. Essa análise é uma alternativa ao programa STRUCTURE [26], comumente utilizada para análise da estrutura populacional. A DAPC permite a análise de uma grande quantidade de dados em menor tempo e com a utilização de menos recursos computacionais. Por meio desta análise pode-se identificar a formação de agrupamentos entre os indivíduos, sendo um método multivariado desenvolvido para identificar e descrever grupos de indivíduos geneticamente relacionados. A DAPC fornece uma avaliação visual de diferenciação populacional, e da contribuição de alelos individuais para cada população formada. A terceira abordagem foi a *Principal Component Analysis* (PCA) [13], na qual foi realizada com o pacote *adegenet* [28] na plataforma R [29].

A matriz de parentesco (k) foi inferida pelo pacote computacional do R [29] *synbreed* [30], utilizando os SNPs resultantes das etapas de filtragem e imputação. O cálculo foi feito pela função “*kin*”, calculando desta forma o “*simple matching coefficient*” [31]. Para este cálculo os genótipos foram codificados como 2 ou 0, que corresponde a contagem de um dos alelos para cada loco.

2.2.4. Desequilíbrio de ligação

Para a análise do Desequilíbrio de Ligação (DL), foi utilizada a estatística r^2 [9], como medida própria do DL, resultando, portanto, em múltiplos testes. Além da mensuração do r^2 , o efeito da estrutura populacional sob o DL foi investigado utilizando três abordagens como sugerido por Mangin et al. [10]. Assim, além do comumente utilizado r^2 (DL mensurado sem nenhuma medida de correção), utilizou-se o r^2_s (DL mensurado levando em consideração a estrutura populacional), r^2_v (DL mensurado considerando o parentesco), e r^2_{vs}

(DL mensurado considerando ambas as medidas de correções, de parentesco e estrutura populacional). A estimativa de todas as medidas de DL foi com o *LDcorSV*[32]na plataforma R [29].

Neste estudo utilizaram-se três abordagens para inferir a estrutura da população: programa STRUCTURE, DAPC e PCA. As matrizes de estruturação, proveniente de cada abordagem, foram utilizadas no pacote *LDcorSV* como medidas de correções para o cálculo do DL. O mesmo foi feito para o coeficiente de parentesco. Assim, foram geradas 8 estimativas diferentes de DL: r^2 , r^2_v (correção para parentesco), r^2s_struc , r^2s_pca e r^2s_dapc (correção para estrutura da população para cada abordagem), r^2ks_struc , r^2ks_pca e r^2ks_dapc (correção do parentesco em conjunto a cada abordagem de estruturação). Para visualizar o decaimento do DL através do genoma foi gerado um gráfico contendo todas as mensurações, na qual os valores de DL entre os pares de marcadores foram plotados contra a distância em centiMorgans (cM), com base nas informações de posição dos marcadores SNPs. A taxa de decaimento do DL no genoma foi estimado segundo a equação proposta por Remington [33].

Com auxílio do pacote *LDheatmap*[34]na plataforma R [29]foram gerados para cada um dos 20 cromossomos disponíveis e para as 8 mensurações de DL, gráficos do tipo *heatmap* [34]para a visualização dos padrões do DL encontrados, envolvendo SNPs fisicamente ligados.

A fim de investigar evidências de dependência funcional entre as regiões e também em relação ao processo de duplicação que ocorreu no genoma da soja foram investigadas pela presença de DL entre os pares de SNP localizados em diferentes cromossomos. Para isso, foi considerado apenas pares de SNPs em que DL foi $\geq 0,9$ e FDR (*False-discovery rate*) $\leq 0,01$. Para verificar a dependência funcional entre essas regiões, foi verificado a presença de QTL comuns compartilhados utilizando-se a ferramenta SoyBase *browser*, disponível em <https://www.soybase.org/>.

2.3. Resultados e discussão

2.3.1. Genotipagem

Com um número inicial de 186.961 marcadores SNPs, resultantes da genotipagem na plataforma da *Affymetrix* (*Axiom® Soybean Genotyping Array*), após a utilização dos parâmetros recomendados do software *Axiom® Analysis Suite*, nenhum SNP foi eliminado. Entretanto, recomenda-se eliminar os marcadores que não são considerados como “*Poly High*

Resolution”, sendo aqueles que não são classificados como acurados e polimórficos nas etapas de sequenciamento. Assim, após a utilização deste parâmetro obteve-se 24.230 SNPs, os quais foram utilizados para a próxima etapa de filtragem. Sendo que, apenas 12.936 marcadores SNPs restaram após a eliminação de marcadores SNPs com mais de 10% de dados perdidos e MAF (*Minimum Minor Allele Frequency*) > 0,05. Em seguida as etapas de filtragem, os dados foram submetidos à imputação no programa BEAGLE 4.1 [25]. Após as etapas de filtragem e imputação de dados, foram obtidos 12.133 marcadores SNPs, os quais foram utilizados para as análises. A distribuição destes 12.133 SNPs ao longo dos 20 cromossomos da soja está demonstrada na Figura 1.

Os SNPs resultantes apresentaram uma boa distribuição ao longo dos 20 cromossomos, sendo que alguns padrões podem ser verificados em todos os cromossomos, como a maior concentração de marcadores SNPs nas regiões pericentroméricas de cada cromossomo.

2.3.2. Estrutura da população e matriz de parentesco

Sabendo-se que a estrutura genética da população e o parentesco são considerados fatores críticos que influenciam drasticamente na confiabilidade dos falsos positivos das análises de mapeamento associativo e estudos de DL [10,11,35], ambos foram estimados para os 95 genótipos e com 12.133 SNPs. Foram utilizadas três abordagens distintas para acessar a estrutura da população: STRUCTURE, DAPC e PCA. Para inferir a relação genética entre os genótipos, foi calculada a matriz de parentesco com o “*simple matching coefficient*”.

A abordagem STRUCTURE indicou a presença de duas subpopulações, $K=2$. (Figura 2). Dos 95 genótipos, 30 foram atribuídos à primeira subpopulação (cor azul na Figura 2), e 11 foram atribuídos à segunda subpopulação (cor vermelha na Figura 2). Além disso, 54 genótipos foram considerados como “*admixed*”. A primeira subpopulação é composta por genótipos com origem mista de todos os continentes. Na segunda subpopulação foram agrupados genótipos predominantemente com origem nas Américas do Norte e do Sul, bem como da Ásia.

A DAPC, utilizando o método *k-means*, na qual atribui a cada indivíduo um agrupamento, foi verificada a presença de 3 grupos bem delimitados. Todos os genótipos na DAPC foram atribuídos aos seus respectivos grupos com 100% de probabilidade, não havendo a presença de indivíduos classificados como “*admixed*”. O fato de a DAPC ter captado melhor a delimitação de grupos pode ser explicado pela análise utilizar combinações

de alelos que sumarizam maior parte da variação genética [15], ao contrário da análise do STRUCTURE que considera todos os alelos. Resultados similares foram relatados para um painel de 170 acessos de girassol (*Helianthus annuus*) [36]. Pela DAPC verificamos que o primeiro agrupamento (cor azul na Figura 3) é formado por 26 genótipos de origem da América do Sul e Ásia. Enquanto que o segundo, (cor vermelha na Figura 3) compreende 28 materiais, que em sua maioria são originários da Ásia. Já último agrupamento (cor verde na Figura 3), apresenta a Ásia, Américas do Norte e do Sul e África como regiões de origem dos 41 genótipos ali alocados. Assim, como pode ser verificado mesmo que os agrupamentos da análise DAPC foram mais claros, estes ainda não apresentaram correspondência com a origem geográfica. Agrupamentos similares foram verificados nas análises STRUCTURE e PCA.

A PCA não revelou nenhuma separação clara de grupos. Os primeiros cinco componentes explicam 27,4 % da variância, sendo: 8,0% para PC1; 5,9% para o PC2; 5,5% para PC3; 4,5% para o PC4 e 3,5% para PC5. Baseado no gráfico da Figura 4, na qual os dois primeiros componentes estão representados, que não houve agrupamentos bem definidos. Por ser o centro de origem e diversidade da soja, a China e o Japão apresentam a maior variabilidade da cultura [37] e, portanto, isto explica a dispersão dos genótipos nas sub-populações encontradas pela abordagem do STRUCTURE, DAPC. Além disso, em ambas as abordagens de estruturação, foram verificadas que a maioria dos genótipos Brasileiros foram alocados em sub-populações, na quais há também os genótipos provenientes dos EUA. A proximidade dos materiais brasileiros aos dos EUA já é bem conhecida: Hiromoto e Vello [38] afirmam que a maioria dos materiais brasileiros é proveniente do germoplasma originário da América do Norte. Apesar dos genótipos apresentarem dissimilaridade entre eles, o que pode ser verificado na Figura 5, que representa a matriz de parentesco, estes não formam agrupamentos genéticos bem definidos.

A partir do cálculo do “*simple matching coefficient*”, que representa a probabilidade de alelos amostrados aleatoriamente a partir de dois indivíduos serem idênticos [39], foi construído um dendrograma em conjunto com seu *heatmap*. Na Figura 5 é demonstrada a relação genética entre os 95 genótipos do painel, sendo que, no *heatmap* quanto mais escuro o tom vermelho, mais similares são os genótipos, e quanto mais clara a tonalidade, menos similares. Os valores da matriz de parentesco variaram de 1,05 (próximo ao branco) a 2,00 (vermelho), e a média foi de 1,4 (laranja). Os valores apresentados demonstram que os genótipos apresentam em geral valores médios de similaridade, o que pode ser verificado pela predominância da cor laranja no *heatmap*. Destacam-se a presença três agrupamentos, sendo que o menor deles contém dois genótipos, PI212602 (Afeganistão) e Paranagoiana (cultivar

comercial brasileiro), e que apresentam os menores valores de parentesco entre os demais genótipos, evidenciado pela cor esbranquiçada no *heatmap*. Em geral, os agrupamentos formados não apresentam correspondência com a proveniência geográfica de cada genótipo, assim, mesmo os genótipos brasileiros foram distribuídos nos três agrupamentos.

Acredita-se que, como os genótipos pertencentes ao painel são provenientes de várias regiões, as análises de estruturação e parentesco não foram capazes de definir agrupamentos genéticos claros e bem definidos. Além disso, a ausência de agrupamentos genéticos bem definidos pode ser consequência da base genética estreita da espécie, resultante do processo de domesticação e melhoramento [37,38]. A não identificação de agrupamentos bem definidos com correspondência a origem geográfica também foi verificada por outros autores [40,41].

2.3.3. Desequilíbrio de ligação

Neste estudo foi utilizado para estimação do DL entre pares de locos fisicamente ligados o cálculo do r^2 de Hill e Robertson [9], sendo o método mais utilizado e recomendado. Entretanto, como já mencionado, vários fatores podem afetar diferentemente o decaimento do DL, sendo que estrutura da população e o parentesco podem ser considerados como um dos grandes complicadores para a mensuração da extensão do DL [4]. Desta forma, a fim de controlar os efeitos da estrutura da população e parentesco, a metodologia proposta por Mangin et al. [10] foi utilizada. Assim, o DL foi estimado através do cálculo do r^2 e, também, utilizando-se de medidas de correções para a estrutura da população (r^2s) e parentesco (r^2v) entre os indivíduos.

FDR de 0,01 foi utilizado como limiar para definição de quais pares de locos apresentam DL significativo. Assim, aqueles definidos como significativos foram utilizados para a construção dos *heatmaps* e para o cálculo da taxa de decaimento do DL sobre o genoma. Os *heatmaps* produzidos para todos os cromossomos, considerando a medida de DL simples e com correções elucidam a presença de blocos de ligações com alto DL em regiões específicas em cada cromossomo, representados em vermelho (Figura 6). À medida que o DL atinge valores médios e baixos, estes são representados por tonalidades entre laranja e amarelo, e entre verde e azul, respectivamente. Na Figura 6 está representado o *heatmap* do cromossomo 17. Neste e também nos demais cromossomos (dados não apresentados), verifica-se regiões de alto DL intercaladas por regiões de baixo DL. Desta forma, como já reportado anteriormente [3], o DL não foi uniformemente distribuído ao longo do genoma.

As estimativas sem correções (r^2) apresentam locos em DL mesmo a longas distâncias físicas (Figura 6), podendo atrapalhar nas interpretações e resultando em DL não confiável [10]. Associações entre locos fisicamente distantes podem ser resultado da superestimação do DL devido alguns fatores, tais como a estrutura da população [10,42], havendo a necessidade de um método mais detalhado e confiável para a estimação do DL. Quando a matriz de parentesco foi levada em consideração, o DL (r^2v) foi afetado drasticamente. Ou seja, houve uma redução da quantidade de associações entre locos fisicamente distantes, sendo que os blocos de altos valores de DL permaneceram intercalados com blocos de baixos valores de DL (Figura 6).

Considerando a abordagem do programa STRUCTURE, podemos verificar no *heatmap* da Figura 6, representado por r^2s_struc que este não foi suficiente para a redução das associações a longas distâncias. O *heatmap* r^2s_struc apresenta o mesmo padrão do *heatmap* de r^2 , e apesar de haver diferenças nos valores de DL, estas não são perceptíveis no *heatmap*, mas podem ser visualizadas na curva de decaimento da Figura 7. O mesmo pode ser verificado quando foi utilizada a DAPC para correção do DL. O *heatmap* r^2s_dapc apresenta sutis diferenças em comparação ao de r^2 (DL sem correções). Já o *heatmap* titulado como r^2s_pca , na qual o DL foi corrigido pela estrutura da população obtida pela PCA, apresentou resultados similares à análise r^2v , com uma grande redução de DL entre locos distantes fisicamente.

Apesar da estrutura da população afetar diretamente na estimação do DL, no caso do painel de acessos avaliados, foi observada uma correção mais sutil em comparação ao uso da estimativa de parentesco entre os genótipos. Desta forma, quando r^2vs foi plotado em um *heatmap*, observa-se que o parentesco em conjunto com qualquer informação sobre estrutura genética (STRUCTURE, DAPC ou PCA) resultou em menor número de associações entre locos distantes fisicamente. De acordo com a Figura 6, é possível notar ainda que quando as medidas de correções são utilizadas os valores de DL relativamente baixos são eliminados (cor verde). Entretanto, os valores de DL considerados como altos (representados em vermelho escuro nos *heatmaps*) permanecem inalterados. A similaridade da distribuição de blocos com altos valores de DL demonstram que as medidas de correções são eficientes para reduzir o viés do cálculo do DL nas populações [11]. E de maneira geral, as correções reduzem os DL considerados como falsos positivos resultantes devido à própria estrutura da população ou mesmo do parentesco entre os indivíduos e mantêm o DL real.

Para a melhor visualização e sumarização dos valores do DL ao longo do genoma, considerando-se as 8 medidas, utilizou-se de recursos gráficos. Os valores das medidas do DL

foram plotados em relação às respectivas distâncias entre os pares de locos SNP [7], que pode ser visualizado na Figura 7. Considerando os valores de r^2 , ou seja, cálculo do DL sem correções, plotados contra as distâncias dos pares de marcadores, observamos que o decaimento do DL se estende por mais de 10 Mb no genoma (considerando que o decaimento em $r^2=0,1$), com r^2 variando de 0,45 a 0,175. Estes valores foram relativamente altos quando comparados aos resultados apresentados por Hyten et al. [3]. Estes autores demonstraram que em alguns segmentos do genoma da soja, utilizando a mesma medida de r^2 , o DL decaiu para 0,1 em uma distância de 90 kb nas *landraces*, 212 kb nos ancestrais selvagens da soja e 574 kb nas cultivares elites. Entretanto, estes autores ainda observaram que em outros segmentos do genoma da soja, r^2 não atingia 0,1 em nenhum momento, concluindo que os segmentos não apresentam um valor de DL consistente. Os resultados descritos por Hyten et al. [3] podem ser explicados pelo número reduzido de marcadores (558 SNPs) e genótipos por grupo (média de 20) utilizados, necessitando assim de estudos com um maior número de marcadores e uma análise no genoma inteiro.

O sistema reprodutivo da planta é um fator importante que afeta o padrão do DL [43]. Plantas autógamas, como a soja, apresentam elevado DL quando comparados a plantas alógamas, como o milho (*Zea mays*). Em espécies autógamas o DL geralmente se estende a longas distâncias [44]. Mesmo que haja recombinação efetiva, pela taxa de fecundação-cruzada, em autógamas a autofecundação é predominante e aumenta o grau de homozigose, limitando o número de heterozigotos que podem ser obtidos nas recombinações [45]. Zhou et al. [46] consideram que o decaimento lento do DL pode ser um fenômeno comum em variedades de soja asiáticas. A maioria dos acessos da Coleção de Germoplasma de Soja do USDA (*United States Department of Agriculture*) e do painel avaliado neste estudo possui origem asiática. Zhang et al. [47], ao analisar o padrão de decaimento do DL ao longo da eucromatina e da heterocromatina em um painel de soja, revelaram que o DL se estende a uma distância bem maior do que aquelas anteriormente relatadas em soja [3]. Além disso, quando o DL foi calculado apenas para a heterocromatina não foi verificado o decaimento para 0,1 até uma distância de 10Mb. Quando Zhang et al. [47] analisaram as regiões de eucromatina, verificaram um decaimento um pouco mais rápido, sendo que o DL atingiu 0,1 em menos de 2Mb, demonstrando que as regiões menos condensadas, e, portanto, que possuem mais genes ativos, estão sujeitas a maiores taxas de recombinações e por conseguinte, possuem um decaimento mais rápido. Em soja, o decaimento a longas distâncias também já foi verificado por Mamidi et al. [48]. Os autores demonstraram que houve

inconsistência nos resultados em duas populações de soja, na qual em uma população o DL decaiu para 0,1 em 7Mb e na outra em 6 Mb.

Ao analisar cada cromossomo individualmente (material não apresentado), percebeu-se que eles não seguem um padrão de decaimento, como apresentado também por Mamidi et al.[48]. Cada um dos cromossomos possui um padrão particular de decaimento, assim como na distribuição do DL ao longo do genoma. Hyten et al. [3] relata o mesmo achado, em que os segmentos do genoma apresentam padrões variados e que foram também reportados em *Arabidopsis* por Hagenblad e Nordborg[49]. Deste modo, o DL é dependente da região do genoma analisado: cromossomos ou regiões diferentes provavelmente irão apresentar diferentes padrões de DL. Da mesma forma, pode-se inferir que populações diferentes, as quais possuem histórias evolutivas diferentes, assim como os cromossomos, apresentam padrões distintos de DL [48].

Como já mencionado, a estrutura da população e o parentesco são considerados uns dos principais complicadores para a estimação do DL e para o elevado número de falsos positivos em análises de associação [4]. Quando adicionada a matriz da estrutura da população, obtida pela DAPC e STRUCUTURE, verifica-se que o DL decai relativamente mais rápido, embora ainda se estenda a longas distâncias (Figura 7). Enquanto a medida r^2 decaiu de 0,45 para 0,175 em 10Mb, o $r^2_{s_struc}$ decaiu de 0,45 a 0,155 e o $r^2_{s_dapc}$ decaiu de 0,45 a 0,151 nos mesmos 10 Mb. Entretanto, quando utilizada a PCA($r^2_{s_pca}$) para correção do DL, verificou-se que a curva iniciou em 0,45 e decaiu para 0,1 em menos de 5Mb. Portanto, o $r^2_{s_pca}$ apresentou-se como a melhor forma de correção do DL, considerando a estrutura genética populacional. O decaimento em 5Mb apresenta-se de acordo com o exposto para a soja por Mamidi et al. [48]. Não obstante, o mesmo foi evidenciado com a utilização da matriz de parentesco como correção. Com a medida de r^2_v , o DL caiu drasticamente, de 0,45 a 0,1 em 5Mb. Assim, como já verificado nos *heatmaps*, quando utilizado das medidas de r^2_{vs} , independentemente do tipo de análise de estruturação, o DL foi ainda mais reduzido. Portanto, r^2_{vs} apresenta-se como a melhor abordagem a ser utilizada, independente da forma a qual a estrutura da população foi inferida e deve ser utilizada para a eliminação dos falsos-positivos. Outros trabalhos, como Hyten et al.[3] e Mangin et al. [11], enfatizam a importância e a necessidade da utilização de alguma medida de correção para estudos de DL, bem como estudos de mapeamento associativo e demais que utilizam a medida do DL. Além disso, destaca-se a medida de $r^2_{vs_pca}$, na qual o DL decaiu para 0,1 em menos de 3,75 Mb. Desta maneira, podemos verificar que apesar da estrutura da população utilizada como medida de correção reduza as associações de locos a longas distâncias, quando comparado a correção

proposta por r^2v , na qual utiliza a matriz de parentesco como correção, essa foi mais efetiva, reduzindo a taxa de decaimento em menores distâncias. Resultados similares a este trabalho foram demonstrados por outros autores que utilizaram medidas de correções [11,35,44,50]. Adetunji et al. [35] avaliaram o DL em linhagens de beterraba (*Beta vulgaris*). Quando o parentesco e estruturação genética não foram utilizados como correções, o DL prolongou-se por mais de 50 cM, mas quando os valores foram corrigidos, o DL diminuiu bruscamente em uma distância de menos de 6 cM. O mesmo pode ser verificado para os resultados dos estudos de associação com milho (*Z. mays*) [51].

Em síntese, a correção dos valores de DL usando informações sobre estrutura e parentesco é fundamental para garantir a confiabilidade dos resultados, reduzindo a quantidade das associações espúrias entre locos não intimamente e funcionalmente relacionados. A não utilização destas correções pode levar a estimativas tendenciosas do DL e, portanto, a uma escolha inadequada da densidade do marcador para estudo de GWAS [11]. Ainda assim, recomenda-se que estudos de GWAS desenvolvam estudos mais aprofundados para a escolha da estimativa de DL conforme o padrão de estrutura e parentesco populacional. Na literatura são citados vários métodos, sendo apenas alguns testados neste trabalho. A escolha do método deve ser feita baseada em vários fatores que variam de acordo com a população de trabalho.

Destaca-se que, mesmo com a utilização de medidas de correções dos vieses do cálculo do DL, este ainda persiste entre locos separados por longas distâncias. A grande extensão do DL pode ser explicada pelo tipo de população em questão, na qual inclui na maioria dos materiais poucos melhorados. No milho, cultura alógama, em que o DL decai rapidamente em cerca de 0,2 kb [52], foi encontrado por Remington et al. [33] e Jung et al. [53] que em algumas populações pouco domesticadas o DL decaiu por maiores distâncias. A presença de blocos de alto DL e decaimento a longas distâncias podem evidenciar que regiões estão sobre altas taxas de recombinações e, portanto, relacionadas aos processos de seleção e adaptação da espécie [12,54]. Logo, blocos com alto DL que se estendem a longas distâncias, configuram um DL que é considerado como ideal para os métodos de GWAS, permitindo uma cobertura eficiente do genoma com um número moderado de marcadores por região. Assim, sugere-se a utilização de ferramentas como GWAS para investigação dos segmentos com altos DL, que possivelmente estão associadas a importantes QTLs.

2.3.4. Desequilíbrio de ligação inter-cromossômico

O DL também pode ser encontrado entre locos localizados em diferentes cromossomos. Este fenômeno pode ser observado quando os vieses causados pela estrutura da população e parentesco no cálculo do DL não são corrigidos [10,42]. Além disso, locos associados a longas distâncias ou mesmo em cromossomos diferentes podem estar envolvidos em funções em comum e, portanto, submetidos às mesmas pressões seletivas [42]. Para verificar a presença de DL inter-cromossômico foi selecionado apenas a medida de DL $r^2_{vs_pca}$, que mais reduziu os vieses de estrutura e parentesco (Figuras 6 e 7). Os valores de DL, com correções de estrutura pela PCA e parentesco, foram verificados a partir da presença de locos associados em cromossomos distintos. Os valores de $r^2_{vs_pca}$ acima de 0,9 e o FDR $\geq 0,01$ foram considerados como *threshold* para verificar a presença de DL significativo entre cromossomos distintos.

Considerando o cálculo do DL sem correções (r^2), detectamos 447 SNPs apresentando alto DL localizados em cromossomos distintos. Contudo, assim como é demonstrado no *heatmap* do cromossomo 17 (Figura 6), ao utilizar as medidas de correções, o DL entre locos distantes é reduzido, bem como a taxa de decaimento. Assim, o mesmo foi observado no DL inter-cromossômico. Foram observados 417 SNPs e 364 SNPs apresentando DL inter-cromossômico para as medidas de r^2_s e $r^2_{v_pca}$, respectivamente. Ao analisar o DL inter-cromossômico apresentado por $r^2_{sv_pca}$ é observada redução de 20,08%, quando comparado ao r^2 . Assim, utilizando a medida de DL $r^2_{vs_pca}$ verificou-se a presença de 354 associações entre locos de diferentes cromossomos, que podem ser resumidas em 37 regiões apresentadas na Tabela 1.

A fim de verificar a dependência funcional, evidenciada pela presença em comum de QTLs já reportados para as regiões, utilizou-se o recurso *SoyBase browser* disponível no <https://www.soybase.org/>. Na Tabela 1 estão sumarizados os QTLs reportados por vários autores, encontrados em ambas as regiões interligadas entre os cromossomos. Com exceção de 5 regiões (8, 25, 30, 33 e 34) todos as demais 32 regiões apresentaram QTLs em comum. A presença de QTLs em comum sugere que as regiões com associações em diferentes cromossomos podem ser responsáveis pelos mesmos caracteres.

A soja possui em genoma relativamente complexo, com várias duplicações antigas e recentes [55,56]. Dados apresentados no *SoyBase* mostram a presença destas duplicações em todos os cromossomos. Grandes duplicações são produzidas geralmente pelos eventos de duplicações do genoma completo, sendo as duplicações em pequena escala aquelas resultantes

do processo de *crossing-over* desigual e até mesmo anomalias cromossômicas [57]. Desta forma, como o genoma da soja apresenta uma série de duplicações, foram também investigados se as regiões ligadas localizadas em cromossomos diferentes são na verdade regiões duplicadas do genoma. Todavia, das 37 regiões, foram encontradas apenas 5 regiões duplicadas, sendo 4 (3,17,21 e 33) delas relacionadas a duplicações antigas e apenas 1 (18) considerada como duplicação recente.

Em suma, estes resultados indicam que quando estimado o DL, mesmo após o controle da estrutura da população e parentesco, esse ainda pode revelar várias regiões situadas em cromossomos diferentes com alto DL, que em sua maioria não são resultados de duplicações e sim participam de processos em comum, possivelmente apresentando dependência funcional.

2.4. Conclusão

O uso da estrutura populacional, bem como as relações entre os genótipos, é necessário para a estimação de DL real em qualquer população de estudo. A escolha do método para inferir a estrutura da população é um passo importante para o estudo de DL e, portanto, para o controle dos falso positivos nas GWAS. Caso esta etapa não seja realizada corretamente, as chances de sucesso podem ser reduzidas em estudo GWAS, e nos esforços aplicados ao melhoramento genético da soja.

REFERÊNCIAS

1. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet.* 2008;9: 477–485. doi:10.1038/nrg2361
2. Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 2002;18: 19–24. doi:10.1016/S0168-9525(01)02550-1
3. Hyten DL, Choi IY, Song Q, Shoemaker RC, Nelson RL, Costa JM, et al. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics.* 2007;175: 1937–1944. doi:10.1534/genetics.106.069740
4. Flint-Garcia SA, Thornsberry JM, Buckler ES. Structure of Linkage Disequilibrium in Plants. *Annu Rev Plant Biol.* 2003;54: 357–374. doi:10.1146/annurev.arplant.54.031902.134907

5. Abdurakhmonov IY, Abdukarimov A. Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources. *Int J Plant Genomics*. 2008;2008: 1–18. doi:10.1155/2008/574927
6. Caldwell KS. Extreme Population-Dependent Linkage Disequilibrium Detected in an Inbreeding Plant Species, *Hordeum vulgare*. *Genetics*. 2005;172: 557–567. doi:10.1534/genetics.104.038489
7. Gaut BS, Long AD. The lowdown on linkage disequilibrium. *Plant Cell*. 2003;15: 1502–1506. doi:10.1105/tpc.150730
8. Gupta PK, Rustgi S, Kulwal PL. Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol*. 2005;57: 461–85. doi:10.1007/s11103-005-0257-z
9. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet*. 1968;38: 226–231. doi:10.1007/BF01245622
10. Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb)*. Nature Publishing Group; 2012;108: 285–291. doi:10.1038/hdy.2011.73
11. Mangin B, Sandron F, Henry K, Devaux B, Willems G, Devaux P, et al. Breeding patterns and cultivated beets origins by genetic diversity and linkage disequilibrium analyses. *Theor Appl Genet*. Springer Berlin Heidelberg; 2015;128: 2255–2271. doi:10.1007/s00122-015-2582-1
12. Cericola F, Portis E, Lanteri S, Toppino L, Barchi L, Acciarri N, et al. Linkage disequilibrium and genome-wide association analysis for anthocyanin pigmentation and fruit color in eggplant. *BMC Genomics*. 2014;15: 896. doi:10.1186/1471-2164-15-896
13. Price AL, N.j.patterson, R.m.plenge, M.e.weinblatt, N.a.shadick. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38: 904–909. doi:10.1038/ng1847
14. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155: 945–959. doi:10.1111/j.1471-8286.2007.01758.x
15. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet*. BioMed Central Ltd; 2010;11: 94. doi:10.1186/1471-2156-11-94
16. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2: 2074–2093. doi:10.1371/journal.pgen.0020190

17. Wojciechowski MF, Mahn J, Jones B. Fabaceae - Legumes. The Tree of Life Web Project, v.14, 2006. <http://tolweb.org/Fabaceae/21093/2006.06.14>. Accessed 12 aug 2016.
18. Hymowitz T. Speciation and Cytogenetics. In: BOERMA HJ, SPECHT JE. (Ed.). Soybeans: improvement, production and uses. 3. ed. Madison: American Society of Agronomy, p. 97-136, 2004.
19. Brzostowski LF, Pruski TI, Specht JE, Diers BW. Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. *Theor Appl Genet*. Springer Berlin Heidelberg; 2017;130: 2315–2326. doi:10.1007/s00122-017-2961-x
20. Wilson RF. Soybean: Market Driven Research Needs. *Genet Genomics Soybean*. 2008; 3–15. doi:10.1007/978-0-387-72299-3_1
21. Mulato BM, Möller M, Zucchi MI, Quecini V, Pinheiro JB. Genetic diversity in soybean germplasm identified by SSR and EST-SSR markers. *Pesqui Agropecuária Bras*. 2010;45: 276–283. doi:10.1590/S0100-204X2010000300007
22. Marconato MB, Pereira EM, Silva FM, Bizari EH, Pinheiro JB, Mauro AO, et al. Genetic divergence in a soybean (*Glycine max*) diversity panel based on agro-morphological traits. *Genet Mol Res*. 2016;15: 1–10. doi:10.4238/gmr15048980
23. Doyle JJ, Doyle JL. A rapid isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*. 1987;19:11–15.
24. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*. *Plant Methods*; 2013;9: 29. doi:10.1186/1746-4811-9-29
25. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am J Hum Genet*. 2007;81: 1084–1097. doi:10.1086/521987
26. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association Mapping in Structured Populations. *Am J Hum Genet*. 2000;67: 170–181. doi:10.1086/302959
27. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*. 2005;14: 2611–20. doi:10.1111/j.1365-294X.2005.02553.x
28. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. 2008;24: 1403–1405. doi:10.1093/bioinformatics/btn129
29. R Core Team: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2013.
30. Wimmer V, Albrecht T, Auinger H-J, Schön C-C. synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*. 2012;28: 2086–2087. doi:10.1093/bioinformatics/bts335

31. Schlee D, Sneath P, Sokal R FW. Numerical taxonomy. The principles and practice of numerical classification. *Syst Zool.* 1975;24: 263.
32. Desrousseaux D, Sandron F, Siberchicot A, Cierco-ayrolles C, Mangin B. Package “LDcorSV.” 2017;
33. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci.* 2001;98: 11479–11484. doi:10.1073/pnas.201394398
34. Shin J-H, Blay S, McNeney B, Graham J. LDheatmap: an R function for graphical display. *J Stat Softw.* 2006;16: Code Snippet 3. doi:10.18637/jss.v016.c03
35. Adetunji I, Willems G, Tschoep H, Bürkholz A, Barnes S, Boer M, et al. Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. *Theor Appl Genet.* 2014;127: 559–571. doi:10.1007/s00122-013-2239-x
36. Filippi C V, Aguirre N, Rivas JG, Zubrzycki J, Puebla A, Cordes D, et al. Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. *BMC Plant Biol.* 2015;15: 52. doi:10.1186/s12870-014-0360-x
37. Wysmierski PT, Vello NA. The genetic base of Brazilian soybean cultivars: Evolution over time and breeding implications. *Genet Mol Biol.* 2013;36: 547–555. doi:10.1590/S1415-47572013005000041
38. Hiromoto DM, Vello NA. The genetic bae of brazilian soybean(*Glycine max*(L.) MERRILL) cultivars. *Genet Brazilian J.* 1986; 295–306.
39. Hoffstetter A, Cabrera A, Huang M, Sneller C. Optimizing Training Population Data and Validation of Genomic Selection for Economic Traits in Soft Winter Wheat. *G3 Genes|Genomes|Genetics.* 2016;6: 2919–2928. doi:10.1534/g3.116.032532
40. Hwang E-Y, Song Q, Jia G, Specht JE, Hyten DL, Costa J, et al. A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics.* 2014;15: 1. doi:10.1186/1471-2164-15-1
41. Brown-Guedira GL, Thompson JA, Nelson RL, Warburton ML. Evaluation of Genetic Diversity of Soybean Introductions and North American Ancestors Using RAPD and SSR Markers. *Crop Sci.* 2000;40: 815. doi:10.2135/cropsci2000.403815x
42. Bouchet S, Servin B, Bertin P, Madur D, Combes V, Dumas F, et al. Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus. *PLoS One.* 2013;8. doi:10.1371/journal.pone.0071377

43. Rafalski A, Morgante M. Corn and humans: Recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet.* 2004;20: 103–111. doi:10.1016/j.tig.2003.12.002
44. Simko I, Haynes KG, Jones RW. Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. *Genetics.* 2006;173: 2237–2245. doi:10.1534/genetics.106.060905
45. Nordborg M, Tavaré S. Linkage disequilibrium: What history has to tell us. *Trends Genet.* 2002;18: 83–90. doi:10.1016/S0168-9525(02)02557-X
46. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol.* 2015;33: 408–14. doi:10.1038/nbt.3096
47. Zhang J, Song Q, Cregan PB, Jiang G-L. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*) [Internet]. *Theoretical and Applied Genetics.* 2015. doi:10.1007/s00122-015-2614-x
48. Mamidi S, Chikara S, Goos RJ, Hyten DL, Annam D, Moghaddam SM, et al. Genome-Wide Association Analysis Identifies Candidate Genes Associated with Iron Deficiency Chlorosis in Soybean. *Plant Genome J.* 2011;4: 154. doi:10.3835/plantgenome2011.04.0011
49. Hagenblad J, Nordborg M. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. *Genetics.* 2002;161: 289–298.
50. Vos PG, Paulo MJ, Voorrips RE, Visser RGF, van Eck HJ, van Eeuwijk FA. Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor Appl Genet.* Springer Berlin Heidelberg; 2017;130: 123–135. doi:10.1007/s00122-016-2798-8
51. Mezouk S, Dubreuil P, Bosio M, Décousset L, Charcosset A, Praud S, et al. Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. *Theor Appl Genet.* 2011;122: 1149–1160. doi:10.1007/s00122-010-1519-y
52. Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc Natl Acad Sci U S A.* 2001;98: 9161–6. doi:10.1073/pnas.151244298
53. Jung M, Ching A, Bhattaramakki D, Dolan M, Tingey S, Morgante M, et al. Linkage disequilibrium and sequence diversity in a 500-kbp region around the *adh1* locus in elite maize germplasm. *Theor Appl Genet.* 2004;109: 681–689. doi:10.1007/s00122-004-1695-8

54. Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler ES. Genetic diversity and selection in the maize starch pathway. *Proc Natl Acad Sci.* 2002;99: 12959–12962. doi:10.1073/pnas.202476999
55. Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van de Peer Y. How many genes are there in plants (... and why are they there)? *Curr Opin Plant Biol.* 2007;10: 199–203. doi:10.1016/j.pbi.2007.01.004
56. Cannon SB, Shoemaker RC. Evolutionary and comparative analyses of the soybean genome. *Breed Sci.* 2012;61: 437–444. doi:10.1270/jsbbs.61.437
57. Freeling M. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annu Rev Plant Biol.* 2009;60: 433–453. doi:10.1146/annurev.arplant.043008.092122

FIGURAS

Figura 1. Densidade dos marcadores baseada em 12.133 SNP distribuídos em 20 cromossomos, considerando 1Mb como tamanho de janela. Os 20 cromossomos da soja estão representados no eixo vertical; o eixo horizontal mostra o comprimento do cromossomo (Mb).

Figura 2. Estrutura da população de soja com base no software Structure. O eixo vertical mostra a probabilidade de cada genótipo pertencer a cada subpopulação. A sub-população 1 e 2 são representadas como azul e vermelho, respectivamente. O eixo horizontal representa os nomes dos genótipos.

Figura 3. Estrutura da população do painel de diversidade de soja revelada pela DAPC. O eixo vertical mostra a probabilidade de cada genótipo pertencer a cada uma das três subpopulações. A subpopulação 1, 2 e 3 são representadas como azul, vermelho e verde, respectivamente. A coordenada do eixo horizontal representa os nomes dos genótipos.

Figura 4. Distribuição dos 95 genótipos de soja baseado em PCA bidimensional. Os genótipos de soja provenientes da África, América do Norte, América do Sul, Ásia e da Europa estão representados em azul, cinza, vermelho, verde e azul claro, respectivamente.

Figura 5. Heatmap e dendrograma baseados na estimação do “*simple matching coefficient*” com base em 12.133 SNPs genotipados em 95 genótipos de soja. O histograma na parte superior mostra a distribuição dos valores na matriz de parentesco.

Figura 6. Heatmaps de desequilíbrio de ligação no cromossomo 17 da soja considerando: r^2 , DL sem qualquer correção; r^2v , DL corrigido pela matriz de parentesco; r^2s_dapc , r^2s_struc , r^2s_pca , DL corrigido pela a estrutura da população fornecida por: DAPC; Software STRUCTURE; e abordagem PCA; r^2vs , DL corrigido pelo parentesco entre os genótipos, e também por cada uma das abordagens de estrutura populacional: r^2vs_dapc ; r^2vs_pca ; r^2vs_struc .

Figura 7. Decaimento do desequilíbrio de ligação no genoma determinado pelas medidas de DL contra a distância entre SNPs: r^2 , sem qualquer correção (vermelho); r^2v , medida corrigida para o parentesco (azul). r^2s_dapc , r^2s_struc , r^2s_pca , medida corrigida para a estrutura da população estimada pelas abordagens: DAPC (verde claro); Software STRUCTURE (laranja); e PCA (roxo). Medidas corrigidas para o parentesco e também cada uma de uma abordagem de estrutura populacional: r^2vs_dapc (verde escuro); r^2vs_pca (rosa); r^2vs_struc (vermelho escuro).

TABELAS

Tabela 1. QTLs relatados no Soybase *browser* co-localizados nas regiões contendo desequilíbrio de ligação inter-cromossômico estimado pela medida de $r^2_{sv_dapc}$. I e II = regiões que apresentaram desequilíbrio de ligação inter-cromossômico, pos (cM) = posição do intervalo onde o DL inter-cromossômico foi detetado. ^aBaseado na lista QTL apresentado no SoyBase (<https://www.soybase.org/>).

(continua)

Nº	I Crom	I pos (cM)	II Crom	II pos (cM)	QTLs reportados em I ^a	QTLs reportados em II ^a
1	1	28,16-32,80	17	74,22-74,24	Seed oil 23-2 Seed weight 18-1.1	Seed oil 39-7 Seed weight 43-2 Seed weight 49-10 Seed weight 49-10
2	1	57,08-63,38	17	74,22-74,24	Seed weight 18-1.1 Seed weight 15-2 Seed weight 34-11 Seed weight 37-5 Seed weight 36-17 Seed oil 39-5	Seed weight 43-2 Seed weight 49-10 Seed oil 39-7
3	1	95,02-95,36	5	70,44-70,91	Seed weight 18-1.1 Plant height 6-8 Plant height 42-1	Seed weight 34-9 Plant height 26-1 Plant height 37-1 Plant height 24-1 Plant height 24-2
4	2	7,08	5	49,06	Seed isoflavone 6-2 Seed daidzein 7-3 Seed daidzein 13-10 Canopy wilt 3-4 Canopy wilt 3-8 Canopy wilt 3-11 Seed genistein 7-3 Seed weight 42-1	Seed isoflavone 11-15 Seed daidzein 13-12 Canopy wilt 3-5 Seed weight 36-9 Seed weight 34-9
5	3	0,05-0,62	5	87,15	Seed oil 43-32	Seed oil 30-1 Seed oil 8-1
6	3	17,44	7	2,59		
7	3	20,1	6	85,25		
8	3	20,45	11	97,62	Seed oil 43-32	Seed oil 24-14
9	4	58,1	7	38,48	Pod maturity 35-1 Pod maturity 35-1 Pod maturity 17-3 Pod maturity 18-1 Seed weight 20-3 Seed weight 11-1 Plant height 20-2 Plant height 20-2	Pod maturity 18-4 Pod maturity 15-1 Pod maturity 3-1 Pod maturity 4-2 Pod maturity 7-1 Seed weight 49-15 Plant height 2-1 Plant height 3-3
10	4	58,1	10	108,72	Flood tolerance 4-4 Pod maturity 35-1 Pod maturity 18-1 Seed linoleic 1-5 Seed weight 11-1 Seed weight 20-3 Pod maturity 18-1 Plant height 20-2 First flower 11-2	Flood tolerance 4-11 Pod maturity 26-3 Seed weight 34-8 Pod maturity 26-3 Plant height 31-2 First flower 24-4 Seed weight 34-8 Seed linolenic 9-4 First flower 24-4

(continuação)

Nº	I Crom	I pos (cM)	II Crom	II pos (cM)	QTLs reportados em I ^a	QTLs reportados em II ^a
11	4	58,1	13	65,93	Seed weight 11-1 Seed weight 20-3 Plant height 20-2 Seed isoflavone 11-16	Seed weight 15-3 Seed weight 49-14 Plant weight 44-2 Plant height 40-2 Seed isoflavone 9-11
12	7	38,48	13	65,93	Seed abortion 1-4 Seed isoflavone 1-7 Seed isoflavone 11-4 Seed total isoflavone 9-9 Seed genistein 3-4 Seed genistein 11-9 Seed genistein 13-2 Seed genistein 2-3 Plant height 19-5 Plant height 2-1 Plant height 3-3 Plant height 6-5	Seed abortion 1-3 Seed total isoflavone 9-11 Seed genistein 11-8 Seed genistein 13-6 Plant height 5-6
13	7	38,48	10	108,72	Pod maturity 3-1 Pod maturity 4-2 Pod maturity 7-1 Plant height 3-3 Plant height 6-5	Pod maturity 26-3 Plant height 31-2
14	9	6,58	11	84,69	Seed oil 1-2 Seed oil 43-21	Seed oil 24-14
15	8	54,02	13	86,25	Seed protein 34-5 Seed protein 34-4 Seed isoflavone 11-5 Seed isoflavone 7-1 Seed daidzein 7-2 Seed daidzein 8-4 Seed glycitein 8-9	Seed protein 36-20 Seed protein 36-18 Seed totalisoflavone 9-11 Seed daidzein 10-3 Seed daidzein 11-4 Seed daidzein 13-5 Seed glycitein 9-6
16	10	108,72	13	65,93	Plant height 31-2 Seed oil 39-16	Plant height 5-6 Plant height 40-2 Seed oil 37-7
17	12	20,29	19	83	Seed weight 43-3	Seed weight 43-4
18	12	20,29	6	86,85	Hypocotyl weight 1-2	Hypocotyl weight 1-3
19	12	52,78	5	27,51	Seed weight 23-2 Seed weight 44-1	Seed weight 36-9 Seed weight 34-9 Seed weight 6-4 Seed weight 37-12 Seed weight 36-13
20	12	52,78	5	40,00	Seed daidzein 7-4 Seed genistein 6-1	Seed daidzein 13-12 Seed genistein 2-4 Seed genistein 13-10
21	14	4,22	4	58,1	Seed weight 29-1 Seed weight 29-4	Seed weight 47-3
22	14	4,22-4,58	7	38,48	Seed weight 29-1 Seed weight 29-4	Seed weight 49-15
23	14	4,22-4,58	10	108,72	Seed weight 29-1 Seed weight 29-4	Seed weight 34-8
24	14	4,22-4,58	13	65,93	Seed weight 29-1 Seed weight 29-4	Seed weight 15-3 Seed weight 49-14
25	14	8,93	9	6,58-6,60		

(conclusão)

Nº	<i>I</i> Crom	<i>I</i> pos (cM)	<i>II</i> Crom	<i>II</i> pos (cM)	QTLs reportados em <i>I</i> ^a	QTLs reportados em <i>II</i> ^a
26	14	8,93	11	84,69	Seed weight 29-1 Seed weight 29-4	Seed weight 32-1 Seed weight 36-11 Seed weight 10-3 Seed weight 35-9
27	14	37,42-41,79	7	39,33-39,42	Seed protein 4-10	Seed protein 7-1
28	14	84,83	17	42,77-56,77	Seed yield 23-10 Seed yield 32-3 Seed yield 31-1	Seed yield 21-6 Seed yield 5-2
29	14	84,83	9	74,58	Seed oil 34-2 Seed oil 43-3 Seed oil 43-2 Seed daidzein 4-1 Seed protein 28-4	Seed oil 43-22 Seed daidzein 10-2 Seed protein 36-28 Seed protein 33-3
30	15	0,06-0,22	5	12,59	Seed linolenic 1-1	Seed linolenic 7-5
31	15	4,51	10	23,31	Seed protein 31-2	Seed protein 21-5
32	15	38,89-73,22	10	49,65-53,87	SCN 44-10 SCN 21-3 SCN 29-9 SCN 38-1 SCN 44-14 SCN 25-1 SCN 18-4 Seed oil 36-3 Seed oil 39-8 Seed oil 23-3 Seed protein 4-13 Seed protein 27-2 Seed protein 27-3 Pod maturity 31-2 Pod maturity 37-2 Pod maturity 27-7 Pod maturity 29-7	SCN 38-4 SCN 34-5 SCN 38-5 Seed oil 43-33 Seed oil 19-3 Seed oil 43-35 Seed protein 34-2 Seed protein 42-1 Seed protein 36-40 Pod maturity 24-9
33	16	72,00	20	7,44		
34	17	22,21-31,86	8	4,17		
35	17	35,02	19	24,75	Pod number 11-1 Seed oil 37-1	Pod number 1-9 Seed oil 43-27
36	17	42,77-56,36	9	74,58	Seed weight 47-2 Seed weight 13-5 Seed weight 49-10 Seed weight 18-2 Seed weight 47-2 Seed oil 23-3 Seed palmitic 4-3	Seed weight 27-3 Seed weight 21-3 Seed weight 3-4 Seed weight 21-3 Seed oil 43-22 Seed palmitic 3-2
37	19	83,00	6	86,85	Seed daidzein 10-4 Seed genistein 10-6	Seed daidzein 11-1 Seed daidzein 13-2 Seed genistein 11-1

3. ARQUITETURA GENÉTICA DE MÚLTIPLOS CARACTERES DE SOJA VIA ANÁLISE DE ASSOCIAÇÃO GENÔMICA AMPLA

RESUMO

A soja é considerada uma das culturas de maior importância mundial por ser o vegetal com maior porcentagem de óleo e proteína em sua composição. Devido sua grande importância, existem muitos esforços para obtenção de cultivares mais produtivos e resistentes a condições adversas. Atualmente, com os avanços nas plataformas de genotipagem, o estudo de associação ampla no genoma (GWAS) é considerado uma ferramenta promissora, possibilitando a obtenção de informações da arquitetura genética de caracteres complexos da soja. Desta forma, o objetivo deste estudo foi utilizar a abordagem GWAS para dissecar a arquitetura genética de dez importantes caracteres em soja. Para isso, foi avaliado em seis ambientes um painel de 95 genótipos de soja, provenientes de diversas localidades do mundo. Utilizou-se a plataforma de genotipagem 180 K Axiom® Soybean Genotyping Array, na qual após o controle de qualidade e imputação de dados resultaram em 12133 marcadores SNPs. Devido à presença da interação genótipo ambiente, as abordagens GWAS foram realizadas por ambiente. No total foram identificados 181 marcadores associados aos dez caracteres avaliados. Destes, 43 SNPs foram associados a valor agrônômico (VA), 12 à massa de cem sementes (MCS), 12 ao acamamento (AC), 11 ao período de enchimento de grãos (PEG), 28 ao número de dias para a maturação (NDM), 13 à altura da inserção da primeira vagem (AIPV), 16 à altura da planta na maturação (APM), quatro à altura da planta no florescimento (APF), 33 associados à produtividade de grãos (PG) e nove ao teor de óleo (TO). A presença de regiões genômicas distintas associadas para um mesmo caráter, em ambientes diferentes, evidencia a complexidade da regulação dos caracteres e da interação QTL e ambiente. Algumas regiões identificadas associadas aos caracteres estudados já foram previamente reportadas por outros estudos, entretanto, a maioria das regiões encontradas são inéditas, sendo candidatas potenciais a serem confirmados seus envolvimento no controle destes caracteres em estudos futuros.

Palavras-chave: GWAS; Múltiplos caracteres; *Glicine max*; QTLs

ABSTRACT

Soybean is considered one of the world's most important crops because it is the vegetable with the highest percentage of oil and protein in its composition. Due to its great importance, there is several efforts for obtaining cultivars more productive and resistant to adverse conditions. Nowadays, with advances in the genotyping platforms, genome-wide association studies (GWAS) are considered a promising tool, enabling obtaining information about the genetic architecture of complex characters in soybean. Therefore, the objective of this study was to use the GWAS approach to dissect the genetic architecture of ten important traits in soybean. We evaluated in six environments a panel of 95 soybean genotypes from diverse localities in the world. We used the 180 K Axiom® Soybean Genotyping Array genotyping platform, which resulted in 12,133 SNP markers after quality control and data imputation. Due to the presence of the genotype-environment interaction, the GWAS approached were performed across environments. In total, 181 markers associated to the ten evaluated traits were identified. Of these, 43 SNP were associated with agronomic value, 12 with mass of hundred seeds, 12 with lodging, 11 with grain filling period, 28 with number of days to maturity, 13 with height of the first pod, four with plant height at flowering, 33 with grain yield and nine were associated with oil content. The presence of distinct genomic regions associated with the same trait, in different environments, shows the complexity of the traits' regulation and the presence of the interaction QTL and environment (QxE). Some regions associated with the evaluated traits were previously reported by other

studies, however, most of them are novel, which are potential candidates to be confirmed their involvement in the control of their traits in future studies.

Keywords: GWAS; Multiple-traits; *Glycine max*; QTLs

3.1. Introdução

A soja cultivada (*Glycine max*) é a oleaginosa de maior importância mundial, sendo a principal fonte vegetal de óleo e proteína [1,2]. No cenário mundial, a soja é responsável por 68% da produção de proteína vegetal e 57% da produção de óleo nas últimas décadas [3]. Em média, o grão de soja apresenta 40,3% de proteínas, 21% de lipídeos, 33,8% de carboidratos e 4,9% de minerais [4]. Por apresentar tal composição química, a soja dispõe de um amplo leque de utilidades, abrangendo, principalmente, a alimentação humana e produção de ração animal, sendo matéria prima para agroindústrias e, mais recentemente, utilizada na produção de bicomcombustível [5].

Em 2050, estima-se que a população alcançará mais de 9 bilhões de habitantes, indicando que os recursos alimentícios, provenientes basicamente da agricultura e pecuária, devem dobrar para garantir a alimentação mundial [6–8]. Atualmente, relata-se que 800 milhões de pessoas sofrem de deficiência alimentar [3], evidenciando a problemática que enfrentamos com a produção de alimentos no mundo. Entretanto, devido à restrição de áreas de cultivo, mudanças climáticas e escassez da disponibilidade de água, o aumento da produção dos alimentos apresenta-se como um grande desafio.

Desenvolver novos cultivares de soja, que possam produzir bem com menos insumos, auxiliando em uma agricultura mais sustentável, e sob condições climáticas mais estressantes e imprevisíveis é essencial para o futuro da segurança alimentar, e representa o maior desafio para os melhoristas [9]. Para acelerar e auxiliar neste processo, técnicas modernas de melhoramento deve ser exploradas. Neste sentido, destaca-se o mapeamento associativo como ferramenta em potencial.

O objetivo do mapeamento é identificar marcadores moleculares que estejam intimamente ligados, devido sua localização próxima ou dentro de genes [10]. Duas abordagens podem ser utilizadas: mapeamento convencional, também chamado de mapeamento de ligação e o mapeamento associativo, conhecido também como *Genome-Wide Association Studies* (GWAS) ou mapeamento por desequilíbrio de ligação. O mapeamento por ligação foi considerado por tempos a abordagem mais comum. Entretanto, devido o estudo de associação genômica ampla não necessitar da formação de populações próprias de

mapeamento como, F_{1S} , F_{2S} , retrocruzamentos, duplo-haplóides, entre outras [11], essa abordagem vem ganhando destaque.

Os estudos de associação genômica ampla podem ser considerados uma ferramenta alternativa para superar as limitações do mapeamento por ligação [12]. Em suma, esta abordagem busca a identificação de correlações significativas entre um loco marcador e o fenótipo do caráter de interesse [12]. Consiste em uma técnica baseada no desequilíbrio de ligação, na qual requer o conhecimento prévio acerca da diversidade, estruturação e desequilíbrio de ligação da população a ser utilizada, a fim de minimizar ocorrências de falso positivos na análise [13]. A abordagem ainda oferece a possibilidade de explorar a variabilidade genética encontrada no germoplasma, proporcionando a redução de custos e tempo para estabelecer populações de mapeamento. E ainda, permite a avaliação de uma infinidade de alelos para as mais diversas características, explorando todo o histórico de recombinação relativo ao tempo de desenvolvimento do germoplasma. Portanto, ao contrário do mapeamento por ligação, a GWAS explora todo histórico de recombinação presente na espécie [12,14], obtendo, assim, maior resolução na análise e, portanto, melhores resultados [15,16].

As análises GWAS, além de possibilitar o acesso a informações a respeito da arquitetura genética de caracteres complexos, as regiões identificadas associadas a esses caracteres possuem potencial aplicação em programas de melhoramento. Os resultados de tais análises, utilizados como cofatores, podem auxiliar no aumento da acurácia das predições em estudo de seleção genômica ampla [17,18] e, portanto, aumentar as taxas de sucesso nas etapas de seleção.

Diante do exposto, nosso principal objetivo foi obter informações a respeito da arquitetura de dez caracteres de importância agrônômica em soja, utilizando a abordagem GWAS em múltiplos ambientes em um painel de 95 genótipos utilizando a plataforma de genotipagem de *180 K Axiom® Soybean Genotyping Array*.

3.2. Material e Métodos

3.2.1. Material vegetal

Para o estudo de GWAS e, a fim de reunir materiais com grande divergência genética, foi elaborado um painel de noventa e cinco genótipos de soja com procedência de diversas regiões geográficas do mundo. Para obter mais informações a respeito dos genótipos

utilizados neste estudo, consultar Mulato et al. [19] e Marconato et al. [20]. Os autores citados utilizaram os mesmos genótipos para elucidar a diversidade genética através de marcadores SSR e caracteres agro-morfológicos, respectivamente. Os 95 genótipos foram avaliados em delineamento experimental alfa-látice 5x19 com 3 repetições e parcelas de 4 linhas de 5 metros cada, espaçamento entre linhas de 0,5 metros, em que apenas as duas linhas centrais das parcelas foram colhidas. No total, os genótipos foram avaliados em seis ambientes: Ponta Grossa (Paraná, 25° 05' S, 50° 09' W) na safra de 2012/2013 (denominado de ambiente 1), Jaboticabal-SP (São Paulo, 21° 15' S, 48° 19' W) nas safras de 2012/2013 e 2013/2014 (denominados de ambientes 2 e 3) e Piracicaba (São Paulo, 22°42' S, 47°38' W) nas safras de 2012/2013 e 2013/2014 (denominados de ambientes 4 e 5) e por fim em Anhembi-SP (São Paulo, 22°50' S, 48°01' W) na safra de 2014/2015 (denominado de ambiente 6).

3.2.2. Genotipagem e controle de qualidade dos marcadores

Para proceder a genotipagem, e, portanto, a obtenção dos marcadores SNPs, os genótipos foram cultivados em casa de vegetação e, posteriormente, coletaram-se uma única folha por planta. As folhas coletadas foram maceradas e procedeu-se a extração de DNA pelo protocolo CTAB [21]. Para a quantificação do DNA das amostras utilizou-se o aparelho Quantifluor®, sendo que esse foi diluído para uma concentração de 100 ng μ L⁻¹ e distribuídos em uma placa, na qual foi enviada para genotipagem pela plataforma da *Affymetrix (180 K Axiom® Soybean Genotyping Array)*.

Ao fim da genotipagem, os SNPs foram filtrados seguindo as recomendações do *Software Axiom® Analysis Suite*: $DQC \geq 0,82$ (*Dish quality control*, medida da resolução da distribuição dos valores de contraste), $QC \text{ call rate (Quality control call rate)} \geq 92$, $Average \text{ call rate for passing} \geq 97$, $Minor \text{ alele cut off} \geq 2$ e marcadores *Poly High Resolution*. Além disso, SNPs com MAF (*Minimum Minor Allele Frequency*) menor que 0,005 e mais de 10% de dados perdidos foram eliminados. Ao final destas etapas de controle de qualidade, os dados resultantes foram submetidos a imputação de dados faltantes no software BEAGLE 4.1. [22]. Todos os marcadores heterozigotos também foram eliminados. Para as análises foram utilizados 12133 marcadores SNPs resultantes das etapas de controle de qualidade e imputação de dados.

3.2.3. Fenotipagem e análises estatísticas

Para a análise GWAS dez caracteres foram avaliados, sendo eles: NDM: número de dias para a maturidade, contados desde a semeadura até a data em que 95% das vagens apresentaram-se maduras; APF: altura da planta (cm) no florescimento; APM: altura da planta (cm) na maturidade, medida desde a base da planta no solo até o ápice da haste principal; AC: acamamento, avaliado na maturidade por uma escala de notas visuais de 1 a 5, sendo a nota 1 correspondente à planta ereta e a 5 à planta totalmente acamada; VA: valor agronômico, avaliado na maturidade através de uma escala de notas visuais de 1 a 5, sendo a nota 1 correspondente à planta sem nenhum valor agronômico e a nota 5 à planta com excelentes características agronômicas (grande número de vagens, altura superior a 60 cm, vigor, sem acamamento, ausência de hastes verdes e retenção foliar, sem debulha de vagens e sem sintomas de doenças); PG: produtividade de grãos, avaliada pelo peso dos grãos (g) da parcela; MCS: Massa de cem sementes, em gramas; TO: teor de óleo, medido por espectroscopia infravermelho, utilizando curva de calibração e equipamento fornecidos pela PERTEN (DA7200); PEG: período de enchimento de granação, número de dias compreendidos entre os estágios R5 e R7; AIPV: altura da inserção da primeira vagem, em centímetros.

A análise de variância (ANOVA) foi realizada para cada caráter utilizando o software R [23], através da função *lm*. A combinação de locais x anos foi considerada como ambientes, resultando assim em seis ambientes. Destaca-se que nem todos os caracteres foram avaliados em todos os ambientes, o que pode ser visualizado na Tabela 1. A estimação da herdabilidade foi realizada de acordo com o proposto por Knapp et al. [24] para isso utilizou-se o pacote *lmer4*[25], na qual todos os fatores do modelo foram considerados como aleatórios para obtenção das variâncias. Foram calculadas as correlações de Pearson, entre os valores genéticos preditos dos caracteres estudados, no pacote *Performance Analytics* disponível no programa R [23], de forma a permitir uma melhor compreensão dos resultados do mapeamento associativo de múltiplos caracteres.

3.2.4. Análise de associação genômica-ampla (GWAS)

Na GWAS foram calculados os valores de BLUP utilizando o pacote do R *lme4*[25]. A combinação de anos e locais para cada caráter avaliado foi considerada como um ambiente. Assim, além do cálculo dos valores de BLUP considerando todos os ambientes, esses também

foram mensurados para cada ambiente em particular, a fim de verificar a estabilidade das associações encontradas. A GWAS foi conduzida com auxílio do pacote computacional GAPIT[26]do software R[23]. O modelo mixed linear model (MLM) foi utilizado, na qual as matrizes de estrutura(Q) e parentesco(K) são empregadas para a correção e redução do número de falsos positivos, devido aos vieses causados pelo parentesco e estruturação da população [26–28]. Foi utilizado de um *threshold* de $-\log_{10}(p) > 3$ como já reportado em outros trabalhos publicados para a soja [27,29–31]. A variância explicada para cada SNP considerado significativo foi calculada pelo quadrado da correlação entre o fenótipo e genótipo [32].

Para inferir a estrutura populacional e, portanto, obtenção da matriz Q , foi utilizada a análise Bayesiana implementada no software STRUCTURE 3.2 [33]. Foram utilizadas 10 corridas independentes com K variando entre 2 e 10, 500.000 burn-ins, 1.000.000 simulações de Monte Carlo Cadeias de Markov (MCMC), modelo de não-mistura (no admixture) e frequências alélicas correlacionadas. Assim, foram gerados 80 datas-set, na qual cada k , variando de 2 a 10, foi testado em 10 corridas. Sendo que a definição do número de sub-populações mais provável foi realizada por meio dos valores de ΔK [34]. A matriz de parentesco (k) foi inferida pelo pacote *synbreed* [35]do R [23], e utilizando os SNPs já filtrados e imputados.

3.3. Resultados e discussão

3.3.1. Análise fenotípica

Para análise fenotípica nem todos os caracteres foram avaliados em todos os ambientes, o que é evidenciado pela ausência de alguns ambientes para alguns caracteres na Tabela 1. Entretanto, considerando os ambientes no quais cada caráter foi avaliado, estes apresentaram variações, como pode ser verificado pela média apresentada em cada local avaliado (Tabela 1), bem como pelo efeito significativo de genótipos apresentado pela ANOVA (Tabela 1). Destaca-se o ambiente 6, na qual a produtividade (PG) teve uma média bem restrita quando comparado aos demais ambientes. O mesmo é elucidado quando comparamos os valores de máximos e mínimos, na qual no ambiente 6 o valor máximo de produtividade de grãos foi reduzido, demonstrando que o ambiente não apresentou condições favoráveis para o cultivo dos genótipos analisados. Além disso, neste ambiente o coeficiente de variação (CV) foi encontrado como o mais elevado, de 53,33% para produtividade (PG).

Em relação aos demais CV dos experimentos, estes variaram de 41,68 % a 3,75%. A maioria dos ambientes apresentou CV abaixo de 20%, demonstrando boa precisão experimental [24], com exceção da produtividade, na qual esse foi mais elevado em todos os ambientes avaliados (Tabela 1). O alto CV apresentado pela produtividade de grãos (PG) pode ser resultante da complexidade do caráter, sendo altamente influenciado pelo ambiente. Além disso, pelo estudo utilizar genótipos exóticos não adaptados, o alto CV para alguns caracteres já era esperado. Em relação à ANOVA, esta foi conduzida para os dez caracteres avaliados, na qual a combinação de anos e locais foi considerada como ambientes. Pelos resultados apresentados para todos os caracteres avaliados observamos a presença de efeito genotípico significativo ($p < 0,01$) na população avaliada (Tabela 1). Sendo que o mesmo foi verificado para o efeito de ambientes, ou seja, a variação ambiental foi significativa ($p < 0,01$) para os dez caracteres. Além disso, a interação genótipos e ambiente (GxE) foi observada como altamente significativa para todos os caracteres (Tabela 1). Em relação à herdabilidade, essa variou de 88% para altura de planta na maturidade (APM) a 25% para período de granação (PEG) nesta população. O teor de óleo (TO) e produtividade de grãos (PG) apresentaram alta herdabilidade com 82% e 62%, respectivamente, estando de acordo com o trabalho de Wang et al.[36]. Em uma população de mapeamento por ligação contendo 85 RILs (*Recombinant Inbred Lines*) Wang et al.[36] verificaram herdabilidade média para produtividade de 65% e 94% para teor de óleo.

As correlações fenotípicas estimadas entre os 10 caracteres estudados podem ser visualizadas na Figura 1. Essa figura representa um correlograma, na qual a diagonal é composta por histogramas dos valores de cada variável. Na parte de cima da diagonal é representada as correlações entre as variáveis e na parte de baixo a distribuição dos valores. Pela Figura 1, observa-se que as correlações apresentaram significância variadas com p-valores de 0,001, 0,01, 0,05, 0,1 (“***”, “**”, “*”, “.”) pelo Teste t. Todavia, apenas as correlações com significância abaixo de 5% serão consideradas como relevantes para este estudo. Desta forma, foram observadas 38 correlações significativas variando de 0,29 a 0,92 para as positivas e de -0,73 a -0,19 para as negativas (Figura 1).

Destaca-se a correlação alta encontrada entre valor agronômico (VA) e produtividade de grãos (PG) (0,66***), bem como dos demais caracteres que influenciam na produtividade da cultura: teor de óleo (TO) (0,63***), massa de cem sementes (MSC) (0,59***), período de granação (PEG) (0,44***). Correlações negativas elevadas foram verificadas entre acamamento (AC) e valor agronômico (-0,73***); e entre acamamento (AC) e teor de óleo (TO) (-0,68***). Também foi verificada a correlação negativa do caráter acamamento com

massa de cem sementes (-0,68***). A correlação negativa alta entre acamamento e esses três caracteres, bem como com produtividade (-0,43***) evidencia as consequências do acamamento das plantas gerando perdas na colheita e, portanto, redução da produtividade. Plantas altas tanto no florescimento, quanto na maturidade apresentam maior probabilidade de acamamento, isto fica evidente pela presença da correlação alta entre acamamento e altura da planta no florescimento e maturidade (0,70*** e 0,63***, respectivamente).

Teor de óleo apresentou correlação significativa com todos os demais caracteres avaliados, indicando a complexidade neste caráter. Altura da inserção da primeira vagem (AIPV) apresentou alta e significava correlação com altura da planta na maturidade (APM) (0,73***) e no florescimento (APF) (0,73***), indicando que plantas com inserção da primeira vagem próximas do solo resultaram em plantas com tamanho reduzido. Nenhuma relação foi evidenciada entre altura da inserção da primeira vagem e produtividade, entretanto Malik et al. [4] verificaram correlação média (0,44**) entre esses caracteres em uma população de 27 genótipos de soja.

A correlação entre caracteres fenotípicos é um parâmetro populacional, portanto, populações compostas por genótipos distintos poderão apresentar correlações distintas e muitas vezes opostas. Isto é verificado quando comparamos a correlação apresentada por Silva et al. [37] entre teor de óleo e produtividade. Os autores observaram correlação negativa entre os caracteres. Contudo, em nosso estudo foi verificado o oposto, teor de óleo e produtividade de grãos apresentaram correlação média (0,50***), indicando que os genótipos desta população com maior produtividade de grãos apresentaram maior teor de óleo. A estimação de parâmetros de correlação entre caracteres em quaisquer populações auxilia na definição das estratégias de seleção indireta, bem como na melhor interpretação dos resultados de mapeamento de múltiplos caracteres.

3.3.2. Estrutura da população e matriz de parentesco

A fim de reduzir o número de associações espúrias na análise GWAS [28,38,39] foi inferida a estrutura populacional e o parentesco dos 95 genótipos de soja com 12133 SNPs. Os resultados da análise de estruturação populacional via Structure indicou que o número mais provável de subpopulações é $k=2$ (Figura S1). De forma geral, os genótipos de diferentes regiões geográficas (países) foram distribuídos nas duas subpopulações (cor azul e vermelha na Fig. S1), demonstrando uma estruturação complexa. Além disso, 54 genótipos foram classificados “*admixed*” e a subpopulação de cor vermelha apresentou predominância de

genótipos originários das Américas do Norte e do Sul e da Ásia. O cálculo do “*simple matching coefficient*” [40] foi utilizado para inferir a matriz de parentesco (K). Os genótipos do estudo revelaram valores médios de similaridade genética entre a comparação dois a dois. Este resultado corrobora a idéia de que genótipos de regiões distintas gerariam um painel diverso e, portanto, com pouca ou média similaridade entre genótipos. Os resultados do *software* Structure e da matriz de parentesco demonstraram que os genótipos não foram agrupados de acordo com suas respectivas localidades geográficas, e apresentaram grupos de origem mista, portanto, grupos complexos. Grupos complexos podem ser resultantes da ausência de *pools* genéticos bem definidos, podendo ser consequência da base genética estreita da espécie, e, propriamente, da escolha dos parentais em programas de melhoramento [41,42]. De modo similar, Brown-Guedira et al. [43] não verificaram a presença de sub-populações com correspondência geográfica utilizando-se marcadores SSR em uma população de 105 genótipos. Além disso, Hwang et al.[30] relataram que apesar da complexidade da estrutura populacional verificada em seu estudo, foi possível desenvolver com sucesso a análise GWAS em soja.

3.3.3. Mapeamento associativo

A fim de investigar se regiões associadas no controle de caracteres agronômicos importantes para soja realizou-se o estudo de GWAS. As análises de GWAS foram realizadas para os dez caracteres avaliados neste estudo. A interação QTLs e ambientes é um importante fator a ser considerado para o estudo de caracteres quantitativos [44–46]. Desta forma, os caracteres agronômicos foram avaliados em diferentes anos e locais, na qual a combinação destes resultaram em 6 ambientes, como já mencionado na Tabela 1. Afim de verificar o efeito de ambientes sobre a estabilidade dos SNPs nas análises GWAS, foram calculados os valores de BLUPs dos genótipos para cada caráter em cada um dos ambientes que esses foram avaliados. Justifica-se a abordagem de GWAS por ambientes devido a significativa presença de interação GxE, apresentada para todos os caracteres (Tabela 1). Além disso, através da análise conjunta, também foram calculados os valores do BLUPs considerando todos os ambientes.

Utilizou-se o modelo MLM, na qual foram observadas associações significativas ($-\log P > 3$) para todos os caracteres em pelo menos um dos ambientes avaliados. Associações foram encontradas localizadas em 19 cromossomos, com exceção dos cromossomos 3 (Tabela

2). No total foram identificados 181 marcadores associados aos dez caracteres avaliados. Sendo que, 43 SNPs foram associados ao valor agrônômico (VA), 12 ao peso de cem sementes (PCS), 12 ao acamamento (AC), 11 ao período de granação (PEG), 28 ao número de dias para a maturidade (NDM), 13 à altura da inserção da primeira vagem (AIPV), 16 à altura da planta na maturidade (APM), 4 à altura da planta no florescimento (APF), 33 associados à produtividade de grãos (PG) e 9 ao teor de óleo (TO). Neste estudo, SNPs associados a um caráter, em um mesmo cromossomo, com menos de 1cM de distância entre eles, foi considerado como uma região. Portanto, a média da variância fenotípica daqueles SNPs foi considerada, e não a variância fenotípica de cada SNP.

Para soja, na literatura, ainda não foi reportada a utilização do caráter valor agrônômico em análises GWAS, uma vez que, a avaliação desse caráter pode ser considerada muito subjetiva. Entretanto, quando bem conduzida, por representar uma boa sumarização dos caracteres de uma planta, pode ser um considerado um caráter promissor para GWAS. Assim, para este trabalho, o valor agrônômico de cada genótipo foi utilizado para GWAS. Considerando os valores de BLUPs das análises individuais e da conjunta, foram identificadas, no total, para valor agrônômico, 18 regiões, correspondentes aos 43 SNPs mencionados. No ambiente 1, verificou-se apenas um SNP localizado no cromossomo 10, na qual a variância fenotípica explicada foi de 12,10%. No ambiente 2, também foram identificadas 3 regiões, essas associadas aos cromossomos 1, 12 e 13, que são responsáveis por 41,09% da variância fenotípica. Já no terceiro ambiente, 3 regiões foram associadas e estão localizadas nos cromossomos 1, 11 e 20, correspondendo a 44,72% da variância fenotípica. Nos demais ambientes avaliados, sendo o 4 e 6, foram identificados, respectivamente, 4 e 2 regiões, explicando 53,72% e 46,97% da variância fenotípica. Além disso, na análise conjunta, identificaram-se apenas 2 regiões nos cromossomos 5 e 13, que juntos corresponderam a 27,46% da variância fenotípica. Destaca-se que nas análises GWAS dos ambientes 1, 3 e 6, essas apresentaram regiões no cromossomo 1 bem próximas, assim como no ambiente 4 apresentou uma região em comum com a análise conjunta, localizada no cromossomo 5 em 78-79cM.

Como ainda não foi reportado nenhum QTL dito como associado especificamente a valor agrônômico em soja, neste trabalho as regiões encontradas para este caráter foram examinadas no *SoyBase* e verificada a presença de QTLs reportados. Valor agrônômico é um caráter avaliado por meio de escala de notas, sendo essa representativa de um conjunto de caracteres visuais como, arquitetura da planta, quantidade de vagens cheias, vigor e sanidade da planta, debulha prematura das vagens, acamamento e retenção foliar na maturidade. Esse

caráter está diretamente relacionado com produtividade, evidenciado também pela correlação alta e positiva apresentada por ambos [47] (Figura 1). Logo, 10 das 13 regiões identificadas para valor agrônômico (VA) apresentam relatos de QTLs para produtividade de grãos (PG) (Tabela 1). A relação da produtividade e valor agrônômico, também fica evidenciada quando comparamos os SNPs encontrados para ambos, na qual 33,6% dos SNPs verificados para produtividade, também são apresentados para valor agrônômico (Tabela 1).

Considerando a massa de cem sementes (MCS), foram encontradas associações nos ambientes 1, 2, 3, 4 e 5. No ambiente 1, 5 SNPs foram identificados na região entre 7,48 cM-8,04 do cromossomo 18, correspondendo à aproximadamente 15 % da variância fenotípica. No ambiente 2, também foi verificado um SNP no cromossomo 18, mas em uma região bem distante da identificada no ambiente 1, não apresentando nenhuma relação. Nos demais ambientes avaliados, foram encontradas 5 regiões nos cromossomos 10, 11, 14 e 16, sendo que nenhuma destas regiões foram identificadas em mais de um local e representaram cerca de 25% a 28% da variância fenotípica. Interessante ressaltar que, várias regiões encontradas para massa de cem sementes também foram associadas com produtividade, altura da planta na maturidade, valor agrônômico e número de dias para maturidade, demonstrado que uma região pode apresentar interdependência entre caracteres. A interdependência entre os caracteres apresentadas é corroborada pela correlação significativa apresentada na Figura 3.

Em relação ao caráter acamamento (AC), o SNP identificado como AX-90376352, localizado na região 71,23 cM do cromossomo 13 foi identificado nas análises do ambiente 1 e na conjunta. No ambiente 1, apenas um SNP foi identificado e correspondeu a 11,67% da variância. Já no ambiente 2, 3 SNPs localizados em diferentes cromossomos, explicaram 40,93% da variância fenotípica. Nos ambientes 4 e 6, ambos apresentaram apenas uma região em associação nos cromossomos 20 e 9, respectivamente. Por fim, na análise conjunta foram identificadas duas regiões nos cromossomos 4 e 13 com 26% da variância fenotípica. Ainda é interessante destacar que, algumas regiões foram também associadas para teor de óleo, valor agrônômico, período de granação e massa de cem sementes apresentaram correlações negativas significativas (Figura 3). Além disso, o caráter acamamento também apresentou regiões compartilhadas com altura da planta no florescimento e altura da planta na maturidade o que corrobora o demonstrado por Wen et al. [48]. Os autores reportaram que algumas regiões identificadas no cromossomo 10, 19 estão fortemente associadas a acamamento e também a altura de planta, maturidade e florescimento, demonstrando dependência desses caracteres. As regiões apresentadas por Wen et al. [48] não foram verificadas associadas a

acamamento neste estudo. Todavia, foram identificadas 7 regiões associadas a acamamento que ainda não foram reportadas na literatura.

Para o caráter período de granação (PEG), verificamos associações nos ambientes 1, 2, 4 e 5, localizadas nos cromossomos 1, 5, 11, 12, 13, 17 e 18. Destaca-se que no ambiente 4 foram identificados 5 SNPs que corresponderam com 67,77% da variância fenotípica. Já na conjunta, uma região associada a PEG foi identificada no cromossomo 13, com apenas 12,39% da variância fenotípica, sendo que essa também foi verificada associada para número de dias para maturidade (NDM).

Para número de dias para a maturidade (NDM), foram identificadas na análise conjunta, duas regiões do cromossomo 18, em 7,47 cM e em 129 cM, e uma região no cromossomo 15, que também foram apresentadas nas análises dos ambientes 2 e 5. No ambiente 2, as regiões identificadas associadas a NDM explicaram 85,55% da variância fenotípica. Também foram identificadas regiões associadas nos ambientes 3 e 4.

Para altura da inserção da primeira vagem (AIPV), o SNP AX-90415118 foi verificado na análise do ambiente 2 e na conjunta e representou em média 14% da variância fenotípica para ambas análises em que esse foi verificado. Além disso, outro SNP também foi identificado em comum nos ambientes 4 e 6. Em geral para AIPV, 5 regiões distintas foram associadas a ela nos cromossomos 5, 8, 9, 15, 16. Cada ambiente, em média, apresentou regiões associadas à AIPV correspondente a 14% da variância fenotípica. Os SNPs apresentados no cromossomo 5, estão localizados em uma região já relatada, contendo um QTL para AIPV por Salas et al. [49].

No geral, para o caráter altura da planta na maturidade (APM), 8 regiões dos cromossomos 2, 4, 10, 11, 12 e 16 foram identificadas, considerando todas as análises dos ambientes em que esse foi avaliado, com exceção do ambiente 5 que o caráter não foi avaliado. Um SNP associado à APM foi verificado na análise de GWAS nos ambientes 1 e 2 e na conjunta, esse está localizado no cromossomo 12 em 8,9 cM e explicou em média 12% da variância fenotípica.

Para o caráter altura da planta no florescimento (APF) verificou-se associações apenas nos ambientes 2, 5 e na conjunta, sendo verificado apenas uma região para cada ambiente, localizadas nos cromossomos 12, 11 e 5, respectivamente. A região do cromossomo 5 e 12 apresentaram associações para APM. Também foi verificada correlação positiva entre as variáveis (Figura 1). Associações compartilhadas entre altura de planta, maturidade e florescimento também já foram relatadas por Wen et al. [48].

O caráter teor de óleo (TO), o qual foi apenas avaliado nos ambientes 2 e 5, apresentou regiões associadas nos cromossomos 9, 13,5 e 18. Já na conjunta, além dos mesmos SNPs verificados no ambiente 2, esse apresentou um SNPs no cromossomo 4 que ainda não foi relatado na literatura. Eskandari et al. [50] apresentaram QTLs para as mesmas regiões encontradas nos cromossomos 13 e 9.

Com relação à produtividade foram identificadas regiões associadas nas análises de todos os ambientes, incluindo também a análise conjunta. Cerca de 26,20%, 56,52%, 42,62%, 50,62%, 53,73%, 13,74% e 55,89% correspondem à somatória das variâncias fenotípicas das regiões associadas ao caráter para cada análise respectivamente. Na análise do ambiente 1, duas regiões do cromossomo 10 foram identificadas, sendo que a localizada em 0,14 cM também foi encontrada na conjunta. No ambiente 2, a região 95cM do cromossomo 1 também foi verificada na análise conjunta, bem como a região 61,23 cM do cromossomo 13. Ainda no ambiente 2, este apresentou outro SNP, AX-90472901, em comum com o ambiente 5, que correspondeu a 14,29% da variância fenotípica. Várias regiões já foram identificadas por outros autores associadas a produtividade de grãos (Tabela 2), contudo, alguns não foram relatadas.

Em suma, considerando os dez caracteres avaliados neste estudo, os locos de SNPs identificados, explicaram, em média, 13,97% da variância fenotípica, sendo que o mínimo encontrado foi de 10,54% para um loco associado à APM e o máximo foi de 20,62% para o mesmo caráter. Entretanto, há alguns caracteres que a somatória das variâncias fenotípicas das regiões encontradas correspondeu a 85,52%. Isto demonstra que as regiões identificadas possuem um grande potencial por explicar mais que 10% da variância fenotípica de um caráter complexo, na qual é governado por vários genes, devendo ser validadas e futuramente utilizadas em programas de melhoramento [17,18]. Entretanto, a utilização dos resultados de GWAS em rotinas de seleção assistida por marcadores é algo ainda desafiador, pois é dependente de um sistema de genotipagem rápido e com baixo custo [51]. Contudo, as regiões identificadas associadas a algum caráter via GWAS pode auxiliar na busca de alelos promissores em bancos de germoplasma, e no aumento da acurácia de predição em modelos de seleção genômica ampla [18].

A GWAS pode ser utilizada em conjunto com o mapeamento de ligação, de forma a confirmar e examinar a variância explicada para aquele fenótipo. Entretanto, vários fatores podem comprometer o sucesso da GWAS [52,53], entre eles o tamanho da população de estudo. Populações de tamanho reduzido podem acarretar problemas devido às frequências de alelos QTL e pouca diversidade genética, reduzindo a probabilidade de sucesso da GWAS

[54,55], bem como aumentar o número de associações falsas [56]. Apesar do tamanho relativamente pequeno da população, a boa qualidade dos experimentos e da fenotipagem, bem como a grande quantidade representativa de ambientes avaliados justificaram a realização de GWAS. Além disso, para superar estes desafios, em nosso estudo, utilizamos um painel de soja contendo genótipos de diferentes localidades do mundo, abrangendo uma grande variabilidade genética. A escolha do painel levou em conta a diversidade genética de todos os caracteres estudados. As etapas de filtragem dos marcadores SNPs utilizados garantiu aumento na qualidade desses, bem como resultou em um número elevado de marcas. GWAS conduzidas com sucesso em populações próximas de 100 genótipos já foram conduzidas em várias culturas [57–59], incluindo soja [60–63]. A presença de QTLs já relatados sugere que esses se apresentam confiáveis o bastante para serem utilizados em programas de seleção assistida por marcadores (MAS). Para os locos ainda não reportados na literatura novos estudos devem ser realizados para a confirmação e averiguação do envolvimento com o controle dos caracteres. Dos 181 locos identificados associados a algum dos caracteres avaliados, 124 deles ainda não foram reportados na literatura e podem ser considerados como potenciais candidatos aos seus respectivos caracteres.

Como a GWAS foi realizada considerando cada ambiente em particular, foi possível verificar que os SNPs, e, portanto, as regiões encontradas associadas, em sua maioria, não são estáveis. A estabilidade pode ser defendida com o aparecimento de um mesmo SNP e/ou região genômica em diferentes análises considerando diferentes ambientes. Relatado por Contreras-Soto et al.[64] e também verificado neste trabalho, os SNPs e regiões identificadas são específicas de cada ambiente avaliado, demonstrando a complexa interação QTL ambiente (QxE). Além disso, a presença da interação QxE evidencia a complexidade dos caracteres estudados. Entretanto, a instabilidade das associações já era esperada, devido à grande variação dos genótipos dentro de cada ambiente e pela presença de interação genótipo ambiente significativa [64]. A presença de regiões divergentes associadas para um mesmo caráter, considerando ambientes distintos é relatada por vários estudos [46,64,65] e apresenta-se como um objeto importante a ser introduzido nos modelos de análise de GWAS e deve ser levada em consideração para aplicação dos QTLs em programas de seleção assistida por marcadores.

3.4. Conclusão

Os resultados apresentados por este estudo forneceram evidência da complexidade da arquitetura genética de dez importantes caracteres em soja. No total 181 SNPs foram identificados associados a estes caracteres, avaliados em seis ambientes. A presença de regiões genômicas distintas associadas para um mesmo caráter, em ambientes diferentes, evidencia a complexidade da regulação dos caracteres e da interação QTL e ambiente. Algumas regiões identificadas associadas aos caracteres estudados já foram previamente reportadas por outros estudos, entretanto, a maioria das regiões encontradas são inéditas e possuem usos potenciais a serem confirmados em estudos posteriores.

REFERÊNCIAS

1. Gepts P. Legumes as a Model Plant Family. Genomics for Food and Feed Report of the Cross-Legume Advances through Genomics Conference. *Plant Physiol.* 2005;137: 1228–1235. doi:10.1104/pp.105.060871
2. Chaudhary J, Patil GB, Sonah H, Deshmukh RK, Vuong TD, Valliyodan B, et al. Expanding Omics Resources for Improvement of Soybean Seed Composition Traits. *Front Plant Sci.* 2015;6: 1–16. doi:10.3389/fpls.2015.01021
3. Food and Agriculture Organization. Accessed on december 2017 at www.fao.org.
4. Malik MFA, Ashraf M, Qureshi AS, Ghafoor A. Assessment of genetic variability, correlation and path analyses for yield and its components in soybean. *Pakistan J Bot.* 2007;39: 405–413.
5. Li Y hui, Zhao S cen, Ma J xin, Li D, Yan L, Li J, et al. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics.* 2013;14. doi:10.1186/1471-2164-14-579
6. Ray JD, Dhanapal AP, Singh SK, Hoyos-Villegas V, Smith JR, Purcell LC, et al. Genome-Wide Association Study of Ureide Concentration in Diverse Maturity Group IV Soybean [*Glycine max* (L.) Merr.] Accessions. *G3.* 2015;5: 2391–403. doi:10.1534/g3.115.021774
7. Varshney RK, Glaszmann J-C, Leung H, Ribaut J-M. More genomic resources for less-studied crops. *Trends Biotechnol.* Elsevier Ltd; 2010;28: 452–60. doi:10.1016/j.tibtech.2010.06.007

8. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol.* 2015;33: 408–14. doi:10.1038/nbt.3096
9. Sajjad M, Khan S., Khan AS. Exploitation of germplasm for grain yield Improvement in spring wheat (*Triticum aestivum*). *Inter J Agric Biol.* 2011;13: 695–700. doi:10.13140/2.1.4612.9285
10. Al-Maskri AY, Sajjad M, Khan SH. Association mapping: A step forward to discovering new alleles for crop improvement. *Int J Agric Biol.* 2012;14: 153–160. doi:10.13140/2.1.1925.9524
11. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods.* 2013;9: 29. doi:10.1186/1746-4811-9-29
12. Gupta PK, Rustgi S, Kulwal PL. Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol.* 2005;57: 461–85. doi:10.1007/s11103-005-0257-z
13. Mezmouk S, Dubreuil P, Bosio M, Décousset L, Charcosset A, Praud S, et al. Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. *Theor Appl Genet.* 2011;122: 1149–1160. doi:10.1007/s00122-010-1519-y
14. Flint-Garcia SA, Thornsberry JM, Buckler ES. Structure of Linkage Disequilibrium in Plants. *Annu Rev Plant Biol.* 2003;54: 357–374. doi:10.1146/annurev.arplant.54.031902.134907
15. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci.* 2001;98: 11479–11484. doi:10.1073/pnas.201394398
16. Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics.* 2010;9: 166–177. doi:10.1093/bfpg/elq001
17. He J, Zhao X, Laroche A, Lu Z-X, Liu H, Li Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front Plant Sci.* 2014;5: 1–8. doi:10.3389/fpls.2014.00484
18. Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, et al. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS One.* 2014;9: 1–12. doi:10.1371/journal.pone.0093017
19. Mulato BM, Möller M, Zucchi MI, Quecini V, Pinheiro JB. Genetic diversity in soybean germplasm identified by SSR and EST-SSR markers. *Pesqui Agropecuária Bras.* 2010;45: 276–283. doi:10.1590/S0100-204X2010000300007

20. Marconato MB, Pereira EM, Silva FM, Bizari EH, Pinheiro JB, Mauro AO, et al. Genetic divergence in a soybean (*Glycine max*) diversity panel based on agro-morphological traits. *Genet Mol Res*. 2016;15: 1–10. doi:10.4238/gmr15048980
21. Doyle JJ, Doyle JL. A rapid isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*. 1987;19:11–15.
22. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am J Hum Genet*. 2007;81: 1084–1097. doi:10.1086/521987
23. R Core Team: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2013.
24. Knapp SJ, Stroup WW, Ross WM. Exact confidence intervals for heritability on a progeny mean basis. *Crop Sci*. 1985;25: 192–194. doi:10.2135/cropsci1985.0011183X002500010046x
25. Bates D, Maechler M, Bolker BM, Walker S. Fitting linear mixed-effects models using {lme4}. *J Stat Softw*. 2015;67: 1–48. doi:10.18637/jss.v067.i01
26. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: Genome association and prediction integrated tool. *Bioinformatics*. 2012;28: 2397–2399. doi:10.1093/bioinformatics/bts444
27. Dhanapal AP, Ray JD, Singh SK, Hoyos-Villegas V, Smith JR, Purcell LC, et al. Association mapping of total carotenoids in diverse soybean genotypes based on leaf extracts and high-throughput canopy spectral reflectance measurements. *PLoS One. BMC Plant Biology*; 2015;10: 1–20. doi:10.1371/journal.pone.0137213
28. Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb)*. Nature Publishing Group; 2012;108: 285–291. doi:10.1038/hdy.2011.73
29. Hao D, Cheng H, Yin Z, Cui S, Zhang D, Wang H, et al. Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. *Theor Appl Genet*. 2012;124: 447–458. doi:10.1007/s00122-011-1719-0
30. Hwang E-Y, Song Q, Jia G, Specht JE, Hyten DL, Costa J, et al. A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics*. 2014;15: 1. doi:10.1186/1471-2164-15-1
31. Zhang J, Song Q, Cregan PB, Nelson RL, Wang X, Wu J, et al. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics*. ???; 2015;16: 217. doi:10.1186/s12864-015-1441-4

32. Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, Kilian B, et al. Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC Plant Biol.* 2012;12: 16. doi:10.1186/1471-2229-12-16
33. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association Mapping in Structured Populations. *Am J Hum Genet.* 2000;67: 170–181. doi:10.1086/302959
34. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol.* 2005;14: 2611–20. doi:10.1111/j.1365-294X.2005.02553.x
35. Wimmer V, Albrecht T, Auinger H-J, Schön C-C. synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics.* 2012;28: 2086–2087. doi:10.1093/bioinformatics/bts335
36. Wang X, Liang G, Marci J, Scott RA, Song Q, Hyten DL, et al. Identification and validation of quantitative trait loci for seed yield , oil and protein contents in two recombinant inbred line populations of soybean. *Molecular Genetics and Genomics.* 2014. pp. 935–949. doi:10.1007/s00438-014-0865-x
37. Silva KB, Bruzi AT, Zuffo AM, Zambiazzi E V., Soares IO, de Rezende PM, et al. Adaptability and phenotypic stability of soybean cultivars for grain yield and oil content. *Genet Mol Res.* 2016;15: 1–11. doi:10.4238/gmr.15026756
38. Mangin B, Sandron F, Henry K, Devaux B, Willems G, Devaux P, et al. Breeding patterns and cultivated beets origins by genetic diversity and linkage disequilibrium analyses. *Theor Appl Genet.* Springer Berlin Heidelberg; 2015;128: 2255–2271. doi:10.1007/s00122-015-2582-1
39. Adetunji I, Willems G, Tschoep H, Bürkholz A, Barnes S, Boer M, et al. Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. *Theor Appl Genet.* 2014;127: 559–571. doi:10.1007/s00122-013-2239-x
40. Hoffstetter A, Cabrera A, Huang M, Sneller C. Optimizing Training Population Data and Validation of Genomic Selection for Economic Traits in Soft Winter Wheat. *G3: Genes|Genomes|Genetics.* 2016;6: 2919–2928. doi:10.1534/g3.116.032532
41. Hiromoto DM, Vello NA. The genetic bae of brazilian soybean(*Glycine max*(L.) MERRILL) cultivars. *Genet Brazilian J.* 1986; 295–306.
42. Wysmierski PT, Vello NA. The genetic base of Brazilian soybean cultivars: Evolution over time and breeding implications. *Genet Mol Biol.* 2013;36: 547–555. doi:10.1590/S1415-47572013005000041

43. Brown-Guedira GL, Thompson JA, Nelson RL, Warburton ML. Evaluation of Genetic Diversity of Soybean Introductions and North American Ancestors Using RAPD and SSR Markers. *Crop Sci.* 2000;40: 815. doi:10.2135/cropsci2000.403815x
44. Feng Y, Lu Q, Zhai R, Zhang M, Xu Q, Yang Y, et al. Genome wide association mapping for grain shape traits in indica rice. *Planta.* Springer Berlin Heidelberg; 2016;244: 819–830. doi:10.1007/s00425-016-2548-9
45. Jia Y, Sun X, Sun J, Pan Z, Wang X, He S, et al. Association mapping for epistasis and environmental interaction of yield traits in 323 cotton cultivars under 9 different environments. *PLoS One.* 2014;9: e95882. doi:10.1371/journal.pone.0095882
46. Lasky JR, Upadhyaya HD, Ramu P, Deshpande S, Hash CT, Bonnette J, et al. Genome-environment associations in sorghum landraces predict adaptive traits. *Sci Adv.* 2015;1: e1400218–e1400218. doi:10.1126/sciadv.1400218
47. Wallace de SL, Sandra HU-T, Eder LG, Cleber VGA, Bruno HPV, Eduardo HB, et al. Agronomic performance and path analysis of roundup ready and conventional soybean from two-way crosses in the northwestern of So Paulo, Brazil. *African J Agric Res.* 2016;11: 4584–4593. doi:10.5897/AJAR2016.11700
48. Wen Z, Boyse JF, Song Q, Cregan PB, Wang D. Genomic consequences of selection and genome-wide association mapping in soybean. *BMC Genomics.* BMC Genomics; 2015;16: 671. doi:10.1186/s12864-015-1872-y
49. Chase PSJCODWK, Mansur L. Genetic mapping of seed shape in three populations of recombinant inbred lines of soybean (*Glycine max* L . Merr .). 2006; 1459–1466. doi:10.1007/s00122-006-0392-1
50. Eskandari M, Cober ER. Using the candidate gene approach for detecting genes underlying seed oil concentration and yield in soybean. 2013; 1839–1850. doi:10.1007/s00122-013-2096-7
51. Pantalião GF, Narciso M, Guimarães C, Castro A, Colombari JM, Breseghello F, et al. Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. *Genetica.* 2016;144: 651–664. doi:10.1007/s10709-016-9932-z
52. Cedric Gondro; Julius van der Werf; Ben Hayes. *Genome-wide association studies and genomic prediction.* Humana Press, editor. Totowa, NJ; 2013.
53. Cao K, Zhou Z, Wang Q, Guo J, Zhao P, Zhu G, et al. Genome-wide association study of 12 agronomic traits in peach. *Nat Publ Gr.* Nature Publishing Group; 2016;7: 1–10. doi:10.1038/ncomms13246
54. Gaut BS, Long AD. The lowdown on linkage disequilibrium. *Plant Cell.* 2003;15: 1502–1506. doi:10.1105/tpc.150730

55. Bernardo R. Parental selection, number of breeding populations, and size of each population in inbred development. *TAG Theor Appl Genet.* 2003;107: 1252–1256. doi:10.1007/s00122-003-1375-0
56. Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics.* 1990;124: 743–756. doi:10.1046/j.1365-2540.1998.00308.x
57. Racedo J, Gutiérrez L, Perera MF, Ostengo S, Pardo EM, Cuenya MI, et al. Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biol. BMC Plant Biology;* 2016;16. doi:10.1186/s12870-016-0829-x
58. Andersen JR, Schrag T, Melchinger AE, Zein I, Lübberstedt T. Validation of Dwarf8 polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.). *Theor Appl Genet.* 2005;111: 206–217. doi:10.1007/s00122-005-1996-6
59. Skøt L, Humphreys MO, Armstead I, Heywood S, Skøt KP, Sanderson R, et al. An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.). *Mol Breed.* 2005;15: 233–245. doi:10.1007/s11032-004-4824-9
60. Zatybekov A, Abugalieva S, Didorenko S, Gerasimova Y, Sidorik I, Anuarbek S, et al. GWAS of agronomic traits in soybean collection included in breeding pool in Kazakhstan. *BMC Plant Biol.* 2017;17. doi:10.1186/s12870-017-1125-0
61. Abugalieva S, Didorenko S, Anuarbek S, Volkova L, Gerasimova Y, Sidorik I, et al. Assessment of soybean flowering and seed maturation time in different latitude regions of Kazakhstan. *PLoS One.* 2016;11: 1–12. doi:10.1371/journal.pone.0166894
62. Yan L, Hofmann N, Li S, Ferreira ME, Song B, Jiang G, et al. Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC Genomics. BMC Genomics;* 2017;18: 1–11. doi:10.1186/s12864-017-3922-0
63. Bastien M, Sonah H, Belzile F. Genome Wide Association Mapping of Resistance in Soybean with a Genotyping-by-Sequencing Approach. *Plant Genome.* 2014;7: 0. doi:10.3835/plantgenome2013.10.0030
64. Mora F, Anto M, Oliveira R De, Higashi W, Scapim CA, Schuster I. A Genome-Wide Association Study for Agronomic Traits in Soybean Using SNP Markers and SNP-Based Haplotype Analysis. 2017; 1–22. doi:10.1371/journal.pone.0171105
65. Begum H, Spindel JE, Lalusin A, Borromeo T, Gregorio G, Hernandez J, et al. Genome-Wide Association Mapping for Yield and Other Agronomic Traits in an Elite Breeding Population of Tropical Rice (*Oryza sativa*). *PLoS One.* 2015;10: e0119873. doi:10.1371/journal.pone.0119873

FIGURAS

Figura 1. Correlações fenotípicas entre os caracteres estudados: APF, altura da planta no florescimento; APM, altura da planta na maturidade; AC, acamamento; VA, valor agrônômico; PG: produtividade de grão; MCS, Massa de cem sementes; TO, teor de óleo; PEG, período de granação; NDM, número de dias para maturidade; e AIPV, altura da inserção da primeira vagem. Significância baseada em probabilidade pelo Teste t (* $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$).

Material Suplementar 1. Figura S1. Estrutura da população de soja com base no software Structure. O eixo vertical mostra a probabilidade de cada genótipo pertencer a cada subpopulação. A sub-população 1 e 2 são representadas como azul e vermelho, respectivamente. O eixo horizontal representa os nomes dos genótipos.

TABELAS

Tabela 1. Estatísticas descritivas, análises de variância e herdabilidade de dez caracteres mensurados em 95 genótipos de soja.

Carater	Ambiente	Min	Máx	Média	Sd	CV%	F _g	F _e	F _{gxe}	h ²
APM	A1	28.67	177	100.75	12.84	14.46	82.93***	187.19***	2.48***	0.88
	A2	32.33	159.33	94.77	7.94	9.12				
	A3	40	153.33	106.84	11.89	10.99				
	A4	20.67	151.67	94.52	10.94	12.34				
	A5	23.33	135	84.72	9.23	11.34				
	A6	17	141.67	77.45	10.01	15.05				
AC	A1	1	5	2.8	0.48	19.3	22.89***	193.18***	2.28***	0.66
	A2	1	4.67	1.7	0.32	17.21				
	A3	1	5	2.85	0.78	28.64				
	A4	1	5	2.63	0.58	23.36				
	A6	1	3.33	1.6	0.56	31.7				
APF	A2	28.67	126.67	82.85	8.06	10.4	24.69***	751.84***	3.84***	0.62
	A5	20.33	105	58.95	8.51	15				
AIPV	A2	4	26.67	13.06	2.25	16.76	16.93***	31.04***	2.45***	0.57
	A4	6	26.33	13.07	1.95	15.37				
	A5	4.33	27	12.45	2.77	21.78				
	A6	1.67	24	11.05	2.72	25.64				
NDM	A2	88.33	142.33	130.12	6.21	4.89	17.91***	305.71***	1.87***	0.55
	A3	116.67	189.33	160.78	9.47	6.15				
	A4	115	186	138	5.11	3.75				
	A5	75.33	148.33	122.68	8.75	7.29				
TO	A2	13.33	23	18.76	0.83	4.49	15.06***	169.14***	2.76***	0.82
	A5	13.67	21.67	17.46	1.36	7.97				
PG	A1	180	4200	2034.89	384.79	20.33	19.89***	372.68***	2.72***	0.62
	A2	274	4467	2072.59	712	41.68				
	A3	83.33	3510	1313.74	245.9	23.74				
	A4	37.33	3697.33	1865.81	416.37	25.69				
	A5	351.33	4387.67	1704.49	598.01	39.06				
	A6	4	1202.67	329.38	148.05	53.33				
P100	A1	8.33	27.67	15.36	1.04	6.6	35.87***	182.17***	2.53***	0.75
	A2	8.67	22.67	14.95	3.33	22.11				
	A3	5.67	25.67	12.42	0.91	7.97				
	A4	6	23.67	13.99	1.29	9.91				
	A5	6.33	22.33	11.64	1.63	13.92				
PEG	A1	24	65.67	41.97	4.35	10.45	5.02***	35.86***	1.65***	0.25
	A2	13.33	57.33	42.44	5.8	14.08				
	A4	18	53	35.94	7.33	20.32				
	A5	22.67	66.67	41.67	10.4	24.94				
VA	A1	1	5	3.01	0.38	13.71	14.10***	137.63***	2.88***	0.49
	A2	1.33	3.33	2.49	0.38	16.18				
	A3	1	5	2.09	0.46	21.98				
	A4	1	5	2.97	0.6	21.79				
	A6	1	4	2.17	0.64	30.49				

*** Efeito significativo ao nível de probabilidade de 0,01; F_G, efeitos de genótipos; F_E, efeitos de ambientes; F_G × F_E, efeitos da interação genótipos e ambiente; APF, altura da planta no florescimento; APM, altura da planta na maturidade; AC, acamamento; VA, valor agrônômico; PG, produtividade de grão; MCS, massa de cem sementes; TO, teor de óleo; PEG, período de granação; NDM, número de dias para maturidade; e AIPV, altura da inserção da primeira vagem.

Tabela 2. Detalhes dos marcadores SNPs associados com dez caracteres avaliados em um painel de genótipos de soja avaliados em múltiplos ambientes. Caract: carater avaliado; A: ambiente em que o carater foi avaliado; C: análise conjunta; SNP: marcador onde foi verificada a associação; Crom: cromossomo; cM: posição onde a associação foi verificada em centiMorgans; Var: variação fenotípica; ^aBaseado na lista QTL apresentado no SoyBase (<https://www.soybase.org/>).

(continua)

Caract	A	SNP	Crom	cM	P-valor	Var (%)	Var total	QTLs reportados ^a
PG	A1	AX-90480405	10	0,14	7,22E-04	13,50	26,20	Seed yield 22-19
		AX-90448602	10	13,76	1,01E-03	12,71		N
		AX-90467518	10	13,73	1,01E-03	12,71		N
	A2	AX-90434500	1	95,71	3,97E-04	15,01	56,52	N
		AX-90319375	1	95,99	3,97E-04	15,01		N
		AX-90472901	16	14,25	5,37E-04	14,29		Seed yield 23-6; Seed yield to Plant height ratio 4-1
		AX-90425892	13	61,23	5,72E-04	14,14		N
		AX-90505916	16	19,96	9,02E-04	13,07		Seed yield to Plant height ratio 1-3
	A3	AX-90403918	15	96,76	4,25E-04	14,75	42,62	N
		AX-90394904	15	96,78	4,25E-04	14,75		N
		AX-90366589	14	10,10	4,93E-04	14,40		N
		AX-90407580	19	6,90	7,30E-04	13,47		N
		AX-90385201	19	6,97	7,30E-04	13,47		N
	A4	AX-90425224	7	82,89	6,91E-04	13,24	50,62	N
		AX-90456035	5	79,56	8,72E-04	12,71		Seed yield 22-10
		AX-90335326	5	79,57	8,72E-04	12,71		Seed yield 22-10
		AX-90389813	5	79,70	8,72E-04	12,71		Seed yield 22-10
		AX-90496976	5	79,64	8,76E-04	12,71		Seed yield 22-10
	A5	AX-90498171	7	6,30	9,89E-04	12,43	53,73	Seed yield 19-5
		AX-90308787	15	15,91	1,08E-03	12,24		N
		AX-90408914	15	15,93	1,08E-03	12,24		N
		AX-90333287	15	15,93	1,08E-03	12,24		N
		AX-90472901	16	14,25	4,63E-04	14,50		Seed yield 23-6; Seed yield to Plant height ratio 4-1
		AX-90341910	8	88,75	6,38E-04	13,74		N
		AX-90312028	11	69,03	8,83E-04	12,99		N
		AX-90504931	1	6,69	1,08E-03	12,51		N
	A6	AX-90499551	18	129,51	5,94E-04	13,74	13,74	N
	C	AX-90434500	1	95,71	3,17E-04	14,99	55,89	N
		AX-90319375	1	95,99	3,17E-04	14,99		N
		AX-90480405	10	0,14	3,58E-04	14,71		Seed yield 22-19
		AX-90466377	1	95,39	4,19E-04	14,34		N
		AX-90513398	14	99,25	5,60E-04	13,68		Seed yield 31-1
		AX-90425892	13	61,23	9,39E-04	12,51		N
VA	A1	AX-90480405	10	0,14	1,40E-03	12,10	12,10	Seed yield 22-19
	A2	AX-90448362	13	0,11	4,51E-04	14,69	41,09	Seed yield 18-3
		AX-90493595	13	0,04	5,62E-04	14,17		Seed yield 18-3
		AX-90486319	12	81,27	7,18E-04	13,59		N
	A3	AX-90498285	12	81,38	7,18E-04	13,59	44,47	N
		AX-90475745	12	81,65	7,18E-04	13,59		N
		AX-90317456	12	81,78	7,18E-04	13,59		N
		AX-90307731	12	81,83	7,18E-04	13,59		N
		AX-90333615	12	81,86	7,18E-04	13,59		N
	A4	AX-90491649	12	81,28	7,18E-04	13,59	53,72	N
		AX-90360535	12	81,38	7,18E-04	13,59		N
		AX-90394566	1	117,43	1,00E-03	12,81		N
		AX-90451047	1	117,45	1,00E-03	12,81		N
		AX-90523899	11	81,89	1,90E-04	17,25		N
	A5	AX-90434167	11	81,89	1,90E-04	17,25	44,47	N
		AX-90397248	11	81,59	4,98E-04	14,86		N
		AX-90455245	20	71,64	7,36E-04	13,91		N
		AX-90434500	1	95,71	9,43E-04	13,31		N
		AX-90319375	1	95,99	9,43E-04	13,31		N
	A6	AX-90521782	5	18,48	6,40E-04	13,75	53,72	N
		AX-90363814	5	78,31	7,04E-04	13,53		Seed yield 22-10
		AX-90465027	5	78,34	7,04E-04	13,53		Seed yield 22-10
		AX-90414937	5	78,29	7,04E-04	13,53		Seed yield 22-10
		AX-90376684	5	78,36	7,04E-04	13,53		Seed yield 22-10

(continuação)

Caract	A	SNP	Crom	cM	P-valor	Var(%)	Var total	QTLs reportados ^a
VA	A4	AX-90470246	15	91,14	7,79E-04	13,30		N
		AX-90423663	8	19,99	8,33E-04	13,14		N
		AX-90492713	8	20,00	8,33E-04	13,14		N
	A6	AX-90462244	1	104,51	8,63E-05	18,75	46,97	N
		AX-90498649	1	104,53	8,63E-05	18,75		N
		AX-90449872	1	104,48	8,63E-05	18,75		N
		AX-90498048	1	104,51	8,63E-05	18,75		N
		AX-90317519	1	104,54	8,63E-05	18,75		N
		AX-90522778	5	87,02	3,79E-04	15,12		N
		AX-90498236	16	13,86	8,90E-04	13,10		Seed yield 23-6;Seed yield to Plant height ratio 4-1
		AX-90415118	5	86,82	9,50E-04	12,95		N
	C	AX-90456035	5	79,56	5,95E-04	14,03	27,47	Seed yield 22-10
		AX-90335326	5	79,57	5,95E-04	14,03		Seed yield 22-10
		AX-90389813	5	79,70	5,95E-04	14,03		Seed yield 22-10
		AX-90358248	5	79,74	6,87E-04	13,69		Seed yield 22-10
		AX-90334315	5	79,80	7,60E-04	12,86		Seed yield 22-10
		AX-90342736	5	79,77	9,81E-04	12,86		Seed yield 22-10
		AX-90370853	5	79,77	9,81E-04	12,86		Seed yield 22-10
		AX-90376352	13	71,26	9,81E-04	13,45		Seed yield to weight ratio 2-3
AC	A1	AX-90519720	12	8,90	1,35E-03	11,67	11,67	N
	A2	AX-90376352	13	71,26	2,72E-04	15,41	40,93	Plant height to Lodging ratio 4-2
		AX-90443142	4	90,93	7,13E-04	13,19		N
		AX-90506630	11	33,90	1,04E-03	12,33		N
	A3	AX-90436413	6	83,97	9,70E-04	12,78	12,77	N
	A4	AX-90487717	20	49,00	9,68E-04	12,81	12,81	N
	A6	AX-90408420	9	98,44	2,65E-04	15,39	15,38	N
		AX-90450888	9	98,53	2,91E-04	15,17		N
		AX-90520819	9	98,68	2,91E-04	15,17		N
		AX-90460435	9	98,53	4,74E-04	14,05		N
	C	AX-90321935	4	80,13	7,54E-04	13,07	26,00	N
		AX-90376352	13	71,26	7,98E-04	12,94		Plant height to Lodging ratio 4-2
MCS	A1	AX-90307081	18	7,48	1,37E-04	17,35	17,34	N
		AX-90309671	18	7,49	2,71E-04	15,71		N
		AX-90320156	18	7,47	2,71E-04	15,71		N
		AX-90451320	18	7,99	7,33E-04	13,37		N
		AX-90485716	18	8,00	7,33E-04	13,37		N
	A2	AX-90374471	18	106,00	7,72E-04	13,59	13,58	Seed weight per plant 6-8
	A3	AX-90422989	14	13,22	2,88E-04	15,29	28,54	N
		AX-90405022	11	82,22	6,91E-04	13,26		Seed weight 10-3
	A4	AX-90365866	10	99,32	9,23E-04	12,99	12,99	Seed weight 34-8
		AX-90419663	10	99,95	9,23E-04	12,99		Seed weight 34-8
	A5	AX-90388828	16	54,65	7,31E-04	13,29	25,23	N
		AX-90385356	16	56,94	1,31E-03	11,95		N
APM	A1	AX-90519720	12	8,90	1,38E-03	11,74	11,73	N
	A2	AX-90519720	12	8,90	5,42E-04	14,08	24,62	N
		AX-90443142	4	90,93	2,55E-03	10,55		Plant height 5-4
	A3	AX-90470288	2	21,30	4,05E-05	20,63	64,80	N
		AX-90508139	2	20,61	1,50E-04	17,34		N
		AX-90318278	2	20,80	3,17E-04	15,52		N
		AX-90489736	2	20,97	3,17E-04	15,52		N
		AX-90463399	10	78,10	3,89E-04	15,03		N
		AX-90375904	16	79,42	4,49E-04	14,69		N
		AX-90422780	2	103,93	4,95E-04	14,46		Plant height 42-1
	A4	AX-90315362	2	103,96	4,95E-04	14,46		Plant height 42-1
		AX-90341953	14	13,62	6,72E-04	13,48	13,47	Plant height 33-6
		AX-90503086	14	13,82	1,16E-03	12,23		Plant height 33-6
	A6	AX-90459046	11	19,28	1,33E-03	11,67	11,67	N
		AX-90305270	11	19,28	1,33E-03	11,67		N
	C	AX-90519720	12	8,90	1,16E-03	12,04	12,05	N
PEG	A1	AX-90413157	1	111,59	8,16E-04	13,06	13,06	N
		AX-90465921	1	111,58	8,16E-04	13,06		N
	A2	AX-90519720	12	8,90	5,42E-04	14,08	14,07	N
	A4	AX-90328966	5	59,94	3,38E-04	14,75	67,78	N
		AX-90437842	5	59,94	3,38E-04	14,75		N
		AX-90312028	11	69,03	5,28E-04	13,73		N
		AX-90420509	11	46,37	6,39E-04	13,30		N
		AX-90452125	18	11,37	7,14E-04	13,05		N
		AX-90339195	11	70,26	7,45E-04	12,95		N

(conclusão)

Caract	A	SNP	Crom	cM	P-valor	Var(%)	Var total	QTLs reportados ^a
PEG	A5	AX-90414653	17	6,15	8,65E-04	12,83	12,82	N
	C	AX-90354480	13	59,98	1,00E-03	12,39	12,39	N
NDM	A2	AX-90399877	2	10,59	1,28E-04	17,16	85,52	N
		AX-90373426	2	10,57	1,28E-04	17,16		N
		AX-90480549	2	10,63	1,28E-04	17,16		N
		AX-90365673	15	21,60	1,46E-04	16,84		Pod maturity 34-4
		AX-90493537	15	21,67	4,15E-04	14,40		Pod maturity 34-4
		AX-90511342	9	44,41	7,02E-04	13,19		Pod maturity 16-3
		AX-90448925	9	44,49	7,02E-04	13,19		Pod maturity 16-3
		AX-90394990	9	44,71	7,02E-04	13,19		Pod maturity 16-3
		AX-90440798	9	44,31	7,02E-04	13,19		Pod maturity 16-3
		AX-90373212	9	44,39	7,02E-04	13,19		Pod maturity 16-3
		AX-90378698	9	44,78	7,02E-04	13,19		Pod maturity 16-3
		AX-90493200	5	74,37	7,03E-04	13,19		N
		AX-90324440	5	74,41	7,03E-04	13,19		N
		AX-90406208	13	59,02	8,78E-04	12,69		N
		AX-90336473	18	129,52	9,76E-04	12,45		N
	A3	AX-90513187	7	1,32	1,34E-03	16,63	16,62	N
	A4	AX-90512489	8	43,00	9,98E-04	12,89	12,88	N
		AX-90491672	8	43,06	9,98E-04	12,89		N
		AX-90314156	8	43,08	9,98E-04	12,89		N
	A5	AX-90309859	18	44,12	2,31E-04	16,12	30,02	Pod maturity 16-2
		AX-90472096	18	45,67	3,75E-04	14,97		Pod maturity 16-2
		AX-90412903	18	45,67	3,75E-04	14,97		Pod maturity 16-2
		AX-90307081	18	7,48	5,90E-04	13,90		N
	C	AX-90336473	18	129,55	2,78E-04	15,65	42,60	N
		AX-90307081	18	7,48	5,05E-04	14,24		N
		AX-90320156	18	7,47	7,96E-04	13,18		N
		AX-90309671	18	7,49	7,96E-04	13,18		N
		AX-90365673	15	21,60	9,75E-04	12,71		Pod maturity 34-4
AIPV	A2	AX-90415118	5	86,82	2,84E-04	15,57	29,10	N
		AX-90420494	15	11,85	6,73E-04	13,54		N
		AX-90424403	15	11,85	6,73E-04	13,54		N
		AX-90341877	15	11,74	9,49E-04	12,75		N
		AX-90485117	5	87,70	1,01E-03	12,60		N
	A4	AX-90311625	9	2,71	4,18E-04	14,85	27,53	N
		AX-90449240	9	2,71	4,18E-04	14,85		N
		AX-90452535	9	2,71	4,18E-04	14,85		N
		AX-90443596	8	99,60	1,05E-03	12,69		N
	A5	AX-90482843	16	12,30	7,05E-04	13,37	13,37	N
	A6	AX-90443596	8	99,60	5,44E-04	13,78	13,78	N
	C	AX-90415118	5	86,82	5,61E-04	14,10	14,09	N
		AX-90522778	5	87,02	9,34E-04	12,91		N
APF	A2	AX-90519720	12	8,906	2,57E-04	16,00	15,99	N
	A5	AX-90506630	11	33,90	1,55E-03	11,50	11,50	N
	C	AX-90415118	5	86,82	5,61E-04	14,10	14,09	N
		AX-90522778	5	87,02	9,34E-04	12,91		N
TO	A2	AX-90468166	13	49,89	1,54E-04	16,49	28,71	Seed oil 37-7
		AX-90514908	13	49,89	1,54E-04	16,49		Seed oil 37-7
		AX-90503862	9	55,13	1,00E-03	12,22		Seed oil 37-10
	A5	AX-90386714	5	1,184	4,03E-05	19,94	34,05	Seed oil 36-1
		AX-90333512	18	123,74	4,65E-04	14,12		N
	C	AX-90468166	13	49,89	2,69E-04	15,33	43,76	Seed oil 37-7
		AX-90514908	13	49,89	2,69E-04	15,33		Seed oil 37-7
		AX-90503862	9	55,13	4,00E-04	14,42		Seed oil 37-10
		AX-90321935	4	80,13	4,76E-04	14,02		N

4. IDENTIFICAÇÃO DE REGIÕES DO GENOMA DIFERENTEMENTE AFETADAS PELA SELEÇÃO EM POPULAÇÕES DE SOJA VISANDO O AUMENTO DA BASE GENÉTICA DA CULTURA EM PROGRAMAS DE MELHORAMENTO DO BRASIL

RESUMO

Considerando que a caracterização das variações genéticas em populações envolvidas em programas de melhoramento é um processo crucial para o entendimento da dinâmica das mudanças ocorridas no genoma devido aos processos de seleção, e que genótipos de origens diversas resultam em uma maior diversidade, estes materiais podem ser utilizados em programas de melhoramento para a ampliação da base genética com introdução de alelos favoráveis. Neste estudo, foi verificada a presença de regiões genômicas diferencialmente afetadas entre uma população de genótipos de origens diversas (*Plant Introductions*-PI) e uma população composta de genótipos de origem brasileira (BR). Para isto, diferentes abordagens foram utilizadas: estimacão da diferenciação populacional nas regiões do genoma por meio do F_{ST} (índice de fixação), XP-CLR (*Cross-population composite likelihood ratio test*) por meio da identificação de regiões do genoma com redução de diversidade nucleotídica, e a presença de blocos de haplótipos. Em ambas as populações (PI e BR) foram observadas uma baixa diversidade genética. Entretanto a população PI apresentou uma maior riqueza alélica do que BR. Regiões diferencialmente selecionadas foram identificadas entre a população BR e PI, demonstrando que essas passaram por processos de seleção divergente. As regiões com sinais de seleção nas populações apresentaram uma grande quantidade de QTLs sobrepostos envolvidos no controle de caracteres de importância para a cultura. No total, 13 regiões genômicas apresentaram sinais de seleção na população BR sugerindo relação com o controle de funções importantes na adaptação as condições ambientais brasileira. Adicionalmente, 18 regiões genômicas na população PI podem representar fontes para a incorporação de novos alelos de importância agrônômica nos programas brasileiros de melhoramento de soja.

Palavras-chave: Incorporação de alelos; *Plant introductions*; Sinais de seleção; Caracterização genética

ABSTRACT

Considering that characterization of genetic variations in populations involved in breeding programs is a crucial process for understanding the dynamics of genomic changes due to selection processes, and that genotypes from different regions harbor great genetic diversity, they may be used in breeding programs in order to increase the genetic base with the introduction of new favorable alleles. In this study, we identified differentially affected genomic regions between a population of genotypes from diverse origins (*Plant Introductions*-PI) and a population of Brazilian genotypes (BR) using three approaches: estimation of population differentiation in the genome by F_{ST} (fixation index) and the XP-CLR approach (cross-population composite likelihood ratio test), identifying genomic regions with reduced nucleotide diversity and the presence of haplotype blocks. Low genetic diversity was observed in both populations (BR and PI). However, the PI population showed a higher allelic richness than BR. Differentially selected regions were identified between BR and PI populations, demonstrating that they underwent divergent selection processes. The regions with signals of selection in the populations presented a large amount of overlapped QTL involved in the control of important traits of the crop. In total, 13 genomic regions showed signs of selection in the BR population, suggesting relation to the control of important

functions to the adaptation to environmental conditions of Brazil. Additionally, 18 genomic regions in the PI population may represent sources for the incorporation of new alleles of agronomical importance in Brazilian soybean breeding programs.

Keywords: Introgression of alleles; Plant Introductions; Signals of selection; Genetic characterization

4.1. Introdução

A soja [*Glycine max* (L.) Merrill] é considerada a cultura de maior importância mundial na atualidade devido seu elevado conteúdo de proteína e óleo. O Brasil, até o momento, é o segundo maior produtor de soja do mundo, perdendo apenas para os Estados Unidos [1]. No Brasil, a cultura é responsável por 48% do total da produção de grãos, sendo a cultura de maior destaque pelo seu número crescente de produção e exportação [2]. Em 2017, 51,60 milhões de toneladas de soja foram destinados às exportações, correspondendo a US\$25,40 bilhões de dólares [2]. Avaliando-se o período de 2005 a 2017 a produção de soja aumentou de 51,12 para 114 milhões de toneladas, um acréscimo considerável de 123% [1,2]. Na safra de 1976/1977a produtividade da soja no Brasil foi de 1,748 t ha⁻¹, sendo que na safra de 2016/2017 foi de 3,364 t ha⁻¹, portanto, houve um aumento de 92,44% [2]. Este aumento relativo da produtividade se deve principalmente da combinação de dois fatores, boas práticas de cultivo e manejo e o desenvolvimento de cultivares mais produtivas, que se desenvolvem bem mesmo em ambientes e condições adversas.

Apesar da grande importância do melhoramento genético de soja para o Brasil, esse é relativamente recente. Apenas em meados de 1940 houve o lançamento da primeira cultivar brasileira, resultante do processo de melhoramento de soja adaptada as regiões do Brasil[3]. A partir do lançamento da primeira cultivar de soja, juntamente com a expansão das áreas de produção, a cultura foi ganhando seu destaque adentro no país. Entretanto, com os avanços nos patamares de produtividade propiciados pelo melhoramento genético houve concomitantemente, perdas na diversidade genética da cultura[4]. Hiromoto e Vello [4] relatam que apenas quatro ancestrais (CNS, Roanoke, Tokyo and S-100) contribuíram com 48,2% da base genética das 69 cultivares brasileiras indicadas para o cultivo em 1986. Isto também foi verificado em 444 materiais lançados entre o período de 1971-2009, na qual 14 ancestrais foram responsáveis por 92,4% da base genética[5]. Desta forma, devido o processo intenso de seleção artificial e natural, é esperado que sinais fossem deixados no genoma. Esses sinais de seleção são resultantes das alterações ocorridas nos padrões de variação

genética do genoma. A caracterização dessas variações genética em populações envolvidas em programas de melhoramento é um processo crucial para o entendimento da dinâmica das mudanças ocorridas no genoma devido aos processos de seleção [6].

A identificação das regiões do genoma afetadas pela seleção permite inferir sobre as funcionalidades dessas regiões e do envolvimento delas no controle de caracteres de importância agrônoma [7–9]. Detectar os sinais de seleção pode revelar segmentos compostos de genes e mutações com grandes efeitos fenotípicos, mesmo que não estejam mais segregando [10], o que impossibilita a detecção pelas abordagens de mapeamento [11]. Em um segmento do genoma contendo sinais de seleção é possível verificar alguns traços indicativos. Sinais de seleção apresentam em geral alterações nas frequências de alelos flancueadores, causando uma redução da variabilidade local, distorção no padrão da variação genética, maior desequilíbrio de ligação e presença de blocos de haplótipos [12–14].

Com o propósito de identificar sinais de seleção no genoma, várias metodologias foram desenvolvidas, sendo que as mais utilizadas são: índice de fixação (F_{ST}) [15], diversidade de nucleotídeos (π) [16], *integrated haplotype score* (iHS) [17], *cross-population extended*

haplotype homozygosity (XP-EHH) [18,19] e *cross-population composite likelihood ratio test* (XP-CLR) [12,20]. As metodologias propostas se diferenciam na forma de detecção dos sinais e nos pressupostos e, portanto, podem ser consideradas como complementares. Assim, cada teste para identificação de regiões sobre seleção pode fornecer resultados diferentes para o mesmo conjunto de dados genômicos e, provavelmente, nenhum deles fornecerá uma conclusão definitiva sobre as hipóteses seletivas [14,21]. Assim, devido os traços distintos deixados no genoma pela presença de sinais de seleção, uma combinação de estratégias múltiplas pode ser uma abordagem robusta, melhorando o poder de detecção [14].

Desta forma, neste estudo os objetivos foram: (i) identificar os padrões de variações genéticas de uma população representada por cultivares brasileiras comparadas a de uma população de genótipos exóticos de diferentes origens geográficas (*Plant Introductions*, PI), (ii) verificar regiões diferencialmente afetadas nas populações, (iii) identificar regiões candidatas contendo sinais de seleção e verificar qual o envolvimento dessas regiões no controle do fenótipo e, portanto, qual sua contribuição no melhoramento genético da soja. Para isto utilizou-se da genotipagem via plataforma da *Affymetrix (180 K Axiom® Soybean Genotyping Array)* e da combinação de várias estratégias de identificação de sinais de seleção para redução de falsos positivos.

4.2. Material e métodos

4.2.1. Material vegetal e genotipagem

Para o estudo foi elaborado um painel contendo 80 PIs (*Plant Introduction*) de soja, selecionadas em estudo prévio realizado por Mulato et al. [22], compreendendo 80 genótipos exóticos de diferentes origens geográficas e 15 cultivares brasileiras, incluindo 9 cultivares comerciais brasileiras: IAC100, Conquista, CD 215, BMX Potência, A7002, Vmax, Sambaíba, Pintado e Paranagoiania, 2 linhagens do programa de melhoramento de soja da Universidade Estadual Paulista “Júlio de Mesquita” (UNESP- Jaboticabal): JAB 00-05-6/763D e JAB 00-02-2/2J3D, 4 linhagens do programa de melhoramento de soja da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ): LQ1050, LQ1505, LQ1421 e LQ1413. Os 95 genótipos utilizados para o estudo já foram averiguados quanto a diversidade genética para caracteres agromorfológicos por Marconato et al. [23]

A extração de DNA das folhas dos 95 genótipos foi realizada pelo protocolo CTAB [24]. Através da quantificação do DNA via aparelho Quantifluor®, as amostras foram diluídas para a concentração de 100 ng μL^{-1} . Posterior a extração, o DNA diluído foi enviado para a genotipagem utilizando a plataforma da *Affymetrix (180 K Axiom® Soybean Genotyping Array)*.

Para garantir a qualidade dos dados, os SNPs resultantes da genotipagem foram filtrados seguindo as recomendações do Software *Axiom® Analysis Suite*. Posteriormente, foram eliminados os marcadores com mais de 10% de dados perdidos. Além disso, apenas o SNPs com MAF (*Minimum Minor Allele Frequency*) maior que 0,005, ou seja, SNPs com frequência menor de 5% foram utilizados. Após a filtragem dos dados, os SNPs com dados perdidos foram submetidos à imputação de dados no software Beagle 4.1. [24].

4.2.2. Diversidade genética

Os parâmetros de diversidade populacional, riqueza alélica, heterozigosidade média esperada [25] foram estimados utilizando o pacote *hierfstat*[27] desenvolvido para o software R [28]. Também foi estimado o conteúdo de informação polimórfica (PIC) médio dos marcadores em cada população.

4.2.3. Screening de indicativos de sinais de seleção

Evidências de regiões do genoma sobre seleção diferencial foram investigadas entre as 80 PIs e os 15 genótipos brasileiros utilizando-se a abordagem *cross-population composite likelihood ratio test* (XP-CLR) [12]. Este método é utilizado na comparação da distribuição das frequências alélicas entre duas populações [12]. Utilizaram-se janelas de 0,05 cM, com passos de 100pb em todo o genoma. O número máximo de SNPs testados em cada janela foi de 100. Além disso, o índice de fixação e de diferenciação da população, conhecido como F_{ST} , foi calculado utilizando o pacote do R *hierfstat*[29]. Para a identificação das assinaturas de seleção recomenda-se a utilização de abordagens complementares [14]. Desta forma, também foi verificado se houve redução de diversidade nucleotídica ao longo do genoma (π PI/ π BR) através do *software* VCFtools [30]. Para averiguar a participação das regiões encontradas envolvidas no processo de seleção, estas foram buscadas na base de dados do SoyBase (<https://www.soybase.org/>) [31] e verificou-se a presença de QTLs.

4.2.4. Estrutura de blocos de haplótipos

Blocos de desequilíbrio de ligação também são considerados como indicativos da presença de regiões sobre seleção [32]. Assim, o DL foi mensurado utilizando-se a medida r^2 , que considera o histórico de recombinação da população e as mutações [33,34]. Foi calculado o DL para as 80 PIs e para os 15 genótipos brasileiros em separado, a fim de identificar possíveis regiões envolvidas nos processos de seleção positiva e adaptativa nos materiais brasileiros, bem como regiões diferencialmente selecionadas nas PIs. Foram utilizados critérios previamente descritos por Gabriel et al. [35] no *software* Haploview para estimação do tamanho dos blocos de haplótipos em cada uma das populações.

4.3. Resultados e discussão

4.3.1. Diversidade genética

A partir da genotipagem dos 95 genótipos pela plataforma da *Affymetrix* (180 K *Axiom® Soybean Genotyping Array*) foram obtidos, após a filtragem e imputação, 12133 SNPs distribuídos nos 20 cromossomos da soja. Parâmetros genéticos foram estimados nas populações de genótipos exóticos (denominada de população PI) e de genótipos de origem do

Brasil (população denominada de BR) separadamente. Através da estimação da diversidade gênica (H_e) e pelo conteúdo de informação polimórfica (PIC) dos marcadores em cada população é possível observar que houve uma pequena redução da diversidade genética na população PI ($H_e=0,24$ e $PIC= 0,62$) quando comparadas à população BR ($H_e=0,29$ e $PIC=0,63$) (Tabela 1). Adicionalmente, cerca de 1200 marcadores SNPs apresentaram-se monomórficos na população BR e, apesar disso, a diversidade genética ainda foi maior. Contudo, a diversidade gênica encontrada, via os valores de H_e , neste estudo, foi relativamente baixa para ambas as populações.

Estudos analisando as cultivares brasileiras encontraram valores de H_e de 0,55 [36] e 0,38[37]. Não obstante, Mamidi et al.[38] avaliaram linhagens avançadas provenientes de programas de melhoramento dos Estados Unidos e observaram que a diversidade gênica (H_e) variou de 0,19 a 0,50, indicando que esta é dependente da população. A variação dos valores de H_e decorrente aos aspectos específicos de cada população evidencia-se pelo resultado apresentado por Tomicic et al. [39], no qual populações de graus distintos de melhoramento apresentaram discrepância nos valores de H_e . Tomicic et al. [39] verificaram que uma população contendo genótipos ancestrais, composta de materiais pouco melhorados, apresentou diversidade gênica mais alta ($H_e = 0,62$) quando comparada a população de genótipos melhorados ($H_e = 0,47$) [39]. A redução da diversidade gênica em populações melhoradas também foi observada por Wen et al.[13], no qual linhagens melhoradas apresentaram valores de H_e próximas a observadas neste estudo ($H_e=0,22$). Além disso, a baixa diversidade em materiais brasileiros foi verificada em vários estudos, sendo resultante do histórico de melhoramento recente e pelo número reduzido de ancestrais responsáveis por grande parte da variação genética [3–5,40].

Embora a diversidade gênica tenha se apresentado maior na população BR, a riqueza alélica foi maior na população PI, indicando que nessa população há um maior número de alelos divergentes nos locos. Isso reflete que a população PI apresentou um maior número de locos polimórficos (1200 a mais do que a população BR). PIs são utilizadas em programas de melhoramento com o objetivo da ampliação da base genética de uma população com a introdução de novos alelos, entretanto, neste estudo foi verificado que apesar da riqueza alélica da população de PIs ter sido maior comparada a população BR, as PIs apresentaram baixa diversidade genética (Tabela 1). A baixa diversidade em PIs já foi evidenciado em outros trabalhos [41,42]. Vários fatores podem acarretar redução da diversidade gênica, como o uso reduzido de parentais em programas de melhoramento ou mesmo a introgressão de alelos limitante pela utilização de genótipos redundantes [13]. Acrescenta-se que a seleção

pode ser um fator responsável para redução da diversidade genética em alguns locais do genoma. A seleção provoca alterações nas frequências alélicas nas regiões vizinhas à região sob seleção [13], o que reflete em uma redução da média da diversidade gênica (H_e) da população, como demonstrado por Tomicic et al. [39]. Contudo, as PIs utilizadas neste estudo são classificadas pelo GRIN (*Germplasm Resources Information Network*) (www.ars-grin.gov) como materiais cultivados, não apresentando demais detalhes sobre o grau de melhoramento desses genótipos, o que sugere que a redução da diversidade pode também ser resultante do processo de domesticação e que as PI passaram por algum processo de seleção nas diversas regiões geográficas de amostragem. Já os materiais brasileiros, são genótipos os quais passaram por processos de melhoramento genético com ciclos de recombinações seguidos da seleção artificial. Desta forma, ambas as populações passaram por processos de desenvolvimento distintos, que de forma direta ou indiretamente envolveu etapas de seleção, ocasionando em uma redução da diversidade genética [13].

4.3.2. Diferenciação entre as populações

O cálculo do F_{ST} tem sido amplamente utilizado com potencial para identificação de sinais de seleção no genoma. Em geral, ele auxilia a revelar a diferenciação entre duas populações, pois em muitos casos a seleção aumenta o grau de diferenciação alélica entre populações [9]. Desta forma, o F_{ST} foi estimado para cada SNP entre as populações PI e BR. Considerando os valores de F_{ST} para os 12133 SNPs, a média de diferenciação entre as populações foi de $F_{ST} = 0,08$, sendo um valor relativamente baixo [43]. A baixa diferenciação entre a maioria dos locos de SNPs analisados podem indicar que ambas as populações são homogêneas, e/ou também que elas apresentam seleção neutra ou direcional nas mesmas regiões do genoma [44]. Contudo, a diferenciação baixa entre as populações deste estudo já era esperada pela relativa base genética estreita da cultura *per se* [45] e, também, pela baixa diversidade gênica apresentada por ambas as populações (Tabela 1). Além disso, devido participação de materiais de origem dos Estados Unidos, China e Japão na composição da população PI e na genealogia dos genótipos resultantes dos programas de melhoramento de soja do Brasil [5], bem como na população denominada de BR, parte dos alelos dos genótipos de origem brasileira são provenientes desses materiais, o que resultou em baixa diferenciação entre as populações do estudo.

Para verificar a diferenciação do genoma, entre a população BR e a PI, como um indicativo da presença de sinais de seleção, foi considerado um *threshold* de $F_{ST} \geq 0,3$, sendo

um valor de grande diferenciação entre populações [43]. Autores como CASA et al. [46] e KAYSER et al.[47] sugerem para que os valores de F_{ST} sejam significativamente associados aos sinais de seleção, estes devem ser o dobro do valor médio encontrado e, portanto, o valor utilizado para este estudo é adequado. Desta forma, foram identificados 168 SNPs que estão localizados em 34 regiões dos cromossomos da soja (Figura 1 e Tabela 2), com exceção dos cromossomos 3, 9, 12, 15, 18 e 19 que não foram identificados nenhum sinal de seleção. Song et al. [48] observaram 42 regiões envolvidas nos processos de seleção impostos nos genótipos Norte Americanos utilizando F_{ST} igual a 0,1. As regiões associadas às assinaturas de seleção correspondem à de 22 Mbp e, portanto, à aproximadamente 2,2% do genoma da soja que contém 1,1 Gbp [49,50]. Considerando os 14 cromossomos, nos quais foram encontrados sinais de seleção, o cromossomo 11 apresentou a maior quantidade de regiões com sinais de seleção.

Como o cálculo do F_{ST} não utiliza uma população como referência e a outra como testada, e sim apenas compara a diferenciação entre os alelos de ambas as populações de uma vez, as regiões identificadas pelo F_{ST} podem não estar diretamente relacionadas ao processo de seleção direcional específica para o ambiente do Brasil[44], podendo estas, serem relacionadas à seleção imposta na população PI. Assim, as regiões encontradas podem estar associadas aos processos de domesticação, adaptação local ou mesmo como a seleção imposta nos programas de melhoramento [48]. De forma a identificar sinais de seleções específicos de cada população, ou seja, diferencialmente afetadas pelos processos de seleção, abordagens complementares foram utilizadas.

Além do cálculo do F_{ST} , para identificar potenciais sinais de seleção entre as PIs e os genótipos brasileiros, foi utilizado o método *cross-population composite likelihood ratio test* (XP-CLR)[51]. Essa análise vem demonstrando maior poder de detecção quando comparado aos métodos de Tajima's D, proposto por Tajima [52] e CLR desenvolvido por Nielsen et al.[51]. Além disso, apresenta-se como um método complementar ao cálculo do F_{ST} por combinar resultados de vários testes de regiões contínuas para redução da taxa de falsos positivos [53]. Além disso, ela permite detectar sinais de seleção devido a processos antigos de afunilamento genético, possibilita utilizar uma população como referência e definir qual a população será a testada. Logo, os resultados desse teste podem ser direcionados a revelar sinais de seleção referentes à população definida como testada [53]. Neste estudo, ambas as populações foram utilizadas como testadas, para verificar regiões diferencialmente selecionadas nas duas populações.

Geralmente, para análise de XP-CLR, utiliza-se de 1% a 0,5% dos valores mais altos para definição do *threshold* dos sinais de seleção [54,55]. A fim de captar regiões com alto grau de diferenciação e, portanto, contendo fortes sinais de seleção, neste trabalho foi utilizado menos de 0,5% dos valores mais altos. Assim, utilizou-se um *threshold* de XP-CLR ≥ 6 para captar os sinais de seleção na população BR, e de XP-CLR ≥ 8 na população PI. Considerando as regiões contendo sinais de seleção relativa à população BR, no total foram detectadas 26 regiões do genoma da soja (Tabela 3 e Figura 2A). Nesse teste, os cromossomos 1, 3, 5, 8, 9 e 19 não apresentaram nenhuma região com valores significativos de XP-CLR relativos à população BR. Já para a população PI, foi verificado 20 regiões distribuídas em 14 cromossomos da soja, com exceção dos cromossomos 5, 9, 12, 15, 17 e 18 (Tabela 4 e Figura 2B). Como podemos verificar, apenas no cromossomo 5 e 9 nenhum sinal foi detectado. As regiões apresentadas na população BR correspondeu a 5,04 Mpb e na população PI à 2,01 Mpb, compreendendo cerca de 0,5% e 0,2% do genoma da soja [49,50], respectivamente. Wen et al.[13] utilizando o método XP-CLR para detecção de sinais de seleção entre 342 materiais não melhorados e 1062 melhorados, apresentaram 1% do genoma com sinais de seleção. No entanto, os autores consideram 1% dos valores mais altos de XP-CLR como *threshold*.

O método utilizado pela análise XP-CLR tem sido amplamente aplicado em diversas culturas como arroz [56,57], pepino [58], colza [55], espinafre [59] e milho [60]. Em soja, o método desenvolvido por Chen et al.[12], já foi avaliado por Wen et al.[13] e por Zhou et al. [54], sendo que esses identificaram 125 e 109 sinais de seleção resultantes do melhoramento nos 20 cromossomos da soja, utilizando como *threshold* valores acima de 9,49 (1%) e 4,6 (5%), respectivamente. Os valores encontrados neste trabalho estão mais próximos do encontrado por Zhou et al. [54], podendo ser explicado pelo menor número de genótipos utilizado quando comparados ao de Wen et al. [13]. Além disso, o grau de diferenciação entre as populações sobre análise é um fator importante para o sucesso da análise XP-CLR [53]. Por Wen et al. [13] ter utilizado 4,64 vezes mais genótipos do que no trabalho de Zhou et al. [54] foi possível captar melhor a diferenciação entre os grupos e, portanto, captou um maior número de sinais de seleção. Desta forma, verificamos que a quantidade de indivíduos utilizados é um fator importante para as análises de detecção de assinatura de seleção, comprometendo o sucesso das análises, assim como já foi verificado para GWAS [54]. Entretanto, apesar do número reduzido de genótipos utilizados neste trabalho e a baixa diferenciação entre os grupos, evidenciada pelo F_{ST} médio, ainda foi possível verificar sinais de seleção. Quando analisados os sinais de seleção encontrados pela análise XP-CLR por

Zhou et al. [54], verificou-se que a população BR também apresentou cinco regiões observadas pelo estudo, sendo as regiões dos cromossomos 9 (72 cM), 11(0,9 cM), 14 (93 cM), 18 (3 cM) e 20 (5,64 cM). A região 93 cM do cromossomo 14 apresentada por Zhou et al. [54] e na população BR, também foi identificada associada à seleção na população PI. Essa região apresenta vários QTLs envolvidos no controle de componentes da semente [54] e demonstra que populações distintas podem apresentar pressões de seleção similares para caracteres importantes.

Outro fator importante para detecção dos sinais de seleção é o grau de melhoramento da cultura estudada [13]. Populações que experimentaram mais ciclos de recombinação natural e artificial quando comparadas aos parentais ou populações pouco melhoradas apresentam um número de sinais de seleção ou de domesticação maior e com valores de XP-CLR bem mais elevados, como pode ser observados em milho [61]. Em espinafre, espécie pouco melhorada, fica bem clara a redução dos valores de XP-CLR (*threshold* de XP-CLR ≥ 1) [59]. Esta variação entre as populações, também pode ser observado nos cromossomos. Cromossomos com regiões que controlam importantes caracteres podem revelar um maior número e indicativos mais fortes da presença de sinais de seleção [13]. Assim, como apresentado neste estudo (Tabela 2-6), cada cromossomo apresenta números variados de sinais de seleção, sendo o mesmo apresentado por diversos autores [13,54,56,58,61].

4.3.3. Redução da diversidade nucleotídica

A identificação de sinais de seleção no genoma é um processo complexo. Os sinais de seleção são identificados pelos indicativos que esses deixam no genoma, como a presença da redução da diversidade nucleotídica na região [13,12,14]. Assim, a fim de fornecer evidências suficientes para a presença destes sinais nos materiais brasileiros e na população de PIs, também foi verificada se há regiões no genoma com reduzida diversidade quando comparada ambas as populações. Para isto foi utilizado o cálculo da diversidade nucleotídica ao longo do genoma (π) [16]. Hufford et al. [61], Saïdou et al. [62], He et al. [63], Benjelloun[64], Qi et al.[58], Meyer et al. [57], Li et al.[65] utilizaram o cálculo da diversidade nucleotídica para auxiliar na identificação de sinais de seleção. Assim, neste estudo o *threshold* para definição de uma região com significativa redução de diversidade nucleotídica utilizou-se 1% dos valores mais altos de *ratio* de π PIs/ π BR ($\geq 2,60$) para definir regiões com reduzida diversidade nucleotídica na população BR em comparação a população

PI, e $\pi_{BR/\pi_{PI}} (\geq 4,98)$ para verificar o contrário, identificando redução da diversidade nucleotídica na população PI [54,55,66].

Pelo cálculo da diversidade nucleotídica em cada uma das populações observou-se que na população BR, em todos os cromossomos, foi maior do que na população PI (Tabela S1). Na população BR, o π médio foi de 0,0030, variando de 0,0025 (cromossomo 10) a 0,0037 (cromossomo 17). Já na população PI, o π médio dos 20 cromossomos foi de 0,0025, variando de 0,0021 (cromossomo 10) a 0,0029 (cromossomo 16). A maior diversidade nucleotídica presente em todos os cromossomos da população BR já era esperada devido à maior diversidade genética encontrada nessa população (Tabela 1). Contudo, apesar do fato dessa população apresentar maiores valores de π , em alguns segmentos do genoma foi possível verificar que houve uma redução da diversidade nucleotídica, apresentando indicativos de possíveis sinais de seleção, sendo também verificado para a população PI.

No total foram identificados 118 SNPs (Figura 2C), os quais apresentaram uma redução da diversidade nucleotídica na população BR, quando comparada a população PI. Os SNPs identificados abrangem 52 regiões distribuídas em 18 cromossomos (Tabela 5), com exceção do cromossomo 1 e 19, e que correspondem a 9,24 Mbp do genoma da soja [49,50]. Além disso, 37 regiões (Tabela 6), que correspondem a 87 SNPs (Figura 2D), demonstraram redução de diversidade na população PI.

Os valores de π encontrados neste estudo foram relativamente altos quando comparados aos apresentados para soja por Qi et al. [58] e Valliyodan et al. [67], demonstrando que os locos avaliados neste estudo apresentaram maior diversidade do que os avaliados pelos autores e, portanto, sugerindo o mesmo para as populações. Maldonado dos Santos et al. [3] ao analisarem as sequências do genoma de 28 cultivares brasileiras verificaram segmentos contendo redução de diversidade nucleotídica no cromossomo 7, próxima da região encontrada no mesmo cromossomo neste estudo na população BR (Tabela 5). Nenhum sinal de seleção apresentado por Maldonado dos Santos et al. [3] foi verificado na população PI, indicando que regiões divergentes foram selecionadas nas diferentes populações, já que ambas são de origem diferente.

4.3.4. Estrutura de blocos de haplótipos

A presença de sinais de seleção pode causar drásticas mudanças na distribuição dos polimorfismos do genoma, na frequência alélica e no DL em volta destas regiões [12–14]. Regiões envolvidas nos processos de seleção estão sobre a ação de constantes recombinações,

sendo que isso acontece quando alelos vizinhos estão em desequilíbrio de ligação. A presença do desequilíbrio de ligação entre alelos vizinhos define um haplótipo, o qual pode permanecer na população até que por algum processo esta ligação entre os alelos vizinhos seja quebrada [44]. Assim, um excesso de desequilíbrio de ligação em uma região do genoma, caracterizada pela presença de haplótipos longos podem ser uma evidência da presença de sinais de seleção [32]. Desta forma, neste estudo foi verificada a presença e extensão dos blocos de haplótipos na população de PIs e nos materiais brasileiros, utilizando-se os critérios para formação de haplótipos definidos por Gabriel et al.[35].

Padrões distintos da extensão de haplótipos foram verificados nos dois conjuntos de genótipos, bem como nos diferentes cromossomos (Figura 4). Considerando todos os cromossomos, a média da extensão dos blocos de haplótipos nas duas populações analisadas foi de 458 kb para a população BR e 316 kb para a população PI. Sendo ambos os valores maiores do que encontrado para linhagens melhoradas de soja de origem dos Estados Unidos (245 kb) por Hyten et al. [68]. Esta diferença pode ser resultado da baixa cobertura do genoma apresentada pelo estudo ao analisar apenas 3 segmentos do genoma da soja [68] e/ou pelos processos evolutivos específicos ocorridos em cada população.

Estudos indicam que blocos de haplótipos que se estendem a longas distâncias podem surgir em resultado a diversos processos genéticos, sendo os principais referentes à seleção artificial e natural [48,66]. Na Figura 4 é possível observar que nos cromossomos 2, 4, 7, 8, 10, 11, 14, 15, 16 e 17 a extensão média dos haplótipos verificados para a população de origem do Brasil foi maior. Destacam-se os cromossomos 2, 4, 6, 7, 10 e 17, onde a média da extensão dos haplótipos verificada foi maior que 750 kb. Isto sugere que nesses cromossomos os processos de seleção podem ter sido mais fortes quando comparados nos demais materiais de origem diversa (população PI). Nos cromossomos 1, 9, 12, 19 e 20, a extensão dos haplótipos foi mais sutil nos materiais brasileiros (Figura 4). Esta variação dos haplótipos nos cromossomos da soja também foi verificada por Hyten et al. [68], o que evidencia que cromossomos diferentes, bem como populações distintas, possuem históricos de evolução e seleção distintos.

Segundo Hyten et al. [45] um aumento do tamanho dos haplótipos é um indicativo da redução da diversidade genética. Esse fenômeno foi observado pelos autores ao comparar a extensão dos blocos de haplótipos entre populações de linhagens elites do EUA e uma população ancestral. O mesmo foi verificado por Song et al. [69]. Deste modo, a fim de averiguar maiores indicativos da presença de sinais de seleção na população BR e PI, foi

verificado a presença das regiões identificadas em comum por pelo menos duas análises utilizadas (F_{ST} , π e XP-CLR) em blocos de haplótipos.

Na Tabela 7 são apresentadas as regiões identificadas em comum pelo menos em duas das análises utilizadas para identificação de indicativos de sinais de seleção relativos à população BR (F_{ST} , π PI/ π BR e XP-CLR), totalizando 18. As 18 regiões estão localizadas, com exceção da região dos cromossomos 12 e 20, em cromossomos que apresentaram média de blocos de haplótipos maiores que 750 kb, sugerindo que cromossomos com médias de haplótipos maior apresentam maior probabilidade da presença de sinais de seleção. No entanto, nos cromossomos 12 e 20, esses apresentaram indicativos da presença de sinais de seleção em apenas uma das análises e, portanto, não estão representados na Tabela 7. Considerando as 18 regiões distribuídas em 13 (2, 4, 5, 6, 10, 11, 12, 13, 14, 16, 17, 18 e 20) dos 20 cromossomos da soja, apenas 8 foram localizadas dentro de blocos de haplótipos. Os haplótipos identificados apresentam em média 196,71 kb de extensão, variando de 401,08 kb a 12,55 kb.

Para a população PI, 26 regiões apresentaram indicativos de sinais de seleção por mais de um teste (F_{ST} , π e XP-CLR), sendo que 8 foram localizadas dentro de blocos de haplótipos (Tabela 8). Esses blocos apresentam em média 122,68 kb, variando de 495,43kb a 8,03kb. Na Figura 4 observamos que no cromossomo 20, bem como no cromossomo 1, os blocos de haplótipos apresentaram extensão média de aproximadamente 500 kb, bem maior do que apresentada pela média dos cromossomos nesta população (316 kb). Logo sugere que esses cromossomos e fica evidente pela presença de 3 regiões contendo sinais de seleção em ambos os cromossomos, estão envolvidos no controle de caracteres importantes.

4.3.5. Regiões candidatas

Considerando que a soja apresenta base genética estreita [45], a incorporação de genótipos exóticos, ou também denominados de PIs, em programas de melhoramento tem sido utilizada como uma estratégia para ampliação da base genética da cultura [70–72]. E que apesar das PI apresentarem maior diversidade genética quando comparadas a materiais elite, essas ainda apresentam baixa diversidade genética em vista de outras espécies de plantas cultivadas [41,42]. Neste estudo foi reunido um conjunto de genótipos exóticos com origens diversas supondo que a distância geográfica dessas pode refletir um aumento da diversidade genética, já que essas passaram por processos divergentes de seleção, bem como de adaptação. As divergências genéticas das PIs foram comparadas às da população de cultivares

brasileiras, de forma que quanto mais geneticamente distinta e diversa for a população PI da população BR, as chances de identificar sinais de seleção divergentes, abrangendo um amplo leque de variação fenotípica torna-se maior. Desta forma, como foi utilizado da comparação entre a população formada pelo conjunto de materiais brasileiros (BR) e a população de PIs (PI), as regiões com sinais de seleção identificadas, tanto pelo F_{ST} , XP-CLR e pelo π *ratio*, podem refletir a seleção imposta pelos melhoristas especificamente para os ambientes do Brasil, que são divergentes dos demais países, devido aos aspectos climáticos e ecológicos, bem como regiões diferencialmente selecionadas nas PIs [39].

Estudo visando a identificação de sinais de seleção no genoma da soja específicos a um ambiente já foram realizados por Tomicic et al. [39] para o Centro-Leste Europeu. Sinais de seleção específicos para o Brasil foram verificados por Maldonado dos Santos et al. [3] comparando cultivares lançados recentemente (depois do ano de 2000) e os mais antigos (antes do ano de 2000). Esse estudo identificou 41 regiões que foram modificadas ao longo do tempo, entretanto, utilizaram apenas 28 genótipos para a divisão em duas sub-populações. Assim, neste estudo apresentamos uma abordagem complementar, pela comparação de regiões diferencialmente selecionadas entre genótipos de todos os continentes e os genótipos brasileiros.

Neste presente estudo, no total (XP-CLR e π) foram identificadas 78 regiões do genoma da soja, associadas a sinais de seleção na população BR e 57 regiões relativas aos processos de seleção na população PI. A maior quantidade de locos candidatos, ou seja, de locos contendo indicativos à presença de sinais de seleção, foram identificados pela abordagem *ratio* de π (52 regiões na população BR e 37 na população PI). O XP-CLR foi a abordagem que apresentou um menor número de regiões (26 regiões na população BR e 20 na população PI). Considerando as três análises para identificação de sinais de seleção envolvidos nos processos de seleção (F_{ST} , XP-CLR e π), a maioria das regiões do genoma com indicativos de sinais de seleção ocorreram apenas uma vez, fato que também ocorreu em demais estudo que abordaram o mesmo tema [3,6,10,14,39,55,63,65,73]. Algumas regiões foram apresentadas em mais de um teste, enquanto outras podem ser consideradas como falsos positivos ou não apresentaram evidências suficientes pelos testes utilizados, necessitando de mais investigações.

Destaca-se que a abordagem do F_{ST} apresentou 34 regiões, as quais podem estar relacionadas aos processos de seleção de qualquer uma das populações analisadas, já que esta abordagem não é direcional para uma população usada como referência e identifica apenas regiões com alto grau de diferenciação entre duas populações. Isso fica evidenciado pela

correspondência de algumas regiões com outros testes específicos para cada uma das populações (Figuras 5A e 5B).

Considerando as 78 regiões identificadas na população BR, apenas 18 foram verificadas em mais de um teste (F_{ST} , XP-CLR e π). Destaca-se que apenas três regiões foram relacionadas pelas três abordagens (F_{ST} , XP-CLR e π) (Figura 5A), sendo as regiões 23,06 cM do cromossomo 2, 55,48 cM do cromossomo 13 e 7,26-7,33 cM do cromossomo 14. Além dessas três regiões, cerca de seis outras foram identificadas contendo redução de diversidade e também apresentaram evidências de sinais de seleção pela análise XP-CLR (Tabela 7). Outras cinco regiões apresentaram coincidências contendo redução de diversidade também foram apresentadas pelo F_{ST} (Tabela 7). Além disso, para a população BR, quatro regiões apresentaram evidências de seleção pelo F_{ST} e XP-CLR.

Na população PI, 26 regiões foram reportadas por pelo menos dois testes (F_{ST} , XP-CLR e π) e estão representadas na Tabela 8. Observa-se que 5 regiões apresentaram redução da diversidade nucleotídica (π) e alta diferenciação entre as populações (F_{ST} , XP-CLR). Cinco regiões foram observadas pelo XP-CLR contendo redução de diversidade nucleotídica na população PI. Considerando π e os resultados do F_{ST} , 19 regiões tiveram correspondência por ambas às abordagens. Além disso, 14 regiões apresentaram alta diferenciação pelo F_{ST} e XP-CLR (Figura 5B).

Ao utilizar da comparação entre populações ancestrais e populações melhoradas, as regiões do genoma encontradas seriam decorrentes somente das alterações provocadas pelo melhoramento da população como verificado por outros autores [54,58,66,74–78]. Contudo, neste estudo utilizamos duas populações de origens diferentes. Logo, os sinais de seleção encontrados em cada uma das populações refletem as regiões diferencialmente afetadas em cada uma delas. As regiões identificadas diferencialmente afetadas nas populações deste estudo foram comparadas com os mapas genéticos disponibilizados no *Soybase* (www.SoyBase.org) para determinar se algum QTL foi já reportado para essas regiões, e, portanto, participam do controle da variação fenotípica de algum caráter. Os QTLs reportados pertencem a 14 categorias distintas, elucidadas na Figura 6, que incluem teor de proteína e óleo, altura de planta, maturidade, produtividade e resistência a estresses bióticos e abióticos. A fim de identificar com fortes indicativos de sinais de seleção, foram consideradas apenas as regiões em que apresentaram dois indicativos: redução de diversidade nucleotídica e/ou alta diferenciação da região ao comparar as populações, via uma das abordagens de F_{ST} ou XP-CLR e/ou localização da região dentro de um bloco de haplótipo.

Da mesma forma que algumas regiões foram apresentadas contendo sinais de seleção em comum entre este estudo com demais [3,54], é interessante destacar que 5 regiões (23,06-23,31 cM do cromossomo 2, 17 cM, 15-17,41 cM do cromossomo 4, 0,13 cM do cromossomo 10, 0,62 cM e 32,38-32,39 cM do cromossomo 11, 93,3 cM do cromossomo 14) foram reportadas em comum associadas a sinais de seleção nas populações PI e BR. Essas regiões foram verificadas separadamente das populações da presença de QTLs.

Considerando apenas as regiões contendo sinais de seleção distintos nas populações BR e PI, 13 e 18 regiões, respectivamente, estas provavelmente estão sob ação de diferentes pressões seletivas dentro das condições ambientais do local de origem. Além disso, essas regiões genômicas podem ser responsáveis pelo controle dos mesmos caracteres agrônômicos, como evidenciado pela presença QTLs na região.

Observou-se que na população BR, regiões candidatas apresentaram sobreposição com uma grande quantidade de QTLs envolvidos no controle de nematóides. O mesmo foi verificado por Maldonado dos Santos et al. [3] ao verificar as alterações no genoma decorridas do processo de melhoramento genético em 28 cultivares brasileiros. O nematóide de cisto da soja (NCS), *Heterodera glycines* Ichinoh, é considerado um dos principais problemas na cultura da soja no Brasil, podendo as perdas atingir 100%. Além das perdas ocasionadas na produtividade da soja, uma vez presente no solo, o nematóide de cisto da soja dificilmente é eliminado, necessitando-se de fontes genéticas de resistência. Na população BR, também foram identificadas regiões sobrepostas a QTLs relacionados à podridão de fitóftora (*Phytophthora sojae*) e síndrome da morte súbita (*Fusarium solani*) representados como a quarta categoria de QTLs mais reportados.

Na população PI, os 18 segmentos apresentam na maioria, como também apresentado na população BR, QTLs reportados para caracteres físicos e composição das sementes como, peso, altura, comprimento, volume e conteúdo de isoflavonóides. Sinais de seleção específicos para materiais melhorados da Europa também apresentaram, na maioria, QTLs para as mesmas categorias [39]. Isso demonstra a importância desses caracteres, e que apesar de ambientes diferentes, e, portanto, populações diferentes, regiões distintas do genoma estão envolvidas no controle de caracteres incluídos na mesma categoria.

A caracterização dos padrões de variação entre populações pode fornecer importantes revelações a respeito das dinâmicas envolvidas no processo de seleção artificial e natural. Regiões com sinais de seleção permitem inferências a respeito da funcionalidade daquela região, bem como da importância dela para o sucesso de programas de melhoramento [6]. Essas regiões geralmente já estão fixadas em algumas populações, o que não permite a

identificação via abordagem GWAS. Além disso, regiões como essas podem ser a chave para aumentar o sucesso de programas de melhoramento.

Os alelos identificados na população PI diferencialmente afetados quando comparada as da população BR, ou seja, que não apresentaram sinais de seleção na população BR, podem ser utilizados para o aumento da variabilidade genética das populações de melhoramento do Brasil, auxiliando no aumento do ganho genético. Além disso, os alelos encontrados na população BR, localmente adaptados ao Brasil, também podem ser utilizados em programas de melhoramento de outras regiões do mundo, como fonte de alelos favoráveis [79]. O envolvimento das regiões encontradas nas populações no controle de caracteres fica evidenciado pela presença de um grande número de QTLs co-localizados [39]. Entretanto, é necessário o desenvolvimento de estudos mais aprofundados para a investigação da contribuição dessas regiões em relação ao fenótipo.

Caracterizar as regiões genômicas e identificar as áreas diferencialmente afetadas em populações pode auxiliar o melhorista a entender as condições ambientais e definir alelos potenciais. Os resultados de estudos desta natureza auxiliam na definição das estratégias utilizadas nos programas de melhoramento. Ao comparar uma população melhorada e outra considerada como fonte de alelos favoráveis é possível verificar quais alelos estão diferencialmente selecionados e se a população de fonte de alelos apresenta variabilidade suficiente, contrastante e favorável para a introdução destes materiais nas etapas de hibridização. Neste estudo as PIs utilizadas apresentaram uma maior riqueza alélica, o que pode contribuir para o aumento da diversidade genética das populações de melhoramento do Brasil. Além disso, as regiões do genoma apresentadas na população PI, que não estão sob seleção na população BR, podem ser investigadas em outros genótipos de interesse, podendo estes serem utilizados como fonte de alelos favoráveis caso provada o envolvimento das regiões no controle de caracteres de interesse em estudos futuros.

4.4. Conclusão

Regiões diferencialmente selecionadas foram identificadas entre a população de genótipos brasileiros (BR) e de genótipos exóticos (PI), demonstrando que essas passaram por processos de seleção divergentes. Regiões observadas na população de PIs apresentam potencial a ser explorado para o aumento da variabilidade genética em programas de melhoramento do Brasil. Adicionalmente, regiões encontradas com sinais de seleção na população de genótipos do Brasil sugerem estarem envolvidas na seleção local.

REFERÊNCIAS

1. Food and Agriculture Organization. Accessed on december 2017 at www.fao.org
2. Companhia Nacional de Abastecimento. Accessed on december 2017 at www.conab.gov.br
3. Maldonado dos Santos JV, Valliyodan B, Joshi T, Khan SM, Liu Y, Wang J, et al. Evaluation of genetic variation among Brazilian soybean cultivars through genome resequencing. *BMC Genomics*. 2016;17: 110. doi:10.1186/s12864-016-2431-x
4. Hiromoto DM, Vello NA. The genetic base of brazilian soybean(*Glycine max*(L.) MERRILL) cultivars. 1986. pp. 295–306.
5. Wyszniński PT, Vello NA. The genetic base of Brazilian soybean cultivars: Evolution over time and breeding implications. *Genet Mol Biol*. 2013;36: 547–555. doi:10.1590/S1415-47572013005000041
6. Aslam ML, Bastiaansen JWM, Megens HJ, Crooijmans RPMA, Nasreen F, Blomberg LA, et al. Genome-wide candidate regions for selective sweeps revealed through massive parallel sequencing of DNA across ten turkey populations. *BMC Genet*. 2014;15: 1–10. doi:10.1186/s12863-014-0117-4
7. Pérez O'Brien AM, Utsunomiya YT, Mészáros G, Bickhart DM, Liu GE, Van Tassell CP, et al. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. *Genet Sel Evol*. 2014;46: 1–14. doi:10.1186/1297-9686-46-19
8. Biswas S, Akey JM. Genomic insights into positive selection. *Trends Genet*. 2006;22: 437–446. doi:10.1016/j.tig.2006.06.005
9. Nielsen R. Molecular Signatures of Natural Selection. *Annu Rev Genet*. 2005;39: 197–218. doi:10.1146/annurev.genet.39.073003.112420
10. Nimmakayala P, Levi A, Abburi L, Abburi V, Tomason YR, Saminathan T, et al. Single nucleotide polymorphisms generated by genotyping by sequencing to characterize genome-wide diversity, linkage disequilibrium, and selective sweeps in cultivated watermelon. *BMC Genomics*. 2014;15: 767. doi:10.1186/1471-2164-15-767
11. Ramey H, Decker J, McKay S, Rolf M, Schnabel R, Taylor J. Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics*. 2013;14. doi:10.1186/1471-2164-14-382
12. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. *Genome Res*. 2010;20: 393–402. doi:10.1101/gr.100545.109

13. Wen Z, Boyse JF, Song Q, Cregan PB, Wang D. Genomic consequences of selection and genome-wide association mapping in soybean. *BMC Genomics*. *BMC Genomics*; 2015;16: 671. doi:10.1186/s12864-015-1872-y
14. Randhawa IAS, Khatkar MS, Thomson PC, Raadsma HW. Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genet*. 2014;15: 1–19. doi:10.1186/1471-2156-15-34
15. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res*. 2002;12: 1805–1814. doi:10.1101/gr.631202
16. Nei M, Li W-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*. 1979;76: 5269–5273. doi:10.1073/pnas.76.10.5269
17. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol*. 2006;4: 0446–0458. doi:10.1371/journal.pbio.0040072
18. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. UKPMC Funders Group Genome-wide detection and characterization of positive selection in human populations. *October*. 2009;449: 913–918. doi:10.1038/nature06250.Genome-wide
19. Tang K, Thornton KR, Stoneking M. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol*. 2007;5: 1587–1602. doi:10.1371/journal.pbio.0050171
20. Nielsen R, Williamson S, Kim Y, Nielsen R, Williamson S, Kim Y, et al. Genomic scans for selective sweeps using SNP data Genomic scans for selective sweeps using SNP data. 2005; 1566–1575. doi:10.1101/gr.4252305
21. Lin K, Li H, Schlötterer C, Futschik A. Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. *Genetics*. 2011;187: 229–244. doi:10.1534/genetics.110.122614
22. Mulato BM, Möller M, Zucchi MI, Quecini V, Pinheiro JB. Genetic diversity in soybean germplasm identified by SSR and EST-SSR markers. *Pesqui Agropecuária Bras*. 2010;45: 276–283. doi:10.1590/S0100-204X2010000300007
23. Marconato MB, Pereira EM, Silva FM, Bizari EH, Pinheiro JB, Mauro AO, et al. Genetic divergence in a soybean (*Glycine max*) diversity panel based on agro-morphological traits. *Genet Mol Res*. 2016;15: 1–10. doi:10.4238/gmr15048980
24. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am J Hum Genet*. 2007;81: 1084–1097. doi:10.1086/521987

25. Nei Masatoshi. Genetic Distance between Populations. Mol Evol Genet Columbia Univ Press New York; 1987;
26. Wright S. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. Evolution (N Y). 1965;19.
27. Goudet J, Jombart T. Estimation and Tests of Hierarchical F-Statistics. R Core Team. 2015; 58. doi:10.2307/2411227>.
28. R Core Team: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2013.
29. Yang R-C. Estimating Hierarchical F-Statistics. Evolution (N Y). 1998;52: 950. doi:10.2307/2411227
30. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330
31. Cannon SB, Shoemaker RC. Evolutionary and comparative analyses of the soybean genome. Breed Sci. 2012;61: 437–444. doi:10.1270/jsbbs.61.437
32. Kim Y, Neilsen R. Linkage disequilibrium as a signature of selective sweeps. Genetics. 2004;167: 1513–1524. doi:10.1534/genetics.103.025387
33. Flint-Garcia SA, Thornsberry JM, Buckler ES. Structure of Linkage Disequilibrium in Plants. Annu Rev Plant Biol. 2003;54: 357–374. doi:10.1146/annurev.arplant.54.031902.134907
34. Mackay I, Powell W. Methods for linkage disequilibrium mapping in crops. Trends Plant Sci. 2007;12: 57–63. doi:10.1016/j.tplants.2006.12.001
35. Gabriel SB. The Structure of Haplotype Blocks in the Human Genome. Science (80-). 2002;296: 2225–2229. doi:10.1126/science.1069424
36. Priolli R, Wysmierski P. Genetic structure and a selected core set of Brazilian soybean cultivars. ... Mol Biol. 2013;36: 382–390. doi:10.1590/S1415-47572013005000034
37. Gwinner R, Alemu Setotaw T, Pasqual M, Dos Santos JB, Zuffo AM, Zambiazzi EV, et al. Genetic diversity in Brazilian soybean germplasm. Crop Breed Appl Biotechnol. 2017;17: 373–381. doi:10.1590/1984-70332017v17n4a56
38. Mamidi S, Chikara S, Goos RJ, Hyten DL, Annam D, Moghaddam SM, et al. Genome-Wide Association Analysis Identifies Candidate Genes Associated with Iron Deficiency Chlorosis in Soybean. Plant Genome J. 2011;4: 154. doi:10.3835/plantgenome2011.04.0011

39. Tomicic M, Djordjevic V, Obreht D, Miladinovic J, Brbaklic L, Mikic A, et al. Tracking footprints of selection associated with soybean adaptation to Central-East Europe environments. *Euphytica*. Springer Netherlands; 2015;203: 701–713. doi:10.1007/s10681-014-1313-z
40. Priolli RHG, Pinheiro JB, Zucchi MI, Bajay M, Vello NA. Genetic Diversity among Brazilian Soybean Cultivars Based on SSR Loci and Pedigree Data. *Brazilian Arch Biol Technol*. 2010;53: 519–531. doi:10.1590/S1516-89132010000300004
41. Keim P, Shoemaker RC, Palmer RG. Restriction fragment length polymorphism diversity in soybean. *Theor Appl Genet*. 1989;77: 786–792. doi:10.1007/BF00268327
42. Keim P, Beavis W, Schupp J, Freestone R. Evaluation of soybean RFLP marker diversity in adapted germ plasm. *Theor Appl Genet*. 1992;85: 205–212. doi:10.1007/BF00222861
43. Hartl DL, Grant A. *Principles of Population Genetics*. 4th Edition. Sinauer Assoc. 2007;14: 544–545. doi:10.2980/1195-6860(2007)14[544b:POPG]2.0.CO;2
44. Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. *Annu Rev Genet*. 2013;47: 97–120. doi:10.1146/annurev-genet-111212-133526
45. Hyten DL, Song Q, Zhu Y, Choi I-Y, Nelson RL, Costa JM, et al. Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci U S A*. 2006;103: 16666–16671. doi:10.1073/pnas.0604379103
46. Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, et al. Diversity and selection in sorghum: Simultaneous analyses using simple sequence repeats. *Theor Appl Genet*. 2005;111: 23–30. doi:10.1007/s00122-005-1952-5
47. Kayser M, Brauer S, Schädlich H, Prinz M, Batzer M a, Zimmerman P a, et al. Y Chromosome STR Haplotypes and the Genetic Structure of U . S . Populations of African , European , and Hispanic Ancestry Y Chromosome STR Haplotypes and the Genetic Structure of U . S . Populations of African , European , and Hispanic Ancestry. *Genome Res*. 2003; 624–634. doi:10.1101/gr.463003
48. Song Q, Hyten DL, Jia G, Quigley C V., Fickus EW, Nelson RL, et al. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. *PLoS One*. 2013;8: 1–12. doi:10.1371/journal.pone.0054985
49. Hwang E-Y, Song Q, Jia G, Specht JE, Hyten DL, Costa J, et al. A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics*. 2014;15: 1. doi:10.1186/1471-2164-15-1
50. Arumuganathan K, Earle ED. Nuclear D N A Content of Some Important Plant Species. 1991;9: 208–218.

51. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. *Population differentiation as a test for selective sweeps*. 2010; 393–402. doi:10.1101/gr.100545.109
52. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989;123: 585–595. doi:PMC1203831
53. Peyrégne S, Dannemann M, Prüfer K. Detecting ancient positive selection in humans using extended lineage sorting. *bioRxiv*. 2016; 1–35. doi:10.1101/092999
54. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol*. 2015;33: 408–14. doi:10.1038/nbt.3096
55. Sun C, Wang B, Yan L, Hu K, Liu S, Zhou Y, et al. Genome-Wide Association Study Provides Insight into the Genetic Control of Plant Height in Rapeseed (*Brassica napus* L.). *Front Plant Sci*. 2016;7: 1–13. doi:10.3389/fpls.2016.01102
56. Xie W, Wang G, Yuan M, Yao W, Lyu K, Zhao H, et al. Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc Natl Acad Sci*. 2015;112: E5411–E5419. doi:10.1073/pnas.1515919112
57. Meyer RS, Choi JY, Sanches M, Plessis A, Flowers JM, Amas J, et al. Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat Genet*. Nature Publishing Group; 2016;48: 1083–1088. doi:10.1038/ng.3633
58. Qi J, Liu X, Shen D, Miao H, Xie B, Li X, et al. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat Genet*. Nature Publishing Group; 2013;45: 1510–1515. doi:10.1038/ng.2801
59. Xu C, Jiao C, Sun H, Cai X, Wang X, Ge C, et al. Draft genome of spinach and transcriptome diversity of 120 *Spinacia* accessions. *Nat Commun*. Nature Publishing Group; 2017;8: 15275. doi:10.1038/ncomms15275
60. He C, Fu J, Zhang J, Li Y, Zheng J, Zhang H, et al. A gene-oriented haplotype comparison reveals recently selected genomic regions in temperate and tropical maize germplasm. *PLoS One*. 2017;12: 1–15. doi:10.1371/journal.pone.0169806
61. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. *Nat Genet*. Nature Publishing Group; 2012;44: 808–811. doi:10.1038/ng.2309
62. Saïdou A-A, Clotault J, Couderc M, Mariac C, Devos KM, Thuillet A-C, et al. Association mapping, patterns of linkage disequilibrium and selection in the vicinity of the PHYTOCHROME C gene in pearl millet. *Theor Appl Genet*. 2014;127: 19–32. doi:10.1007/s00122-013-2197-3

63. He Q, Kim KW, Park YJ. Population genomics identifies the origin and signatures of selection of Korean weedy rice. *Plant Biotechnol J*. 2017;15: 357–366. doi:10.1111/pbi.12630
64. Benjelloun B, Alberto FJ, Streeter I, Boyer F, Coissac E, Stucki S, et al. Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (*Capra hircus*) using WGS data. *Front Genet*. 2015;6. doi:10.3389/fgene.2015.00107
65. Li LF, Li YL, Jia Y, Caicedo AL, Olsen KM. Signatures of adaptation in the weedy rice genome. *Nat Genet*. Nature Publishing Group; 2017;49: 811–814. doi:10.1038/ng.3825
66. Li Y hui, Zhao S cen, Ma J xin, Li D, Yan L, Li J, et al. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics*. 2013;14. doi:10.1186/1471-2164-14-579
67. Valliyodan B, Dan Qiu, Patil G, Zeng P, Huang J, Dai L, et al. Landscape of genomic diversity and trait discovery in soybean. *Sci Rep*. Nature Publishing Group; 2016;6: 1–10. doi:10.1038/srep23598
68. Hyten DL, Choi IY, Song Q, Shoemaker RC, Nelson RL, Costa JM, et al. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics*. 2007;175: 1937–1944. doi:10.1534/genetics.106.069740
69. Song Q, Hyten DL, Jia G, Quigley C V., Fickus EW, Nelson RL, et al. Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. *G3: Genes|Genomes|Genetics*. 2015;5: 1999–2006. doi:10.1534/g3.115.019000
70. Thorne JC, Fehr WR. Incorporation of High-Protein, Exotic Germplasm into Soybean Populations by 2- and 3-way Crosses1. *Crop Sci*. 1970;10: 652. doi:10.2135/cropsci1970.0011183X001000060012x
71. Vello NA, Fehr WR, Bahrenfus JB. Genetic Variability and Agronomic Performance of Soybean Populations Developed from Plant Introductions1. *Crop Sci*. 1984;24: 511. doi:10.2135/cropsci1984.0011183X002400030020x
72. Thompson JA, Nelson RL, Vodkin LO. Identification of Diverse Soybean Germplasm Using RAPD Markers. *Crop Sci*. 1998;38: 1348. doi:10.2135/cropsci1998.0011183X003800050033x
73. Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet*. 2010;42: 1053–1059. doi:10.1038/ng.715
74. Li Y-H, Reif JC, Jackson SA, Ma Y-S, Chang R-Z, Qiu L-J. Detecting SNPs underlying domestication-related traits in soybean. *BMC Plant Biol*. 2014;14: 251. doi:10.1186/s12870-014-0251-1

75. Guo J, Wang Y, Song C, Zhou J, Qiu L, Huang H, et al. A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): Implications from microsatellites and nucleotide sequences. *Ann Bot.* 2010;106: 505–514. doi:10.1093/aob/mcq125
76. Xue S, Bradbury PJ, Casstevens T, Holland JB. Genetic architecture of domestication-related traits in maize. *Genetics.* 2016;204: 99–113. doi:10.1534/genetics.116.191106
77. Arro J, Park JW, Wai CM, VanBuren R, Pan YB, Nagai C, et al. Balancing selection contributed to domestication of autopolyploid sugarcane (*Saccharum officinarum* L.). *Euphytica.* Springer Netherlands; 2016;209: 477–493. doi:10.1007/s10681-016-1672-8
78. Fang L, Gong H, Hu Y, Liu C, Zhou B, Huang T, et al. Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. *Genome Biol. Genome Biology*; 2017;18: 1–13. doi:10.1186/s13059-017-1167-5
79. Concibido VC, La Vallee B, Mclaird P, Pineda N, Meyer J, Hummel L, et al. Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. *Theor Appl Genet.* 2003;106: 575–582. doi:10.1007/s00122-002-1071-5

FIGURAS

Figura 1. Distribuição dos valores de F_{ST} , estimados pela comparação da população PI e BR, nos 20 cromossomos da soja. No eixo horizontal é representada a posição dos SNPs nos cromossomos da soja e no eixo vertical foram plotados os valores de F_{ST} . Os SNPs acima da linha azul, com $F_{ST} \geq 0,3$, foram utilizados para a análise de sinais de seleção.

Figura 2. Distribuição das regiões candidatas nos 20 cromossomos da soja contendo sinais de seleção diferencial entre as populações BR e PI. (A) regiões identificadas pela análise XP-CLR na população BR; (B) regiões identificadas pela análise XP-CLR na população PI; (C) regiões com redução de diversidade nucleotídica na população BR, detectadas via π_{BR}/π_{PI} ; (D) regiões com redução de diversidade nucleotídica na população PI detectadas via π_{PI}/π_{BR} .

Figura 3. Extensão média dos blocos de haplótipo nos 20 cromossomos da soja avaliados na população de genótipos brasileiros (BR) e na população de genótipos de origens diversas (PI). No eixo horizontal são representados os cromossomos, sendo que no eixo vertical é representado o tamanho médio dos blocos de haplótipos em kb.

Figura 5: Diagrama de Venn das regiões contendo indicativos de sinais de seleção nas populações detectadas pelas análises de F_{ST} , XP-CLR, $\pi(\pi)$ e verificada a co-localização dessas regiões dentro de blocos de haplótipos (BH). (A) Regiões contendo sinais de seleção na população BR. (B) Regiões de sinais de seleção encontrados na população PI.

Figura 4: Distribuição dos QTLs reportados no SoyBase co-localizados em regiões de sinais de seleção na população de genótipos brasileiros (BR), na população de PIs (PI) e regiões encontradas em comum contendo sinais de seleção em ambas as populações (BR e PI). No eixo vertical é representado o número de QTLs reportados. Cada barra vertical representa uma categoria de QTL classificado de acordo com o SoyBase: QTL_fungal, relacionados a controle de fungos; QTL_inorganic, envolvidos no controle de compostos inorgânicos; QTL_insect, controle de inseto; QTL_leaf-stem, caracteres da folha e caule; QTL_misc, caracteres diversos; QTL_nematode, controle de nematóides; QTL_oil, envolvidos no teor óleo; QTL_other-seed, caracteres da semente; QTL_pod, caracteres da vagem; QTL_protein, teor de proteína; QTL_reprod-period, caracteres de período reprodutivo; QTL_root, caracteres da raiz; QTL_viral, controle de vírus; QTL_whole-plant, caracteres da planta; QTL_yield, produtividade de grãos.

TABELAS

Tabela 1. Paramêtros de diversidade genética estimados na população de cultivares brasileiras (BR) e de genótipos exóticos (PI).

População	Genótipos	PIC	H _e	RA
BR	15	0,63	0,29	1,90
PI	80	0,62	0,24	1,95

PIC: conteúdo de informação polimórfica; H_e: heterozigosidade esperada; RA: Riqueza alélica

Tabela 2. Regiões dos cromossomos da soja contendo alta diferenciação entre a população de cultivares brasileiras (BR) e a população de genótipos exóticos (PI).

Cromossomo	Início (cM)	Fim (cM)	F _{ST}
1	8,91	8,91	0,38
1	89,67	90,24	0,44
1	98,76	98,79	0,34
2	22,41	23,07	0,34
2	27,31	27,31	0,3
4	17,15	17,41	0,34
5	1,11	1,11	0,31
5	74,79	75,02	0,32
6	9,96	9,96	0,35
6	28,59	28,61	0,33
7	77,75	78,16	0,38
8	26,25	26,25	0,34
8	30,83	31	0,34
8	78	78	0,33
10	0,14	0,14	0,42
10	14,76	15,16	0,33
10	94,62	94,73	0,33
11	0,06	0,82	0,42
11	21,47	21,47	0,33
11	32,38	32,39	0,38
11	43,59	43,59	0,33
11	64,9	64,9	0,35
13	0,04	0,17	0,37
13	44,77	44,77	0,34
13	52,56	63,43	0,46
13	86,65	87,6	0,42
14	6,26	6,26	0,42
14	93,3	93,3	0,42
16	13,4	18,77	0,44
17	28,32	48,57	0,46
17	63,88	64,79	0,31
20	67,38	69,22	0,4
20	75,61	75,95	0,34
20	85,11	85,11	0,33

Cromossomo: cromossomos onde a região foi verificada; Início: posição de início da região representada em cM; Fim: posição final da região representada em cM; F_{ST}: valores de diferenciação entre a população PI e BR.

Tabela 3. Regiões dos cromossomos da soja contendo alta diferenciação pela abordagem XP-CLR na população de cultivares brasileiras (BR), quando comparada a população de genótipos exóticos (PI).

Cromossomo	Início	Fim	XP-CLR
2	23,06	23,31	8,22
4	10,37	10,37	6,24
4	17,15	17,41	6,25
6	38,09	38,09	6,84
6	78,93	78,93	6,48
6	88,34	88,34	6,12
7	5,92	5,93	6,45
7	13,08	13,08	6,28
9	72,81	72,81	6,53
10	0,13	0,13	6,4
11	0,62	0,62	7,22
12	8,72	9,13	6,71
13	55,48	55,48	6,16
14	6,25	6,26	9,72
14	93,29	93,31	7,72
15	7,36	7,36	6,51
15	14,95	15,45	9,26
15	85,8	85,8	8,17
15	97,35	97,35	7,29
16	52,81	52,95	6,32
16	67,5	67,51	7,08
17	53,02	60,38	6,65
18	3,66	3,66	6,1
18	21,17	23,07	6,88
18	105,6	105,6	6,88
20	5,64	5,64	6,52

Cromossomo: cromossomos onde a região foi verificada; Início: posição de início da região representada em cM; Fim: posição final da região representada em cM; F_{ST} : valores de diferenciação entre a população PI e BR.

Tabela 4. Regiões dos cromossomos da soja contendo alta diferenciação pela abordagem XP-CLR na população de genótipos exóticos (PI), quando comparada a população de cultivares brasileiras (BR).

Cromossomo	Início	Fim	XP-CLR
1	8,90	8,92	10,73
2	22,41	23,09	16,93
2	27,30	27,33	12,66
3	87,01	87,23	10,24
4	17,15	17,47	13,33
6	30,04	30,04	9,21
7	77,83	77,92	9,46
8	26,24	26,26	10,41
10	0,14	0,14	8,31
10	93,35	94,74	15,08
11	21,77	21,80	16,95
13	0,04	0,19	18,04
14	4,70	4,72	8,53
14	6,26	6,28	21,63
14	93,29	93,30	8,26
16	50,65	50,67	9,19
16	52,17	52,82	10,20
19	100,02	100,02	8,07
20	67,37	67,40	10,33
20	75,69	76,05	10,11

Cromossomo: cromossomos onde a região foi verificada; Início: posição de início da região representada em cM; Fim: posição final da região representada em cM; F_{ST} : valores de diferenciação entre a população PI e BR.

Tabela 5. Regiões dos cromossomos da soja contendo redução de diversidade nucleotídica (π_{PI}/π_{BR}) na população de cultivares brasileiras (BR), quando comparada a população de genótipos exóticos (PI).

Cromossomo	Início	Fim	π_{PI}/π_{BR}
2	21,7	21,7	3,20
3	0,34	0,52	2,83
3	81,84	81,84	2,73
4	89,47	89,47	2,62
4	96,44	96,45	2,62
5	74,31	74,31	2,83
5	80,13	81,49	3,12
6	1,58	1,58	2,73
6	4,23	4,34	2,73
6	20,39	20,53	2,83
6	22,20	22,2	2,62
6	35,36	35,36	2,73
6	88,4	88,89	3,03
6	92,21	92,21	2,73
7	85,56	85,56	2,83
8	34,41	34,41	3,20
9	17,16	17,20	2,83
9	76,06	76,55	3,03
9	80,73	86,91	2,83
10	12,72	12,72	3,28
10	26,05	26,12	3,12
10	77,84	77,92	2,83
11	10,59	10,6	2,62
11	21,77	21,77	3,89
11	29,87	29,87	3,12
11	33,51	33,52	3,12
11	73,96	73,96	2,73
12	8,53	8,53	2,93
12	20,15	20,64	3,12
13	53,78	56,38	2,93
14	7,26	7,33	3,12
14	17,85	17,93	2,73
14	19,01	19,04	2,83
14	25,64	25,64	2,62
14	39,22	39,22	2,62
15	10,68	11,81	3,12
15	44,66	44,71	2,93
15	93,29	93,35	2,62
16	11,09	11,09	2,73
16	50,65	52,87	3,90
16	56,57	56,57	2,62
16	60,21	61,99	3,20
16	66,92	66,92	2,73
17	20,48	20,48	2,62
17	26,28	26,31	2,73
17	28,27	28,27	2,62
18	3,30	6,28	2,73
18	99,90	99,90	2,73
20	4,70	4,70	2,73
20	68,89	68,97	3,61
20	81,96	82,13	2,73
20	91,81	91,98	2,62

Cromossomo: cromossomos onde a região foi verificada; Início: posição de início da região representada em cM; Fim: posição final da região representada em cM; F_{ST} : valores de diferenciação entre a população PI e BR

Tabela 6. Regiões dos cromossomos da soja contendo redução de diversidade nucleotídica (π_{BR}/π_{PI}) na população de genótipos exóticos (PI), quando comparada a população de cultivares brasileiras (BR).

Cromossomo	Início	Fim	π_{BR}/π_{PI}
1	4,51	6,70	5,57
1	11,91	12,72	6,75
1	89,97	90,24	9,37
1	98,76	98,79	8,25
1	110,09	110,09	6,75
4	96,06	96,06	5,57
5	72,52	72,52	6,75
6	23,93	23,93	6,75
7	13,32	13,32	6,75
7	76,61	78,16	5,57
8	25,91	25,91	6,75
8	30,83	31,00	8,25
8	44,14	44,14	5,57
9	4,83	4,90	6,75
9	6,93	7,19	5,57
10	0,14	0,14	16,29
10	13,17	15,16	6,75
10	79,78	81,16	5,57
10	94,73	94,73	6,33
11	0,06	0,62	16,29
11	0,68	0,82	8,25
11	20,94	20,94	6,75
11	32,39	32,39	6,33
12	78,70	79,13	6,75
13	41,35	41,65	6,75
13	44,77	44,77	8,25
13	48,71	48,71	6,75
13	62,75	62,75	5,57
13	86,60	88,62	5,57
14	93,30	93,32	16,29
16	10,59	10,67	5,57
16	13,40	13,40	8,25
16	18,77	20,33	9,37
20	66,23	66,23	6,75
20	69,22	69,78	8,25
20	75,69	75,99	16,29

Cromossomo: cromossomos onde a região foi verificada; Início: posição de início da região representada em cM; Fim: posição final da região representada em cM; F_{ST} : valores de diferenciação entre a população PI e BR.

Tabela 7. Regiões do genoma da soja com fortes indicativos da presença de sinais de seleção na população BR.

Cromossomos	Início	Fim	F _{ST}	XP-CLR	π_{PI}/π_{BR}	Haplótipos	SS
2	21,06	23,31	0,34	8,22	3,20	219,10	SS
4	17,15	17,41	0,34	6,25	-	337,94	SS
5	74,79	75,02	0,32	-	2,83	63,93	SS
6	88,34	88,34	-	6,12	3,03	-	SS
10	0,13	0,13	0,42	6,4	-	-	-
11	0,62	0,62	0,42	7,22	-	401,84	SS
11	21,47	21,47	0,33	-	3,89	78,56	SS
11	32,38	32,39	0,38	-	3,12	18,68	SS
12	8,72	9,13	-	6,71	2,93	-	SS
13	55,48	55,48	0,46	6,16	2,93	-	SS
14	7,26	7,33	0,42	9,72	3,12	-	SS
14	93,30	93,3	0,42	7,72	-	-	-
16	50,65	52,87	-	6,32	3,90	601,69	SS
16	66,50	67,51	-	7,08	2,73	-	SS
17	28,27	28,27	0,46	-	2,62	-	SS
18	3,30	6,28	-	6,1	2,73	12,55	SS
20	4,64	5,64	-	6,52	2,73	-	SS
20	68,89	68,97	0,40	-	3,61	-	SS

Cromossomo: cromossomos onde a região foi verificada; Início: posição de início da região representada em cM; Fim: posição final da região representada em cM; SS: considerado contendo fortes indicativos da presença de sinais de seleção; “-“: o valor verificado foi abaixo do threshold utilizada para o respectivo teste e/ou não foi verificada a região nos testes.

Tabela 8. Regiões do genoma da soja com fortes indicativos da presença de sinais de seleção na população PI.

Cromossomo	Início	Fim	F_{ST}	XP-CLR	π_{BR}/π_{PI}	Haplótipos	SS
1	8,90	8,92	0,38	10,73	-	8,03	SS
1	89,96	90,24	0,44	-	9,37	-	SS
1	98,75	98,79	0,34	-	8,25	-	SS
2	22,41	23,09	0,34	16,93	-	95,71	SS
2	27,30	27,33	0,3	12,66	-	9,43	SS
4	17,15	17,47	0,34	13,33	-	-	-
7	76,61	78,16	0,38	9,46	5,57	60,19	SS
8	26,24	26,26	0,34	10,41	-	98,54	SS
8	30,83	31,00	0,34	-	8,25	300,75	SS
10	0,14	0,14	0,42	8,31	16,29	-	SS
10	13,17	15,16	0,33	-	6,75	-	SS
10	94,73	94,73	0,33	15,08	6,33	258,06	SS
11	0,06	0,62	0,42	-	16,29	169,00	SS
11	0,68	0,82	0,42	-	8,25	495,36	SS
11	20,94	20,94	0,33	16,95	6,75	-	SS
11	32,39	32,39	0,38	-	6,33	13,03	SS
13	0,04	0,19	0,37	18,04	-	-	-
13	44,77	44,77	0,34	-	8,25	-	SS
13	62,75	62,75	0,46	-	5,57	46,63	SS
13	86,60	88,62	0,42	-	5,57	-	SS
14	6,26	6,28	0,42	21,63	-	-	-
14	93,30	93,32	0,42	8,26	5,57	-	SS
16	13,40	13,40	0,44	-	8,25	-	SS
20	67,37	67,40	0,4	10,33	-	11,75	SS
20	69,22	69,78	0,34	-	8,25	-	SS
20	75,69	75,99	0,33	10,11	16,29	28,41	-

Cromossomo: cromossomos onde a região foi verificada; Início: posição de início da região representada em cM; Fim: posição final da região representada em cM; SS: considerado contendo fortes indicativos da presença de sinais de seleção; “-”: o valor verificado foi abaixo do threshold utilizada para o respectivo teste e/ou não foi verificada a região nos testes.

Arquivo suplementar 1. Tabela S1. Diversidade nucleotídica (π) média nos cromossomos da soja verificada para cada uma das populações.

Cromossomo	π BR	π PI
1	0,0029	0,0022
2	0,0031	0,0025
3	0,0031	0,0025
4	0,0029	0,0025
5	0,0034	0,0030
6	0,0034	0,0028
7	0,0031	0,0024
8	0,0033	0,0022
9	0,0029	0,0026
10	0,0026	0,0022
11	0,0031	0,0024
12	0,0027	0,0024
13	0,0031	0,0024
14	0,0031	0,0028
15	0,0029	0,0028
16	0,0034	0,0030
17	0,0038	0,0026
18	0,0031	0,0027
19	0,0029	0,0023
20	0,0032	0,0025
Média	0,0031	0,0025